Air pollution exposure assessment in sparsely monitored settings; applying machine-learning methods with remote sensing data in South Africa.

INAUGURALDISSERTATION

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

Von

Oluwaseyi Olalekan Arowosegbe

Basel, 2022

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

PD Dr. Kees de Hoogh, Prof. Dr. Martin Röösli und Dr. Sara Adar.

Basel den 21. Juni 2022

Prof. Dr. Marcel Mayor

Dekan der Philosophisch-Naturwissenschaftlichen Fakultät

To my parents and wife

Table of Contents

ACKNOWLEDGEMENTS

It is been a journey of over three years and I would like to appreciate those who have made my PhD experience a worthwhile experience. Firstly, I want to thank my supervisors Dr Kees de Hoogh and Professor Martin Röösli who mentored and supervised this project. I am immensely grateful for the opportunity to work with both of you and the entire Environmental Exposures and Health research group at Swiss Tropical and Public Health Institute, Basel, Switzerland. I also acknowledge the support of Professor Gueladio Cisse for his support and giving me the opportunity to assist tutor students during the climate change and health courses.

This project would have been possible without the support of South Africa collaborators starting with Professor Aqiel Dalvie, Professor Mohamed Jeebhay and Dr Toyib Olaniyan. I also appreciate the cooperation of the staff at South Africa Weather Services for their support.

Settling down to life in Basel was made easy through the support of Christine Mensch, Nora, Laura and Dagmar. I sincerely appreciate your efforts and support throughout my time in Basel. I am immensely grateful to Moses Kebalepile for giving me a good head start at the beginning of this project.

During my time in Basel, I met great minds and kind people at work and home. I am grateful to my colleagues at the former PhD building. I enjoyed having thoughtful conversations with everyone that I shared this space with and am glad that I crossed paths with you guys. Thank you Apolline Saucy and Behzard Valipour for always listening to my questions and proffering solutions. I would not forget my Basel Muttenz apartment family and Mr Fred Nwoke, for a long time you guys filled the family gap. It was a privilege sharing apartment with everyone at the Hofackerstrasse apartment between 2018 and 2021. I will never forget the numerous memorable moments will share and I will continue to relive some of these moments. My appreciation goes to the Schläppi family, many thanks for hosting me during my first Christmas season in Switzerland in 2018 and for your subsequent invitation to Guttannen. I hope to visit again in the future. My special gratitude to Temitope Adebayo, Babatunde Owolodun and Doris Osei Afriyie for your care and support.

I want to deeply appreciate the continued support of my family - Dad, Mom, Suyi, Ife, Femi and my uncle Professor Sunday Arowosegbe and his family. Thank you for your kindness, prayers, patience and perseverance. I hope I make you all proud. To my wife Olufunke Bankole, I appreciate your love, commitment and perseverance.

Lastly, I want to recognize the funders of this project. The Swiss confederation through the Swiss Government Excellence and South Africa National Research Fund for their contribution towards the Joint Swiss-South Africa Chair in Global Environment Health collaboration.

ABSTRACT

Air pollution is one of the leading environmental risk factors to human health – Both short and long-term exposure to air pollution impact human health accounting for over 4 million deaths. Although the risk of exposure to air pollution has been quantified in different settings and countries of the world. The majority of these studies are from high-income countries with historical air pollutant measurements data and corresponding health outcomes data to conduct such epidemiological studies. Air pollution exposure levels in these high-income settings are lower than the exposure levels in low-income countries. The exposure level in sub-Saharan Africa (SSA) countries has continued to increase due to rapid industrialization and urbanization. In addition, the underlying susceptibility profile of SSA population is different from the profiles of the population in high-income settings. However, a major limitation to conducting epidemiological studies to quantify the exposure response relationship between air pollution and adverse health outcomes in SSA is the paucity of historical air pollution measurement data to inform such epidemiological studies.

South Africa an SSA country with some air quality monitoring stations especially in areas classified as air pollution priority areas have historical particulate matter less than or equal to 10 micrometres in aerodynamic diameter (PM_{10} $\mu g/m^3$) measurement data. PM_{10} is one of the most monitored criteria for air pollutants in South Africa. The availability of satellite-derived aerosol optical depth (AOD) at high spatial and temporal resolutions provides information about how particles in the atmosphere can prevent sunlight from reaching the ground. This satellite product has been used as a proxy variable to explain ground-level air pollution levels in different settings.

This thesis main objective was to use satellite-derived AOD to bridge the gap in ground monitored PM_{10} across four provinces of South Africa (Gauteng, Mpumalanga, KwaZulu-Natal and Western Cape). We collected PM_{10} ground monitor measurements data from the South Africa Weather Services across the four provinces for the years 2010 – 2017. Due to the gaps in the daily PM_{10} across the sites and years. In study I, we compared methods for imputing daily ground-level PM_{10} data at sites across the four provinces for years 2010 – 2017 using random forest (RF) models.

The reliability of air pollution exposure models depends on how well the models capture the spatial and temporal variation of air pollution. Thus, study II explored the spatial and temporal variations in ground monitor PM_{10} across the four provinces for the years 2010 – 2017. To explore the feasibility of using satellite-derived AOD and other spatial and temporal predictor variables, Study III used an ensemble machine-learning framework of RF, extreme gradient boosting (XGBoost) and support vector regression (SVR) to calibrate daily ground-level PM_{10} at 1 × 1 km spatial resolution across the four provinces for the year 2016.

In conclusion, we developed a spatiotemporal model to predict daily PM_{10} concentrations across four provinces of South Africa at 1 × 1 km spatial resolution

for 2016. This model is the first attempt to use a satellite-derived product to fill the gap in ground monitor air pollution data in SSA.

LIST OF SCIENTIFIC PAPERS

I. **Arowosegbe, O.O**., Röösli, M., Künzli, N., Saucy, A., Adebayo-Ojo, T.C., Jeebhay, M.F., Dalvie, M.A. and de Hoogh, K., 2021. Comparing Methods to Impute Missing Daily Ground-Level PM_{10} Concentrations between 2010–2017 in South Africa. International journal of environmental research and public health, 18(7), p.3374.

II. **Arowosegbe, O.O**., Röösli, M., Adebayo-Ojo, T.C., Dalvie, M.A. and de Hoogh, K., 2021. Spatial and Temporal Variations in PM_{10} Concentrations between 2010–2017 in South Africa. International journal of environmental research and public health, 18(24), p.13348.

III. **Arowosegbe, O.O**., Röösli, M., Künzli, N., Saucy, A., Adebayo-Ojo, T.C., Schwartz, J., Kebalepile, M., Jeebhay, M.F., Dalvie, M.A. and de Hoogh, K., 2022. Ensemble averaging using remote sensing data to model spatiotemporal PM_{10} concentrations in sparsely monitored South Africa. Environmental Pollution, 119883.

LIST OF ABBREVIATIONS

AOD	Aerosol optical depth
CAMS	Copernicus Atmosphere Monitoring Service
ECMWF	European Centre for Medium-Range Weather Forecasts
GIS	Geographic Information System
LUR	Land-use regression
LMICs	Low and Middle Income countries
MAIAC	Multi-Angle Implementation of Atmospheric Correction
MAIA	Multi-Angle Imager
MODIS	MoDerate resolution Imaging Spectroradiometer
NO_2	Nitrogen dioxide
NO_x	Nitrogen oxides
O_3	Ozone
PM	Particulate matter
$PM_{0.1}$	Ultrafine particle with aerodynamic less than or equal to 0.1 micrometers
$PM_{2.5}$	PM with aerodynamic less than or equal to 2.5 micrometers
PM_{10}	PM with aerodynamic less than or equal to 10 micrometers
RF	Random Forest
RMSPE	Root mean squared prediction error
RMSE	Root mean squared error
SAAQIS	South Africa Air Quality Information Systems
SAWS	South Africa Weather Services
SSA	sub-Saharan Africa

SVR	Support Vector Regression
VIIRS	Visible Infrared Imaging Radiometer
WHO	World Health Organization
XGBoost	Extreme gradient boosting

1 Introduction

Globally, air pollution has been identified as an environmental risk factor to human health. In 2019, air pollution moved from 5th to 4th risk factor for mortality, reemphasizing air pollution as a major environmental risk factor. Ambient air pollution accounts for about 63% of the 6.67 million deaths associated with air pollution in 2019 (Murray et al., 2020). Several epidemiological studies have documented a body of evidence on the association between air pollution and adverse health outcomes across several countries on the different continents of the world (Adebayo-Ojo et al., 2022; Pope 3rd, 2000; Sheehan, Lam, Navas-Acien, & Chang, 2016; Southerland et al., 2022; Stafoggia et al., 2022). This evidence has been used to inform air quality policy actions to mitigate the adverse effect of poor air quality. Historically, the Donora and London smog episodes of 1948 and 1952 are earlier seminal works on the influence of exposure to air pollution and provide earlier examples of the importance of air pollution data to drive scientific actions to tackle air pollution challenges (Logan, 1953; Shrenk, Heimann, Clayton, Gafafer, & Wexler, 1949). The threat posed by exposure to air pollution has attracted the attention of both local and international regulatory bodies and has led to the development and continued review of air quality management guidelines. However, the 2021 World Health Organization Air Quality Guidelines not only provide evidence of the adverse effect of exposure to air pollution on human health but also highlight the disparities in exposure to air pollution across the world. The upward trend in the levels of air pollution in low and middle-income countries remains unabated (World Health Organization, 2021).

The limited air quality monitoring networks in sub Saharan (SSA) SSA countries are a big challenge for air quality management (Amegah & Agyei-Mensah, 2017; Health Effects Institute, 2020). The levels of air pollution in most SSA countries are high with above 90% of the population living in areas exceeding WHO air quality guidelines (World Health Organization, 2021). The availability of reliable air quality is central to all air pollution mitigation actions and have implication for epidemiological studies. For example, reliable air quality data is necessary to quantify exposure–response relationships between air pollution and adverse health outcomes. Also, science driven air quality management plans are informed by reliable air quality data (Health Effects Institute, 2020). Although, several air pollution exposure modelling approaches can complement ground-monitored air quality data. The majority of air pollution exposure modelling approaches are defined based on the state of existing ground-monitored air quality data or source of emission inventories (Vienneau, De Hoogh, & Briggs, 2009). This thesis will focus on the availability of air pollutant data, specifically particulate matter less than or equal to 10 micrometers in aerodynamic diameter (PM_{10} μg/m3), the spatial and temporal characteristics of PM_{10} in an SSA country and also explore the feasibility of applying hybrid approaches combining satellite data with ground monitored PM_{10} data, spatial and temporal variables that could help explain the spatial and temporal variation in PM_{10} concentration across the geographical domain.

1.1 Air Pollutants (Particulate Matters)

Particulate matters are air pollutants that are a mixture of solid, liquid and gaseous particles that can suspend in the air and derived from different sources. They are generally classified into sub-categories based on their sizes; particulate matter less than or equal to 2.5 micrometers in aerodynamic diameter ($PM_{2.5}$ µg/m3) also known as fine particles emanating primarily from the combustion of fossil fuels from cars, domestic and industrial activities or anthropogenic sources. Secondarily, $PM_{2.5}$ can be formed through atmospheric chemical reactions of precursor gases such as ammonia, sulfur oxides, nitrogen oxides and organic gases.

Particulate matter less than or equal to 10 micrometers in aerodynamic diameter (PM_{10} µg/m^3) is also referred to as inhalable particles. PM_{10} are mostly formed from dust from construction sites, landfills, transboundary transportation, waste burning and agricultural activities. Lastly, ultrafine particles less than or equal to 0.1 micrometers in aerodynamic diameter ($PM_{0.1}$ µg/m^3) are smaller than both PM_{10} and $PM_{2.5}$. $PM_{0.1}$ primary sources also include anthropogenic activities of fossil fuels combustion. The smaller the particle size, the more dangerous they are to the body because of their ability to penetrate deeper into the body is relative to their size (Hamanaka & Mutlu, 2018; World Health Organization, 2021). Consequently, $PM_{2.5}$ and PM_{10} are priority air pollutants regulated by both local and international health/environmental organizations and are mostly used as a proxy indicator for air pollution. Notably, the revised 2021 WHO Air Quality Guideline provided qualitative recommendation practice for $PM_{0.1}$ based on available evidence (World Health Organization, 2021).

1.2 Spatial and Temporal Variation of Air pollutant

An important characteristic of air pollutants is that they vary in space and time. The complex interplay between meteorological variables and air pollutants sources influences the dispersion and build-up of air pollutants. Thus, the concentration of air pollutants we breathe in differs based on daily meteorological variables and proximity to emission sources. Although, the sources of air pollutants are ubiquitous. However, these sources may differ based on land use and cover features of different areas. For example, sources of air pollution in residential areas consist of the domestic use of fossil fuels for cooking, heating and population density. There are different sources of emission and can be highly variable within and between geographical boundaries based on the distribution of emission sources and meteorological variables. Consequently, the emission source profiles of urban areas are different from rural areas. Urban areas are often characterized by multiple sources of emission due to population density and economic activities linked to urbanization. Thus, it is important to account for sources of emission while modelling air pollution exposure. This can be achieved by using emission inventories data and proxy emission variables such as road density, population density while modelling air pollution exposure.

Meteorological variables also play a significant role in the formation and dispersion of air pollution concentration (Dayan & Levy, 2005). These variables help explain the day to day variation in air pollution. Air temperature aids the movement of warm air near the ground level to the higher troposphere allowing the movement of pollution through the process called convection. However, during the winter season, the layer of warm air can act as a cover keeping cold at ground level thereby trapping cool air and air pollution near the ground level in a process referred to as thermal inversion. Wind speed and direction also influence the dispersion and dilution of air pollutant, wind at high speeds lead to a great dispersion of air pollutants from the source. While precipitation has a scavenging effect on air pollution by washing and dissolving air pollutants from the atmosphere. Meteorological variables can also act synergistically to influence air pollution, the interaction between temperature and sunlight plays a role in the formation of photochemical smog from pollutants.

2 Air pollution exposure models

The applicability of air pollution exposure data for epidemiological studies remains central to air pollution exposure modelling. The relative availability of few conventional reference grade monitoring stations to characterize the population's exposure to air pollution around the geographical extent of where people live and work has led to the application of different methods to estimate human exposure to air pollution.

2.1 Ground-monitored approaches

A simple approach to modelling air pollution exposures is the use of air pollution data of available stations to investigate the association between air pollution and health outcomes. Conventional reference grade monitoring stations have provided air pollution data for exposure assessments. The reference stations data are usually used as a proxy for individual and population exposure based on their proximity to the population's residential address (Wong, Yuan, & Perlin, 2004). However, because air pollution can vary significantly within a small spatial domain, generalizing exposure data from non-representative and limited monitoring stations to individuals or populations away from the monitoring stations can result in exposure misclassification. Consequently, biasing the association between air pollution and health outcomes when used in epidemiological studies. Nonetheless, spatial interpolation techniques such as kriging and inverse distance weighting have been employed to extrapolate exposure concentration in space (Beelen et al., 2008; Künzli et al., 2005). However, their insufficient ability to account for spatial variability of air pollution due to the spatial coverage and representativeness of monitoring sites have restricted their applicability in epidemiology studies (Di, Koutrakis, & Schwartz, 2016) to areas without sufficient air pollution monitoring.

2.2 Land-use regression

Land Use Regression (LUR) is one of the common models used to assess air pollution exposure in epidemiological studies (short and long term). Conceptually, LUR combines air pollution data from selected locations with corresponding land use predictors variables such as roads, population and elevation data that provide insight on the characteristics of air pollution of the selected locations providing air pollution data in statistical frameworks (Briggs et al., 1997; Nieuwenhuijsen, 2015). The final models are subsequently used to estimate air pollution concentrations based on the characteristics of the predictors at the defined spatial level for the epidemiological studies. Initially, traditional LUR focused on capturing spatial variation of air pollutants. However, spatiotemporal LUR models have been developed to account for temporal variability of air pollution by including temporal predictors such as chemical transport models estimates to explain the day-day variation of air pollution (de Hoogh et al., 2016; Di et al., 2016; Kloog et al., 2015). To account for the non-linear relationship between predictors in spatiotemporal LUR, statistical frameworks such as generalized additive models have been used to model the relationship between air pollution concentration and spatiotemporal predictors (Wood, Pya, & Säfken, 2016).

2.3 Dispersion models

In contrast to the previously discussed methods to exposure models that rely on statistical frameworks to relate the relationship between air pollutants and other spatiotemporal predictors (Jones, Thomson, Hort, & Devenish, 2007), dispersion model depend on mathematical formulas that explain the atmospheric processes that drive the movement of air pollutants from emission sources to receptors. The dispersion model uses mathematical assumptions to model air quality concentration based on emissions and meteorological variables (Gibson, Kundu, & Satish, 2013). This model can be used to estimate air pollution concentrations at receptor locations of interest. They are traditionally used by air quality regulatory agencies to track their progress towards the national or international air quality standards (Arya, 1999). Dispersion models have also been used in epidemiological studies to assess air pollution exposure (Ancona et al., 2015). They can also provide air pollution concentration estimates at a high spatial and temporal resolution – an alluring feature of epidemiological interest. However, the use of dispersion models in epidemiological studies has reduced because most dispersion models are based on the correctness of the underlying mathematical assumptions and not actual measurements to model air pollution exposure from available emission sources. Thus, the development of hybrid models that allows the combination of actual measurements and other predictors limits the misclassification of exposure assignment for epidemiological studies (Esmen & Marsh, 1996).

2.4 Hybrid models

The development of new methods or techniques to improve the reliability of air pollution exposure has led to the use of remote-sensing data such as the Aerosol Optical Depth (AOD) (Lyapustin et al., 2011). The spatial and temporal coverage of AOD – a columnar measurement of light extinction or absorption by aerosol particles suspended in the atmosphere have been explored by researchers to create spatial and temporal continuous air pollution exposure maps. AOD is used as a proxy indicator for the ground level of air pollution under the assumption that the number of particles present in the column of air from the atmosphere to the ground level is a function of the level or extent of light extinction or absorbed by suspended aerosols (de Hoogh, Héritier, Stafoggia, Künzli, & Kloog, 2018; Schneider et al., 2020). Therefore, AOD could be an indicator of ground-level air pollution levels. AOD is measured by the moderate resolution imaging spectroradiometer (MODIS) onboard NASA Terra and Aqua satellites and ESA Sentinel satellite (Lyapustin et al., 2011).

The availability of AOD and other satellite products has been complemented by the increasing application of machine learning methods for air pollution exposure assessment in a hybrid statistical framework combining AOD and other spatiotemporal predictors to calibrate ground-level air pollution levels (Schneider et al., 2020; Stafoggia et al., 2017). The ability of machine learning methods to capture the either known or unknown relationship between spatiotemporal predictors of air pollutants is the major advantage of this statistical framework for exposure assessment compared to other multivariable methods. To this end, individual machine learning learners such as random forest and gradient boosting have been used to model the spatial and temporal variation of PM while using AOD as an input variable (Mandal et al., 2020; Schneider et al., 2020; Shtein et al., 2019; Stafoggia et al., 2019). To maximize the predictive potential and the bias–accuracy tradeoff of individual learners or algorithm, averaging the predictions from more than one machine learning learner have also become popular.

3 Epidemiological evidence of the effects of air pollution on health outcomes.

Air pollution is associated with a myriad of adverse health outcomes (Adar et al., 2018; Adar, Filigrana, Clements, & Peel, 2014). Indeed, the exposure of humans to air pollution could be detrimental to human health even at a low concentration (Stafoggia et al., 2022). These associations are scientifically assessed in an epidemiological framework that examines both the short and long-term relationship between air pollution and health outcomes. The common statistical approach employed in shortterm epidemiological studies is a time-series analysis where the air pollution exposure, health outcome and possible confounder variables are aggregated at the city level (Adebayo-Ojo et al., 2022). Because of the distribution of the health outcomes of interest e.g mortality or hospitalization from cardiorespiratory outcomes, a Poisson multivariable regression analysis is used to model the association between daily counts of health outcomes and acute exposure to air pollutants while accounting for potential confounding variables of seasonality, day of the week, meteorological variables (Schwartz, Dockery, & Neas, 1996). Alternatively, the case-crossover method is used to investigate the relationship between the daily mean of air pollutants and health outcomes when air pollution exposure data is available at a fine spatial and temporal resolutions and health outcomes data at the individual level (Jaakkola, 2003). By design, the case-crossover approach accounts for possible confounding variables because all cases with the outcome of interest contribute person-time of days without the outcome of interest – serving as his or her control (Jaakkola, 2003; Lee & Schwartz, 1999). This allows the comparison of air pollution exposure contrast between days with the outcome of interest and days without the outcome of interest. Recently, the Case Time Series design: a novel self-matched modelling approach for the association between short-term exposure to environmental risk factors and acute health outcomes has been introduced as an epidemiological design that combines the longitudinal structure of time-series analysis and self-matched design (Gasparrini, 2021).

Chronic exposure to air pollution also has health consequences on human health. Long-term epidemiological studies are used to investigate if long-term exposure to air pollution is associated with adverse health outcomes in the population that resides in areas with increased air pollution levels compared to those residing in areas with lower air pollution levels. (Adar et al., 2018; Stafoggia et al., 2022) Conceptually, long-term epidemiological studies on air pollution compare the air pollution exposure contrast of those spatially exposed with those less spatially exposed to the long-term air pollution level in cohort study designs. To achieve this, spatially and temporally resolved air pollution concentration is combined with health outcomes data. The multivariable models used commonly account for both individual and area-level variables that might confound the association between long-term exposure to air pollution and adverse health outcomes.

4 Research Gaps

The trend of ambient air pollution has continued on an upward trajectory in SubSaharan Africa (SSA) countries due to a rapid increase in population, urbanization and the resultant increase in emissions from vehicles, industries, unpaved roads, bush burning and biomass (Amegah & Agyei-Mensah, 2017; Health Effects Institute, 2020; World Health Organization, 2020, 2021). Concurrently, SSA is undergoing an epidemiological transition in disease from communicable diseases to non-communicable diseases. Lower respiratory infections is one of the leading cause of death and disability in Africa accounting for about 1 million death per year (World Health Organization, 2020). This suggests air pollution, is an important risk factor contributing to the epidemiological transition in African countries (Gouda et al., 2019). However, there are limited epidemiological studies from African countries and a major reason for the paucity of evidence is the lack of historical air pollution data needed for both short and long-term epidemiological studies.

The earliest epidemiological studies relied on the aggregation of air pollution exposure data at large geographical domains and temporal e.g annual averages for epidemiological studies under the assumption that air pollution does not vary substantially in space (Künzli et al., 2005). However, several studies have reported that this assumption is generally unrealistic (Eeftens et al., 2012). The impact of air pollution exposure misclassification can bias the association between exposure to air pollution and adverse health outcome. Thus, spatially and temporally resolved air pollution exposure are preferred for epidemiological studies. In addition, spatially and temporally resolved air pollution exposure data provides an opportunity to account for individuals' risk profiles such as age and area-level factors such as socio-economic factors that can influence the association between air pollution and adverse health outcomes (Stafoggia et al., 2022).

This thesis aims to assess the feasibility of using remote sensing data, land use, chemical transport model data and other spatial and temporal predictor data to characterize PM_{10} concentration across four provinces of South Africa (Mpumalanga, Gauteng, Western Cape and KwaZulu-Natal).

Specific objectives are:

- To compare methods of imputing daily missing ground-level PM_{10}.
- To investigate the spatial and temporal variation of PM_{10} concentration across the four provinces
- To provide spatially and temporally resolved PM_{10} estimates across the four provinces.

5 References

Adar, S. D., Chen, Y.-H., D'Souza, J. C., O'Neill, M. S., Szpiro, A. A., Auchincloss, A. H., . . . Kaufman, J. D. (2018). Longitudinal analysis of long-term air pollution levels and blood pressure: a cautionary tale from the multi-ethnic study of atherosclerosis. *Environmental health perspectives, 126*(10), 107003.

Adar, S. D., Filigrana, P. A., Clements, N., & Peel, J. L. (2014). Ambient coarse particulate matter and human health: a systematic review and meta-analysis. *Current environmental health reports, 1*(3), 258-274.

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 CardioRespiratory Disease Hospital Admissions in Cape Town, South Africa. *International Journal of Environmental Research and Public Health, 19*(1), 495.

Amegah, A. K., & Agyei-Mensah, S. (2017). Urban air pollution in Sub-Saharan Africa: Time for action. *Environmental Pollution, 220*, 738-743.

Ancona, C., Badaloni, C., Mataloni, F., Bolignano, A., Bucci, S., Cesaroni, G., . . . Forastiere, F. (2015). Mortality and morbidity in a population exposed to multiple sources of air pollution: A retrospective cohort study using air dispersion models. *Environmental research, 137*, 467-474.

Arya, S. P. (1999). *Air pollution meteorology and dispersion* (Vol. 310): Oxford University Press New York.

Beelen, R., Hoek, G., van Den Brandt, P. A., Goldbohm, R. A., Fischer, P., Schouten, L. J., . . . Brunekreef, B. (2008). Long-term effects of traffic-related air pollution on mortality in a Dutch cohort (NLCS-AIR study). *Environmental health perspectives, 116*(2), 196202.

Briggs, D. J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebret, E., . . . Van Der Veen, A. (1997). Mapping urban air pollution using GIS: a regression-based approach. *International Journal of Geographical Information Science, 11*(7), 699-718.

Dayan, U., & Levy, I. (2005). The influence of meteorological conditions and atmospheric circulation types on PM 10 and visibility in Tel Aviv. *Journal of Applied Meteorology, 44*(5), 606-619.

de Hoogh, K., Gulliver, J., van Donkelaar, A., Martin, R. V., Marshall, J. D., Bechle, M. J., . . . Eeftens, M. (2016). Development of West-European PM2. 5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data. *Environmental research, 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. *Environmental Pollution, 233*, 1147-1154.

Di, Q., Koutrakis, P., & Schwartz, J. (2016). A hybrid prediction model for PM2. 5 mass and components using a chemical transport model and land use regression. *Atmospheric Environment, 131*, 390-399.

Eeftens, M., Beelen, R., de Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., . . . 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. *Environmental science & technology, 46*(20), 11195-11205.

Esmen, N., & Marsh, G. (1996). Applications and limitations of air dispersion modeling in environmental epidemiology. *Journal of Exposure Analysis and Environmental Epidemiology, 6*(3), 339-353.

Gasparrini, A. (2021). The Case Time Series Design. *Epidemiology, 32*(6), 829-837. doi:10.1097/ede.0000000000001410

Gibson, M. D., Kundu, S., & Satish, M. (2013). Dispersion model evaluation of PM2. 5, NOx and SO2 from point and major line sources in Nova Scotia, Canada using AERMOD

Gaussian plume air dispersion model. *Atmospheric Pollution Research, 4*(2), 157-167.

Gouda, H. N., Charlson, F., Sorsdahl, K., Ahmadzada, S., Ferrari, A. J., Erskine, H., . . . Aminde, L. N. (2019). Burden of non-communicable diseases in sub-Saharan Africa, 1990–2017: results from the Global Burden of Disease Study 2017. *The Lancet Global Health, 7*(10), e1375-e1387.

Hamanaka, R. B., & Mutlu, G. M. (2018). Particulate matter air pollution: effects on the cardiovascular system. *Frontiers in endocrinology, 9*, 680.

Health Effects Institute. (2020). *The State of Global Air*. Retrieved from Boston, MA:Health Effects Institute:

Jaakkola, J. (2003). Case-crossover design in air pollution epidemiology. *European Respiratory Journal, 21*(40 suppl), 81s-85s.

Jones, A., Thomson, D., Hort, M., & Devenish, B. (2007). The UK Met Office's next-generation atmospheric dispersion model, NAME III. In *Air pollution modeling and its application XVII* (pp. 580-589): Springer.

Kloog, I., Sorek-Hamer, M., Lyapustin, A., Coull, B., Wang, Y., Just, A. C., . . . Broday, D. M. (2015). Estimating daily PM2. 5 and PM10 across the complex geo-climate region of Israel using MAIAC satellite-based AOD data. *Atmospheric Environment, 122*, 409416.

Künzli, N., Jerrett, M., Mack, W. J., Beckerman, B., LaBree, L., Gilliland, F., . . . Hodis, H. N. (2005). Ambient air pollution and atherosclerosis in Los Angeles. *Environmental health perspectives, 113*(2), 201-206.

Lee, J.-T., & Schwartz, J. (1999). Reanalysis of the effects of air pollution on daily mortality in Seoul, Korea: A case-crossover design. *Environmental health perspectives, 107*(8), 633-636.

Logan, W. P. (1953). Mortality in the London fog incident, 1952. *Lancet*, 336-338.

Lyapustin, A., Wang, Y., Laszlo, I., Kahn, R., Korkin, S., Remer, L., . . . Reid, J. (2011). Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm. *Journal of Geophysical Research: Atmospheres, 116*(D3).

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. *Atmospheric Environment, 224*, 117309.

Murray, C. J., Aravkin, A. Y., Zheng, P., Abbafati, C., Abbas, K. M., Abbasi-Kangevari, 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. *The Lancet, 396*(10258), 1223-1249.

Nieuwenhuijsen, M. J. (2015). *Exposure assessment in environmental epidemiology*: Oxford University Press, USA.

Pope 3rd, C. (2000). Epidemiology of fine particulate air pollution and human health: biologic mechanisms and who's at risk? *Environmentl health perspectives, 108*(suppl 4), 713723.

Schneider, R., Vicedo-Cabrera, A. M., Sera, F., Masselot, P., Stafoggia, M., de Hoogh, K., . . . Gasparrini, A. (2020). A satellite-based spatio-temporal machine learning model to reconstruct daily PM2. 5 concentrations across Great Britain. *Remote sensing, 12*(22), 3803.

Schwartz, J., Dockery, D. W., & Neas, L. M. (1996). Is daily mortality associated specifically with fine particles? *Journal of the Air & Waste Management Association, 46*(10), 927939.

Sheehan, M. C., Lam, J., Navas-Acien, A., & Chang, H. H. (2016). Ambient air pollution epidemiology systematic review and meta-analysis: A review of reporting and methods practice. *Environment international, 92*, 647-656.

Shrenk, H., Heimann, H., Clayton, G. D., Gafafer, W., & Wexler, H. (1949). Air pollution in Donora, Pa: epidemiology of the unusual smog episode of October 1948. Preliminary report. *Public health bulletin, 306*.

Shtein, A., Kloog, I., Schwartz, J., Silibello, C., Michelozzi, P., Gariazzo, C., . . . Just, A. C. (2019). Estimating daily PM2. 5 and PM10 over Italy using an ensemble model. *Environmental science & technology, 54*(1), 120-128.

Southerland, V. A., Brauer, M., Mohegh, A., Hammer, M. S., van Donkelaar, A., Martin, R. V., . . . Anenberg, S. C. (2022). Global urban temporal trends in fine particulate matter (PM2· 5) and attributable health burdens: estimates from global datasets. *The Lancet Planetary Health*.

Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., De Hoogh, K., De'Donato, F., . . . Renzi, M. (2019). Estimation of daily PM10 and PM2. 5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. *Environment international, 124*, 170179.

Stafoggia, M., Oftedal, B., Chen, J., Rodopoulou, S., Renzi, M., Atkinson, R. W., . . . Vienneau, D. (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. *The Lancet Planetary Health, 6*(1), e9-e18.

Stafoggia, M., Schwartz, J., Badaloni, C., Bellander, T., Alessandrini, E., Cattani, G., . . . Lyapustin, A. (2017). Estimation of daily PM10 concentrations in Italy (2006–2012) using finely resolved satellite data, land use variables and meteorology. *Environment international, 99*, 234-244.

Vienneau, D., De Hoogh, K., & Briggs, D. (2009). A GIS-based method for modelling air pollution exposures across Europe. *Science of The Total Environment, 408*(2), 255266.

Wong, D. W., Yuan, L., & Perlin, S. A. (2004). Comparison of spatial interpolation methods for the estimation of air quality data. *Journal of Exposure Science & Environmental Epidemiology, 14*(5), 404-415.

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association, 111*(516), 1548-1563.

World Health Organization. (2020). *Leading causes of death and disability*. Retrieved from Geneva: https://www.who.int/data/stories/leading-causes-of-death-and-disability2000-2019-a-visual-summary

World Health Organization. (2021). *WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. .* Retrieved from Geneva: https://apps.who.int/iris/handle/10665/345329

6 Paper 1: Comparing Methods to Impute Missing Daily Ground-Level PM_{10} Concentrations between 2010–2017 in South Africa.

Oluwaseyi Olalekan Arowosegbe [1,2], Martin Röösli [1,2], Nino Künzli [1,2], Apolline Saucy [1,2], Temitope Christina Adebayo-Ojo [1,2], Mohamed F. Jeebhay [3], Mohammed Aqiel Dalvie [3] and Kees de Hoogh [1,2,]*

[1]Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute,
Socinstrasse 57, CH-4002 Basel, Switzerland; oluwaseyiolalekan.arowosegbe@swisstph.ch (O.O.A.);
martin.roosli@swisstph.ch (M.R.); nino.kuenzli@swisstph.ch (N.K.);
apolline.saucy@swisstph.ch (A.S.); temitope.adebayo@swisstph.ch (T.C.A.-O.)

[2]Faculty of Science, University of Basel, CH-4003 Basel, Switzerland

[3]Centre for Environmental and Occupational Health Research, School of Public Health and Family Medicine, University of Cape Town, Rondebosch, 7700 Cape Town, South Africa;
mohamed.jeebhay@uct.ac.za (M.F.J.); aqiel.dalvie@uct.ac.za (M.A.D.)

*Correspondence: c.dehoogh@swisstph.ch

This paper was published in International Journal of Environmental Research and Public Health,18, 7 (2021).

International Journal of
*Environmental Research
and Public Health*

Comparing Methods to Impute Missing Daily Ground-Level PM_{10} Concentrations between 2010–2017 in South Africa

Oluwaseyi Olalekan Arowosegbe [1,2], Martin Röösli [1,2], Nino Künzli [1,2], Apolline Saucy [1,2], Temitope Christina Adebayo-Ojo [1,2], Mohamed F. Jeebhay [3], Mohammed Aqiel Dalvie [3] and Kees de Hoogh [1,2,]*

[1] Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute, Socinstrasse 57, CH-4002 Basel, Switzerland; oluwaseyiolalekan.arowosegbe@swisstph.ch (O.O.A.); martin.roosli@swisstph.ch (M.R.); nino.kuenzli@swisstph.ch (N.K.); apolline.saucy@swisstph.ch (A.S.); temitope.adebayo@swisstph.ch (T.C.A.-O.)

[2] Faculty of Science, University of Basel, CH-4003 Basel, Switzerland

[3] Centre for Environmental and Occupational Health Research, School of Public Health and Family Medicine, University of Cape Town, Rondebosch, 7700 Cape Town, South Africa; mohamed.jeebhay@uct.ac.za (M.F.J.); aqiel.dalvie@uct.ac.za (M.A.D.)

* Correspondence: c.dehoogh@swisstph.ch

Citation: Arowosegbe, O.O.; Röösli, M.; Künzli, N.; Saucy, A.; Adebayo-Ojo, T.C.; Jeebhay, M.F.; Dalvie, M.A.; de Hoogh, K. Comparing Methods to Impute Missing Daily Ground-Level PM_{10} Concentrations between 2010–2017 in South Africa. *Int. J. Environ. Res. Public Health* **2021**, *18*, 3374. https:// doi.org/10.3390/ijerph18073374

Academic Editor: Tiziana Schilirò

Received: 31 January 2021 Accepted: 22 March 2021
Published: 24 March 2021

Abstract: Good quality and completeness of ambient air quality monitoring data is central in supporting actions towards mitigating the impact of ambient air pollution. In South Africa, however, availability of continuous ground-level air pollution monitoring data is scarce and incomplete. To address this issue, we developed and compared different modeling approaches to impute missing daily average particulate matter (PM_{10}) data between 2010 and 2017 using spatiotemporal predictor variables. The random forest (RF) machine learning method was used to explore the relationship between average daily PM_{10} concentrations and spatiotemporal predictors like meteorological, land use and source-related variables. National (8 models), provincial (32) and site-specific (44) RF models were developed to impute missing daily PM_{10} data. The annual national, provincial and site-specific RF cross-validation (CV) models explained on average 78%, 70% and 55% of ground-level PM_{10} concentrations, respectively. The spatial components of the national and provincial CV RF models explained on average 22% and 48%, while the temporal components of the national, provincial and site-specific CV RF models explained on average 78%, 68% and 57% of ground-level PM_{10} concentrations, respectively. This study demonstrates a feasible approach based on RF to impute missing measurement data in areas where data collection is sparse and incomplete.

Keywords: air pollution; Random Forest; imputation; particulate matter; environmental exposure; South Africa

1. Introduction

Ambient particulate air pollution is a major environmental risk to health. An estimated 4.14 million mortality in 2019 was associated with exposure to ambient air pollution [1]. Routine ambient air quality measurements at a sufficient spatial and temporal scale are essential for the management and evaluation of ambient air pollution regulations, policies and mitigation measures. They are also crucial for calibrating air pollution statistical models for accurate exposure assessment in epidemiological studies investigating the link between air pollution and health. However, in low- and middle-income countries (LMIC), routine air pollution monitoring stations are sparse due to the limited financial, human and technical capacities to manage these monitoring networks [2,3]. The lack of air pollution measurements in LMIC obstructs the development of aforementioned air pollution models for estimating ambient air pollution exposures and thus informing population health studies.

Particulate matter less than or equal to 10 micrometers in aerodynamic diameter (PM_{10} μg/m^3) is associated with acute and chronic adverse health outcomes and it is of high public health significance globally [1,4,5]. PM_{10} is one of the criteria air pollutants in most countries including South Africa, it is measured in South Africa by an air quality monitoring network managed by three levels of government (National, Provincial and Metropolitan/Local) and privately managed air quality monitoring stations [6]. However, due to limited air quality management capacities, these monitoring stations are concentrated around the designated air pollution priority areas. To date, four areas (located in four of the nine provinces) of South Africa have been designated an air pollution priority area; Vaal triangle, Highveld, South Durban Basin and Waterberg based on historical evidence of poor ambient air quality due to the presence of possible source of air pollution [7]. The quality of available data is a major concern with only a small number of South Africa's ambient air pollution monitoring stations accredited by South African National Accreditation System [8].

In South Africa, air quality measurements are often missing due to various reasons such as vandalization of monitoring facilities, and periodic interruption of measurements due to electrical shut down or breakdown of monitoring equipment. This has led to a significant number of monitoring stations being out of operation for months or years resulting in long time-series of PM_{10} measurements missing [9]. Inconsistent air quality data hampers epidemiological studies in South Africa from investigating the association between air pollution and health. Previous studies in South Africa have documented the trends in air pollutants for raising public health awareness about the need for air pollution control [10–13].

Univariable methods of unconditional mean or median, nearest neighbour have been compared with multivariable methods from regression models using other environmental predictors' for imputing daily PM_{10} measurements [14,15]. Multivariable methods were reported to be more robust in performance when the proportion of missing data are higher than 10% [14]. The relationship between $PM_{2.5}$ and PM_{10} at co-located monitoring sites was explored using multivariable methods with the aim to predict $PM_{2.5}$ at sites with PM_{10} data only in Switzerland and India [16,17]. However, this approach is not feasible in South Africa due to the paucity of $PM_{2.5}$ data as it was only designated a criteria air pollutant in 2012 [18].

Random forest (RF)—a machine learning method can be classified as a multivariable method that aggregates the predictions of several regression trees to improve the performance of single regression models. Several studies have been published using RF and other multivariable models to predict missing air pollutants in areas with no or sparse monitoring networks [16,17,19–21]. However, this study aims to leverage on the spatial and temporal dependence characteristics of air pollutants [22,23], by combining observed PM_{10} data with spatial and temporal predictors as well as chemical transport estimates of PM_{10}, ozone (O_3) and nitrogen dioxide (NO_2) in a RF model to predict missing daily PM_{10} observation in some monitoring stations across four provinces of South Africa for years 2010–2017. The result of this analysis will be subsequently used to construct models to predict PM_{10} in areas without monitoring sites.

2. Materials and Methods

2.1. Methods

The RF machine learning method was employed to accommodate the non-linear relationship between PM_{10} measurements and covariates. For each year we constructed RF models at 3 geographical scales to predict missing daily PM_{10} data: (1) one national model, using all daily PM_{10} measurements from the four provinces combined; (2) four provincial models using daily PM_{10} monitoring measurements from sites within each province; and (3) site specific models exclusively using daily PM_{10} measurements from individual sites.

2.2. Monitoring Sites

The focus of this investigation was on PM_{10} monitoring sites in South Africa, which are located in Mpumalanga, Gauteng, Western Cape and KwaZulu-Natal (Figure 1). These stations are managed by the Department of Environmental Affairs, South Weather Services, provincial, local governments and private industries. Hourly PM_{10} data from the four provinces were obtained from the South African Air Quality Information System (SAAQIS) for 61 monitoring sites (27 in Gauteng, 17 in Mpumalanga, 10 in Western Cape, 7 in Kwazulu-Natal) for the study period 1 January 2010–30 December 2017. Air quality monitoring stations instruments were serviced and calibrated bi-weekly, undergoing a full calibration annually, using National Metrology Institute of South Africa certified gases. The number of sites per year varies across the study period. Figure 2 shows the data completeness of the PM_{10} observations obtained from the SAAQIS by province between 2010 and 2017. SAAQIS provides PM_{10} data for research purposes in South Africa upon completion of the required data disclosure forms. SAAQIS can be reached via their website (https://saaqis.environment.gov.za/. Accessed on 22 October 2018).

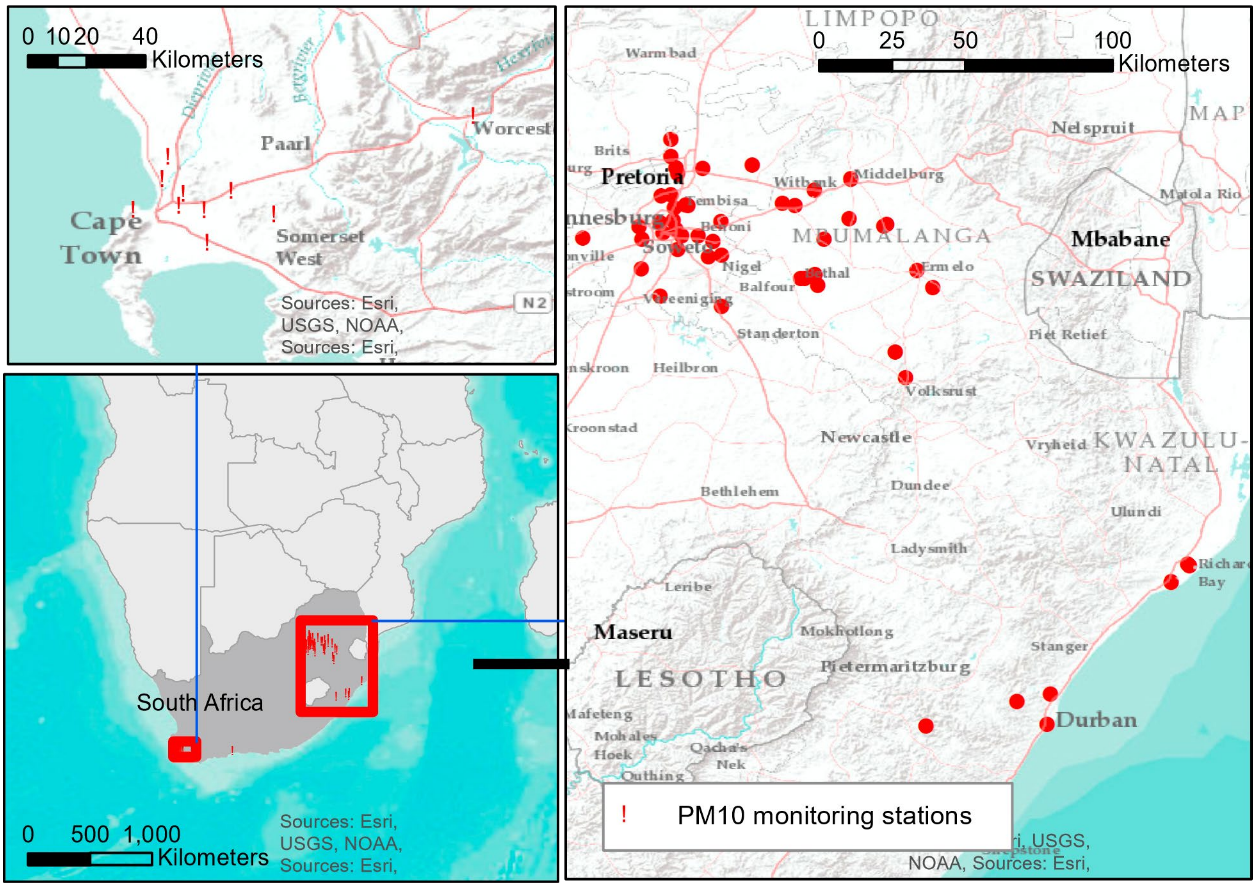


Figure 1. The spatial distribution of particulate matter (PM_{10}) monitoring stations across the four provinces of South Africa operating at some point during 2010–2017.

To ensure quality of the PM_{10} data, the following quality check filters were applied. All negative values or observations greater or less than four times the interquartile range of each monitoring stations were considered outliers and were subsequently removed. A threshold of 75% hourly data per day was used to aggregate hourly data to a daily mean concentration.

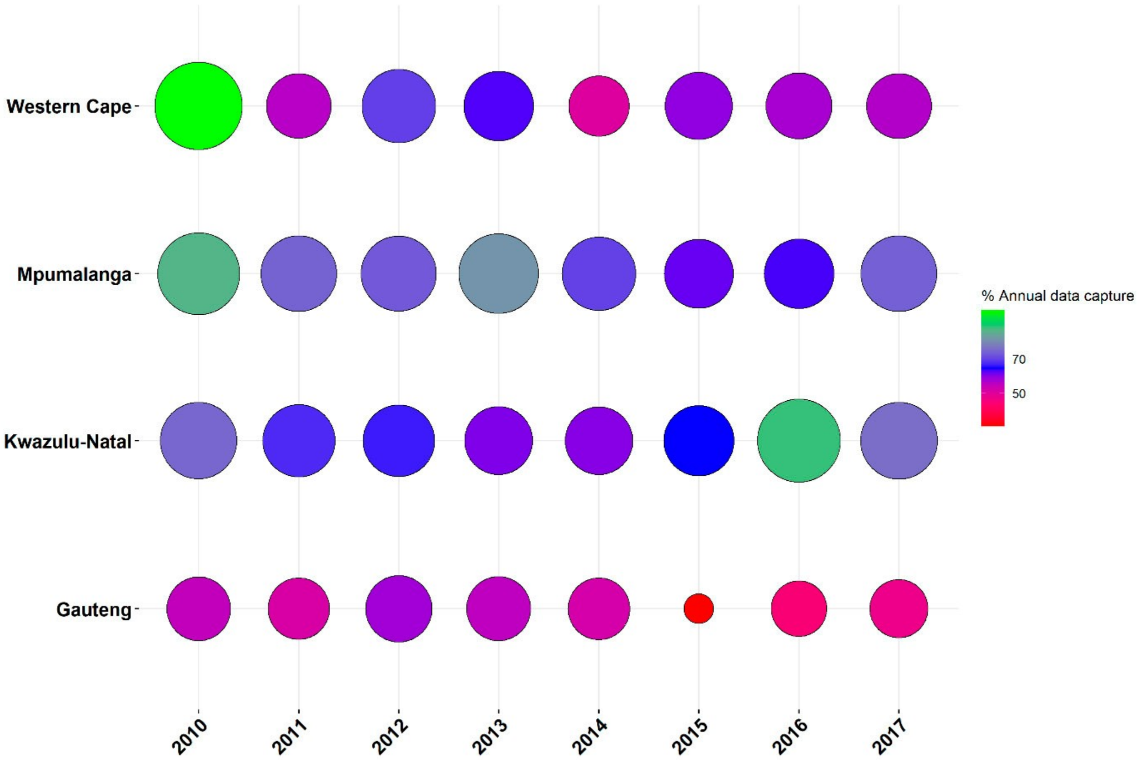


Figure 2. PM_{10} data availability by year and by province—the size and colour of the circles indicate percentage of data capture per year.

2.4. Temporal Parameters

Daily meteorological parameters of total precipitation, boundary layer height, temperature, the component of the horizontal wind towards east (U wind component) and the component of the horizontal wind towards north (V wind component) at a spatial resolution of 0.125 × 0.125° (approximately 10 × 10 km^2) for the hour 12:00:00 were downloaded from the European Centre for Medium-Range Weather Forecasts Reanalysis 5th Generation (ERA5) global climate reanalysis dataset for the year 2010–2017 for South Africa. The U and V wind components were subsequently used to calculate wind speed (ws) and wind direction (wd) respectively using the formulas below:

$$\mathrm{wd} = a\ tan\ 2(-u_{10},\ -v_{10}) \times \frac{\pi}{180} \quad (1)$$

$$\mathrm{ws} = \sqrt{u_{10}^2 + v_{10}^2} \quad (2)$$

In addition to Copernicus Atmosphere Monitoring Service (CAMS) Reanalysis PM_{10} estimates, columnar daily ensemble estimates of pollutant gases of nitrogen dioxide, ozone were also downloaded from the CAMS data store at 0.125 × 0.125° (approximately 10 × 10 km^2). All temporal predictors were resampled at a 1 × 1 km^2 resolution, matching the 1 × 1 km^2 resolution of the raster specifically constructed for this study. The monitoring stations locations were subsequently linked to this raster to extract the temporal predictors.

A number of spatial geographic information system (GIS) predictor variables were calculated for this study at the aforementioned 1 × 1 km^2 grid (see Table 1). South Africa's road network was obtained from OpenStreetMap. For each 1 × 1 km^2 grid cell, we calculated the sum of road length for two categories: major roads and other roads. Land cover data were extracted from the 2018 South Africa National Land cover dataset. The initial 72 land use classes were re-categorized into five major categories: residential; industrial; built-up; agriculture; and water bodies. South Africa's climatic zones were extracted based on the South Africa Bureau of Standards 2005 classification. Population density was obtained from the Socioeconomic data and Application Center (SEDAC) dataset. For the light at night, data extracted from the Visible Infrared Imaging Radiometer Suite-Day/Night Band (VIIRS-DNB) was extracted and averaged at the 1 × 1 km^2 resolution. Elevation and impervious surface were extracted from respectively the Shuttle Radar Topography Mission Digital Elevation Database version 4.1 and the National Oceanic and Atmospheric Administration database.

Table 1. Spatial and temporal predictors used for random forest models

Variable	Description	Source	Resolution
Population density	Mean population within 1 × 1 km^2 grid cell	SEDAC	~1 km
Landcover	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 × 1 km^2 Intersected aggregate	VIIRS-DNB	750 m
Impervious Surface	1 × 1 km^2 Intersected aggregate after removing no data, clouds, shadows data	NOAA	30 m
Elevation	1 × 1 km^2 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, Sub-tropical coastal, Arid interior	South Africa Bureau of Standards 2005	6 Zones
Meteorological variables (daily modelled planetary boundary layer height, temperature, precipitation, wind speed, wind direction, relative humidity, vertical velocity	Daily global ECMWF re-analysis estimates	ERA5-reanalysis	10 × 10 km
Modeled Tropospheric estimates of NO_2, PM_{10}, O_3	Daily Chemical transport model estimate	Chemical transport model Copernicus Atmosphere Monitoring Service (CAMS)	10 × 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 Centre for Medium-Range Weather Forecasts Reanalysis 5th Generation).

2.6. Random Forest Model

RF is a non-parametric machine learning algorithm and an ensemble method that can be used to perform regression for continuous outcome variable (e.g., PM_{10}). Imputation of missing daily PM_{10} data for stations with at least 70% of annual PM_{10} was achieved by combining the measured PM_{10} and spatial and temporal predictor variables at three geographical scales; national, provincial and site specific.

To impute missing PM_{10}, all possible monitoring stations with valid PM_{10} measurements were included in RF analysis. RF was used to estimate the PM_{10} concentration for the missing days by exploring the relationship between observed PM_{10} and spatial and temporal predictors. RF leverages on averaging several independent bootstrap ensemble trees to reduce the variance in the predicted PM_{10} by [24,25]:

1. Randomly resample the data with replacement to create training and validation sets of same sample size as the original dataset.
2. Repeatedly construct regression trees on the training sets and predict on the validation sets.
3. At each trees node, the best predictors from the random subsets of predictors were subsequently used to partition the nodes of respective trees.
4. The final estimate of PM_{10} is the average of individual trees of PM_{10} predictions in a process called bagging.

In this study, the RF parameters number of variables randomly sampled as candidates at each split (mtry) and number of trees to grow (ntree) and minimum number of observations in a terminal node (min.node.size) were selected based on the combinations that minimized out of bag prediction error in the one-third sample left out for validation. Throughout this study, 500 trees were considered. Generally, mtry was tuned at each terminal nodes with two and respective predictors to de-correlate the trees. RF models are less sensitive to parameter tuning for low dimensional data [26]. Similarly, using minimum number of predictors that substantially contribute to explaining the variance in PM_{10} could prevent overfitting the models as RF is prone to overfitting when spatial and temporal variables are included as predictors [27,28].

The feature importance of the models was ranked based on predictors that reduced prediction error when used as splits over the ensemble trees in the RF models. For all the RF models, the faster implementation of RF via the ranger packages was accessed from the caret package in R [29].

2.7. Model Validation

Spatial and temporal cross-validation was used to assess the daily PM_{10} models prediction errors in time and space. Spatial leave one location out cross-validation (LOLO CV) was used to evaluate the national and provincial models. The national model was split into four folds using the province as splitting criterion. Thus, a model was trained on data from all but one province (n − 1). The hold-out provinces sites were iteratively used to estimate the prediction errors of using these models to predict for sites not included in the training data. Sites were used as the splitting criterion for the different provincial models. To account for possible spatial autocorrelation in the models, a complete time-series of observations of a site was sequentially withheld (n − 1) for cross-validation. Spatial LOLO CV was not possible for the site-specific models. Temporal leave time out cross-validation
(LTO CV) was used to assess the model's performance in time. Day of the year was used to split the dataset 10 fold. All three models were sequentially trained on all but one held-out fold. All the models cross-validation were implemented using CAST (Caret Applications for Spatial-Temporal Models) package—a caret package wrapper for spatial and temporal cross-validation [28].

2.8. Error Metrics

Coefficient of determination (R^2), the square of the correlation coefficient between the observed and predicted daily PM_{10} observation was used to evaluate the variance explained by the models. For all the models but sites models, we computed three R^2 measures to assess the models performance. The model building R^2 describing the overall models ability to explain the variance between observed and predicted daily PM_{10} observation. Spatial and temporal R^2 to quantify the contribution of the spatial and temporal level to the total variance of daily PM_{10} model predictions on held-out stations and days.

Root mean squared error (RMSE), the square root of the mean quadratic differences between observed and predicted daily PM_{10}.

Mean absolute error (MAE), the average over the absolute differences between the observed daily PM_{10} and predicted daily PM_{10} were also calculated to provide summary estimates of the models prediction errors.

3. Results

3.1. National Model

The RF models combined spatial and temporally predictor variables with ground monitored PM_{10} from all the four provinces to construct national models for 2010–2017 (Table 2). Figure 3 shows the top 15 ranked variable of importance based on the predictors that reduced prediction error when used as splits over the ensemble trees in the RF models. Temporal predictors of chemical transport model-based estimates of PM_{10}, humidity, Julian date and the spatial variable population emerged as influential variables across 2010–2017. The national RF models for 2010 to 2017 explained between 77% and 79% of the variation in daily PM_{10} concentrations. Spatial CV was used to assess the robustness of the models. The R^2 of the spatial and temporal cross validation varies between 0.11 and 0.35 (RMSE:

17.72–29.47 μg/m^3) and 0.77 and 0.79 (RMSE 12.31–16.43 μg/m^3), respectively.

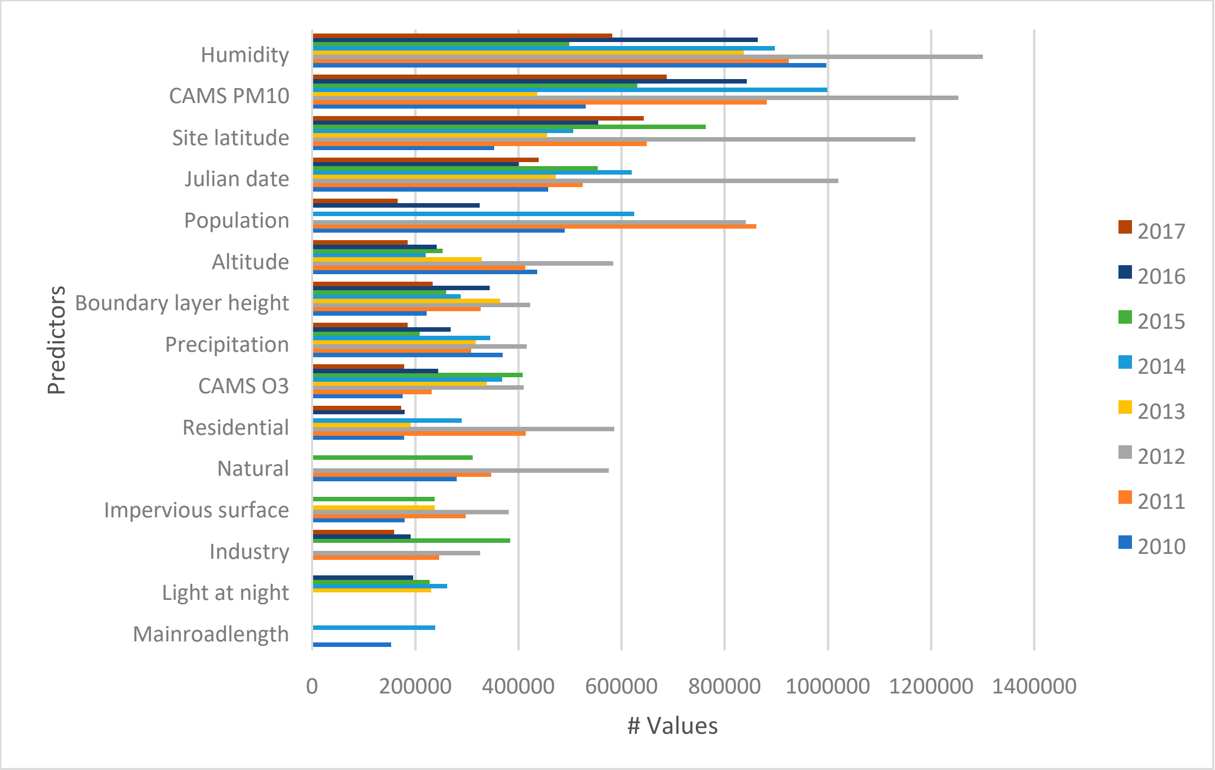


Figure 3. National model variable of importance.

3.2. Provincial Model

The provincial model explored the relationship between PM_{10} and predictor variables by each province across 2010–2017. Supplementary Material Figures S1–S4 highlight chemical transport model-based estimates of PM_{10}, humidity, total precipitation, sites coordinates as variables of importance for explaining the intra-province PM_{10} variability. The contribution of these variables also varied across the study period and provinces— underlying the heterogeneity in the provincial characteristics of PM_{10} concentration.

The performance of the provincial models while predicting PM_{10} for held-out sites varied across provinces and study period (Table 2). The CV results of the RF models for Gauteng, for example, explained between 26% and 52% of spatial variability and between 52% and 79% of temporal variability in measured PM_{10} concentrations. Mpumalanga RF models slightly improved on the Gauteng models with R^2 ranges of 0.39–0.69 (spatial) and 0.73–0.78 (temporal).

Table 2. Summary of model performance statistics over the period 2010–2017 for the national, provincial and site-specific models showing the range of R^2, root mean squared error (RMSE) and mean absolute error (MAE) for the years included.

	Model Building			Spatial LOLO CV			Temporal LTO CV			Data Availability	
	R_2 (Range)	RMSE (Range)	MAE (Range)	R_2 (Range)	RMSE (Range)	MAE (Range)	R_2 (Range)	RMSE (Range)	MAE (Range)	No of Unique Sites	Years
National	0.77–0.79	12.1–16.76	8.69–11.38	0.11–0.35	17.72–29.47	13.62–23.65	0.77–0.79	12.31–16.43	8.85–11.39	20–44	2010–2017
Provincial *											
Mpumalanga	0.73–0.81	14.03–19.35	9.63–12.13	0.39–0.69	22.06–36.21	13.5–29.59	0.73–0.78	13.55–19.21	9.85–12.01	5–17 *	2010–2017
Gauteng	0.49–0.79	10.34–23.36	9.24–16.75	0.26–0.52	19.72–34.25	15.69–29.42	0.52–0.79	15.11–23.43	9.94–16.87	6–18 *	2010–2017
Western Cape	0.29–0.71	6.74–8.73	5.11–6.72	0.35–0.54	7.38–11.22	5.76–8.86	0.44–0.66	6.66–23.29	5.18–17.92	1–11 *	2010–2017
KwaZulu-Natal	0.55–0.79	7.36–9.53	5.29–8.11	0.29–0.57	8.54–19.95	6.95–16.82	0.47–0.78	7.37–10.71	5.46–8	3–6 *	2010–2017
Site-specific ** Beliville											
	0.42–0.47	5.81–9.16	4.51–7.26	NA	NA	NA	0.45–0.49	5.67–9.02	4.45–7.03	NA	2012, 2013, 2015–2017
Bodibeng	0.54–0.63	16.89–19.42	13.61–15.07	NA	NA	NA	0.57–0.67	16.36–18.91	13.32–14.87	NA	2012–2013
Brackenham	0.41–0.49	8.06–8.95	6.31–7.10	NA	NA	NA	0.46–0.49	7.81–8.95	6.25–7.15	NA	2011, 2015–2017
Booysens	0.45–0.67	22.13–22.82	17.99–20.77	NA	NA	NA	0.5–0.71	22.10–25.74	17.87–20.53	NA	2012,2014
Camden	0.38–0.62	10.64–23.27	8.69–17.85	NA	NA	NA	0.39–0.65	10.29–22.43	9.61–17.15	NA	2013, 2015, 2017
CBD	0.38–0.59	6.35–9.55	4.93–7.45	NA	NA	NA	0.41–0.64	6.28–9.23	4.98–7.21	NA	2011–2013, 2015–2017
City Hall	0.45	10.29	7.69	NA	NA	NA	0.48	9.78	7.43	NA	2010
Elandsfontein	0.39–0.52	11.72–12.49	9.38–9.68	NA	NA	NA	0.45–0.57	11.17–11.79	8.99–9.38	NA	2016–2017
Ermelo	0.48–0.76	9.20–18.96	7.69–15.31	NA	NA	NA	0.51–0.77	9.12–19.98	7.54–13.89	NA	2010–2016
Etwatwa	0.63	24.03	18.74	NA	NA	NA	0.69	23.78	18.56	NA	2012
Ferndale	0.68–0.74	3.63–5.42	2.84–3.92	NA	NA	NA	0.65–0.77	3.49–5.38	2.76–3.88	NA	2010–2012
Foreshore	0.32–0.49	5.29–9.76	4.1–7.22	NA	NA	NA	0.33–0.49	5.27–9.58	4.13–7.08	NA	2011–2013,2015–2017
Gangles	0.48–0.74	11.86–13.4	9.22–10.11	NA	NA	NA	0.51–0.75	11.23–11.88	8.96–9.71	NA	2010, 2011, 2013,2014
Germiston	0.42	19.65	14.96	NA	NA	NA	0.44	19.07	14.79	NA	2011
George	0.55–0.56	7.09–8.41	5.49–6.56	NA	NA	NA	0.58	6.95–8.12	5.39–6.34	NA	2010, 2013
Goodwood	0.46–0.57	6.77–8.78	5.26–8.24	NA	NA	NA	0.49–0.59	6.60–8.49	5.29–7.80	NA	2011–2012, 2014–2016
Grootvlei	0.41–0.44	10.76–11.32	8.70–8.87	NA	NA	NA	0.42–0.49	10.65–11.12	8.63–8.82	NA	2011, 2013
Hendrina	0.39–0.71	11.12–17.02	8.32–13.62	NA	NA	NA	0.43–0.74	11.18–16.56	8.36–12.96	NA	2010–2012,2015–2016
Middleburg	0.67–0.81	7.81–19.25	6.08–14.73	NA	NA	NA	0.70–0.82	7.49–18.63	5.92–14.25	NA	2010–2016
Olievenhoutbosch	0.57	34.23	27.01	NA	NA	NA	0.59	34.16	26.98	NA	2012
Orange Farm	0.45–0.69	10.78–19.81	8.57–15.56	NA	NA	NA	0.49–0.71	10.23–19.49	8.28–15.62	NA	2010,2017
Rosslyn	0.55–0.61	5.91–11.49	4.77–9.30	NA	NA	NA	0.52–0.67	5.86–11.05	4.47.8.93	NA	2012–2014
Secunda	0.63–0.77	7.73–25.21	5.86–19.96	NA	NA	NA	0.67–0.77	7.47–24.64	5.75–19.7	NA	2010–2013
Witbank	0.72–0.83	9.21–22.33	7.63–17.27	NA	NA	NA	0.73–0.83	8.79–21.87	7.34–16.75	NA	2010,2013–2016
Komati	0.45–0.83	8.52–28.02	6.61–21.51	NA	NA	NA	0.46–0.84	8.29–27.11	6.5–20.91	NA	2011–2012,2014–2017

Leandra	0.29–0.36	6.63–14	4.86–10.38	NA	NA	NA	0.35–0.4	6.35–13.64	4.81–10.31	NA	2011–2012
Newtown	0.43	22.07	17.52	NA	NA	NA	0.47	21.68	17.27	NA	2012
Phola	0.54–0.65	22.44–28.89	17.83–22.55	NA	NA	NA	0.57–0.65	22.02–28.88	17.48–22.72	NA	2013–2014,2016–2017
Stellenbosch	0.35–0.56	6.34–7.31	4.85–5.67	NA	NA	NA	0.37–0.61	6.26–7.14	4.83–5.62	NA	2012–2013
Tableview	0.36–0.4	5.63–7.04	4.43–5.81	NA	NA	NA	0.38–0.43	5.54–7	4.31–5.6	NA	2011–2013
Tembisa	0.71	17.78	14.09	NA	NA	NA	0.73	17.35	13.89	NA	2011

Table 2. *Cont.*

	Model Building			Spatial LOLO CV			Temporal LTO CV			Data Availability	
	R_2 (Range)	RMSE (Range)	MAE (Range)	R_2 (Range)	RMSE (Range)	MAE (Range)	R_2 (Range)	RMSE (Range)	MAE (Range)	No of Unique Sites	Years
Thokoza	0.56	41.30	29.22	NA	NA	NA	0.57	40.25	28.76	NA	2011
Wallacedene	0.47–0.51	5.53–11.26	4.28–8.9	NA	NA	NA	0.47–0.54	5.52–10.82	4.29–8.69	NA	2012, 2015–2017
Wattville	0.52	39.10	29.09	NA	NA	NA	0.57	37.16	28.57	NA	2012
Club	0.59–0.67	11.01–14.87	8.76–11.86	NA	NA	NA	0.62–0.69	10.7–14.88	8.55–11.99	NA	2012–2014, 2016–2017
Ekandustria	0.46–0.59	11.14–16.83	8.88–13.09	NA	NA	NA	0.50–0.64	10.58–16.43	8.5–12.83	NA	2013–2014
Embalenhle	0.56–0.73	16.48–22.18	11.34–14.69	NA	NA	NA	0.59–0.73	13.31–22.18	11.03–17.86	NA	2012,2014,2016–2017
Verkykkop	0.44–0.49	6.63–9.71	5.53–7.88	NA	NA	NA	0.47–0.48	6.56–9.49	5.33–7.72	NA	2013,2016–2017
Randwater	0.32–0.73	12.99–15.99	9.82–15.83	NA	NA	NA	0.36–0.75	12.08–15.63	9.57–12.19	NA	2013–2017
Esikhaweni	0.43–0.58	9.07.9.45	7.36–7.4	NA	NA	NA	0.44–0.60	8.95–9.35	7.17	NA	2016–2017
Chicken Farm	0.44	13.14	10.44	NA	NA	NA	0.48	12.71	10.21	NA	2017
Kwazamokuhle	0.65	18.10	14.44	NA	NA	NA	0.67	17.10	13.84	NA	2017
Kriel Village	0.62	17.27	13.55	NA	NA	NA	0.66	16.89	13.41	NA	2017
Bosjesspruit	0.51	13.05	10.44	NA	NA	NA	0.55	12.58	10.27	NA	2017

* The provincial models included all possible sites with PM_{10} observation; ** The sites models included the monitoring stations with at least 70% annual PM_{10} observation. NA: Not applicable. These are individual site models—Spatial cross-validation (CV) cannot be perform for models with less than two sites. LOLO: Leave one location out spatial cross-validation; LTO: Leave time out temporal cross-validation. Range: The minimum and maximum values of the statistics metrics from the models across 2010.

3.3. Site-Specific Models

Site-specific or individual site models were used to assess the relationship between PM_{10} and temporal predictor variables if the site have at least 70% annual PM_{10} data. The site-specific models were explored independently from each other. The models for Witbank monitoring station performed best with explaining PM_{10} variability between 72% and 83% (Table 2). Leandra monitoring station performed worst with a range of explained PM_{10} variability between 29% and 36%. The temporal variables of chemical transport model-based estimates of PM_{10}, humidity, Julian date, wind speed, temperature, total precipitation are important variables for explaining PM_{10} variability of the different sites (Supplementary Material Figure S5).

3.4. Models Prediction

Table 3 compares the distribution of observed PM_{10} values against the CV predicted PM_{10} for the three models (national, provincial and site-specific) for days with PM_{10} measurements. The site-specific models outperformed the national and provincial models in capturing the variability in PM_{10}. The mean and the standard deviation of the predicted PM_{10} from the provincial and site-specific models are somewhat comparable to that of the observed PM_{10} concentrations. The range of the predicted mean PM_{10} concentrations from the national models differs substantially from the observed PM_{10} concentrations.

Table 3. Range of the observed versus predicted PM_{10} concentrations (in μg/m^3) for the 3 different models (National, Provincial and Site-specific) averaged over all sites and years (2010–2017) by province for the mean, standard deviation (SD) and 5th, 25th, 50th, 75th and 95th percentiles).

Province		**Mean**	**SD**			**Percentiles**		
		μg/m^3	μg/m^3	5	25	50	75	95
Mpumalanga	Observed	35.70–50.90	17.70–29.10	9.30–15.30	21.40–30.30	32.90–46.20	47.70–71.20	68.20–102.80
	National	34.60–48.60	6.30–11.10	23.70–34.20	29.20–41.10	34.30–47.80	39.50–56.80	45.70–66.50
	Provincial	34.20–46.30	10.40–17.40	17.10–24.70	24.90–33.60	32.20–44.30	42.30–60.40	53.00–75.80
	Site-specific	35.70–52.00	11.40–19.50	18.60–26.10	26.80–37.10	34.30–49.80	43.30–66.90	55.50–85.40
Gauteng	Observed	53.40–58.30	28.40–31.30	16.20–20.30	31.10–35.20	47.50–52.10	71.10–77.10	107.60–115.00
	National	36.30–41.60	10.20–12.90	21.30–24.40	27.00–31.00	34.80–40.70	44.60–52.00	54.00–62.40
	Provincial	52.90–59.40	16.90–17.90	30.80–35.50	40.30–45.40	50.20–56.50	66.10–73.30	81.20–90.00
	Site-specific	53.00–58.40	17.40–19.70	29.30–33.50	37.90–43.10	49.70–54.80	65.60–72.30	84.70–93.20
Western Cape	Observed	19.50–26.70	8.10–11.60	8.50–12.70	13.40–18.70	18.50–25.20	24.30–33.30	35.00–48.10
	National	31.90–49.10	7.10–11.20	22.00–35.90	26.00–41.00	29.90–46.80	36.60–55.40	45.20–71.60
	Provincial	20.00–28.00	39.00–5.50	13.50–20.40	16.70–24.10	20.00–28.00	22.70–31.80	26.90–37.10
	Site-specific	19.50–26.70	4.80–6.60	11.80–17.90	15.90–21.80	18.80–26.20	22.40–30.70	28.00–38.40
KwaZulu-Natal	Observed	24.20–29.80	11.01–14.01	9.50–13.50	15.90–20.01	22.10–26.60	30.70–37.10	45.70–56.60
	National	31.60–43.80	8.20–12.90	21.10–28.40	24.50–33.40	29.00–40.40	37.60–53.00	47.60–66.00
	Provincial	23.90–32.90	5.20–9.50	15.60–21.60	19.20–25.90	22.50–31.60	27.10–39.40	35.40–49.50
	Site-specific	24.20–30.50	6.01–10.02	15.30–19.70	19.10–23.30	23.00–28.30	28.00–36.00	36.00–50.80

4. Discussion

This study explored methods for imputing missing daily PM_{10} measurements in South Africa, while considering the spatial distribution pattern of the sparsely PM_{10} monitoring stations across four provinces of South Africa. The RF models, representing three different geographical domains, exhibit markedly different predictive performances for predicting missing daily PM_{10} measurements across four provinces of South Africa.

The performance of the national models and provincial models decreased considerably when used to predict daily PM_{10} in the LOLO validation. Table 3 indicates that the provincial and site-specific models predicted PM_{10} concentrations do not differ substantially from the observed PM_{10} concentrations in terms of mean and standard deviation. In addition, we constructed a national model for the entire eight years (2010–2017) to compare the performance of this model to the yearly models (not presented in Table 2). The overall performance of the model R^2 of 0.67 (RMSE, 17.70) suggest a reduced performance when compared to the range of the yearly models R^2 of 0.77–0.79 (RMSE 2.10–16.76). The cross-validated spatial R^2 of 0.24 (RMSE, 23.47) is within the range of yearly models R^2 (0.11–0.35), RMSE (17.72–29.47). The better performance of the yearly models might be because most of the PM_{10} sites did not provide measurements consistently through the eight years. Also, the levels of PM_{10} between the years are different due to changing PM_{10} related emission variables. The national model, despite high overall $R^{2'}$s (0.77–0.79), performed poorly in the LOLO CV (R^2 0.11–0.35). This was also reflected in the poor ability to predict the observed PM_{10} concentration (Table 3). This is perhaps not surprising given the large geographical domain of South Africa. The distances between the provinces are substantial (e.g., approximately 1000 km between Western Cape and the other three provinces) and, therefore, they exhibit different local emission characteristics driven by social and economic factors, but also by different climatological differences. The air pollution priority areas of Mpumalanga, Gauteng and KwaZulu-Natal provinces are home to the majority of coal reserves, mining and steel facilities in South Africa. The combined impact of these anthropogenic sources with other local sources of PM_{10} and different climatic zones is likely to result in spatial variation in PM_{10} concentration levels between the provinces resulting in distinct provincial characteristics of PM_{10}, which are not transferable between the provinces.

Our provincial models were based on few monitoring stations relative to the size of the four provinces. For example Western Cape Province, the largest province among the four provinces (area = 129,462 km^2), has only 10 operating sites to capture the variability in PM_{10}. The lack of sufficient representative monitoring sites to capture intra-province variability in PM_{10} could explain the relative poor performance of the provincial and national models. Previous studies also reported on the limitation of regulatory monitoring networks in capturing small-scale spatial variations of pollutant concentrations due to the sparse distribution of the few monitoring stations [30,31].

The site-specific models' PM_{10} predictions did not differ substantially from the distribution pattern of the observed PM_{10} (Table 3). The site-specific RF models, only using temporal predictor variables, were able to capture the observed temporal variability in PM_{10} better than the national and provincial models. Previous studies in India and Switzerland have explored the association between $PM_{2.5}$ and PM_{10} in co-located sites to impute missing daily $PM_{2.5}$ observations. These studies were able to develop imputation models explaining 89% (Switzerland) and 92% (India) variability in $PM_{2.5}$ [16,17]. These two studies were able to use sufficient PM_{10} and $PM_{2.5}$ measurements at co-located sites to inform their models and then apply these to PM_{10} only sites to impute $PM_{2.5}$. In South Africa, there were insufficient co-located sites to follow this approach. Despite this disadvantage, we were able to explain PM_{10} variance by between 29% and 83% in the site-specific models.

This finding highlights the paucity of air quality monitoring data in South Africa where only four provinces provided PM_{10} measurements used for this study. Increasing the number of air pollution monitoring sites in South Africa and improving the data capture will provide more power to model more improved and reliable exposure estimates. Nonetheless, the RF variable of importance ranking across the four provinces indicates that chemical transport model estimates of

PM_{10} and meteorological variables contributed considerably to explaining ground-level PM_{10} across our study area and study period.

5. Conclusions

This study compared three models (national, provincial and site-specific) combining spatial, temporal and chemical transport model-based estimates of PM_{10}, O_3 and NO_2 with observed PM_{10} concentrations to predict missing daily PM_{10} concentrations across 44 monitoring sites in four provinces of South Africa between 2010–2017. Given the extent of air quality monitoring currently conducted in South Africa, the site-specific and provincial models showed a better performance compared to the national models in capturing the variability of ground-level PM_{10}. Thus, our study provides evidence that a model constructed with sites from a province is less generalizable to another province.

The results of this study, complete time-series of daily PM_{10} concentrations containing a mix between measured and imputed PM_{10} concentrations, will be used in subsequent air pollution exposure studies aimed at informing population health studies in South Africa.

Supplementary Materials: The following are available online at https://www.mdpi.com/1660-460 1/18/7/3374/s1, Figure S1: Mpumalanga province random forest variable of importance, Figure S2: Gauteng province random forest variable of importance, Figure S3: Western Cape province random forest variable of importance, Figure S4: KwaZulu-Natal province random forest variable of importance, Figure S5: Site-specific model's random forest variable of importance.

Author Contributions: The following statements should be used Conceptualization, O.O.A., K.d.H. and M.R.; methodology, O.O.A., K.d.H.; formal analysis, O.O.A.; data curation, O.O.A., K.d.H. and A.S.; writing—original draft preparation, O.O.A. and K.d.H.; writing—review and editing, O.O.A., K.d.H., M.R., A.S., M.A.D., M.F.J., N.K., T.C.A.-O.; supervision, K.d.H., M.R., M.A.D.; funding acquisition, M.R., M.A.D., M.F.J., N.K. and K.d.H. All authors have read and agreed to the published version of the manuscript.

Funding: 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.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: We would like to thank the staff of the different monitoring stations and the staff at South Weather Services who provided data for this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. 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.; et al. 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* **2020**, *396*, 1223–1249. [CrossRef]
2. Amegah, A.K.; Agyei-Mensah, S. Urban air pollution in Sub-Saharan Africa: Time for action. *Environ. Pollut.* **2017**, *220*, 738–743. [CrossRef]
3. Fayiga, A.O.; Ipinmoroti, M.O.; Chirenje, T. Environmental pollution in Africa. *Environ. Dev. Sustain.* **2018**, *20*, 41–73. [CrossRef]
4. Pope, C.A., III. Epidemiological basis for particulate air pollution health standards. *Aerosol Sci. Technol.* **2000**, *32*, 4–14. [CrossRef]
5. Pope, C.A., III; Dockery, D.W. Health effects of fine particulate air pollution: Lines that connect. *J. Air Waste Manag. Assoc.* **2006**, *56*, 709–742. [CrossRef] [PubMed]
6. Khumalo, T.J. *2017 State of Air Report and National Air Quality Indicator*; Department of Environmental Affairs: Pretoria, South Africa, 2017.
7. Department of Environmental Affairs. *2nd South Africa Environment Outlook*; Department of Environmental Affairs: Pretoria, South Africa, 2016.
8. Scott, G.M. Development of a Methodology for the Delineation of Air Quality Management Areas in South Africa. Ph.D. Thesis, University of KwaZulu-Natal, Westville, South Africa, 2010.
9. Garland, R.M.; Naidoo, M.; Sibiya, B.A.; Oosthuizen, R. Air quality indicators from the Environmental Performance Index: Potential use and limitations in South Africa. *Clean Air J.* **2017**, *27*, 33–41. [CrossRef]
10. Feig, G.; Garland, R.M.; Naidoo, S.; Maluleke, A.; Van der Merwe, M. Assessment of changes in concentrations of selected criteria pollutants in the Vaal and Highveld priority areas. *Clean Air J.* **2019**, *29*. [CrossRef]
11. Feig, G.T.; Naidoo, S.; Ncgukana, N. Assessment of ambient air pollution in the Waterberg Priority Area 2012–2015. *Clean Air J.* **2016**, *26*, 21–28. [CrossRef]
12. 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.; et al. An air quality assessment in the industrialised western Bushveld Igneous Complex, South Africa. *S. Afr. J. Sci.* **2012**, *108*, 1–10. [CrossRef]
13. Olaniyan, T.; Jeebhay, M.; Röösli, M.; Naidoo, R.N.; Künzli, N.; de Hoogh, K.; Saucy, A.; Badpa, M.; Baatjies, R.; Parker, B.; et al.
The association between ambient NO_2 and $PM_{2.5}$ with the respiratory health of school children residing in informal settlements: A prospective cohort study. *Environ. Res.* **2020**, *186*, 109606. [CrossRef]
14. Junger, W.; De Leon, A.P. Imputation of missing data in time series for air pollutants. *Atmos. Environ.* **2015**, *102*, 96–104. [CrossRef]
15. Gómez-Carracedo, M.P.; Andrade, J.M.; López-Mahía, P.; Muniategui, S.; Prada, D. A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometr. Intell. Lab. Syst.* **2014**, *134*, 23–33. [CrossRef]
16. De Hoogh, K.; Héritier, H.; Stafoggia, M.; Künzli, N.; Kloog, I. Modelling daily $PM_{2.5}$ concentrations at high spatio-temporal resolution across Switzerland. *Environ. Pollut.* **2018**, *233*, 1147–1154. [CrossRef] [PubMed]
17. Mandal, S.; Madhipatla, K.K.; Guttikunda, S.; Kloog, I.; Prabhakaran, D.; Schwartz, J.D.; Geo Health Hub India Team. Ensemble averaging based assessment of spatiotemporal variations in ambient $PM_{2.5}$ concentrations over Delhi, India, during 2010–2016. *Atmos. Environ.* **2020**, *224*, 117309. [CrossRef]
18. The Law Library of Congress. *Regulation of Air Pollution*; Global Legal Research Center: Washington, DC, USA, 2018.
19. Stafoggia, M.; Bellander, T.; Bucci, S.; Davoli, M.; De Hoogh, K.; De'Donato, F.; Gariazzo, C.; Lyapustin, A.; Michelozzi, P.; Renzi, M.; et al. Estimation of daily PM_{10} and $PM_{2.5}$ concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. *Environ. Int.* **2019**, *124*, 170–179. [CrossRef]
20. Stafoggia, M.; Johansson, C.; Glantz, P.; Renzi, M.; Shtein, A.; Hoogh, K.D.; Kloog, I.; Davoli, M.; Michelozzi, P.; Bellander, T. A Random Forest Approach to Estimate Daily Particulate Matter, Nitrogen Dioxide, and Ozone at Fine Spatial Resolution in Sweden. *Atmosphere* **2020**, *11*, 239. [CrossRef]
21. Goldberg, D.L.; Gupta, P.; Wang, K.; Jena, C.; Zhang, Y.; Lu, Z.; Streets, D.G. Using gap-filled MAIAC AOD and WRF-Chem to estimate daily $PM_{2.5}$ concentrations at 1 km resolution in the Eastern United States. *Atmos. Environ.* **2019**, *199*, 443–452. [CrossRef]
22. Wong, D.W.; Yuan, L.; Perlin, S.A. Comparison of spatial interpolation methods for the estimation of air quality data. *J. Expo. Sci. Environ. Epidemiol.* **2004**, *14*, 404–415. [CrossRef]

23. Li, H.Z.; Gu, P.; Ye, Q.; Zimmerman, N.; Robinson, E.S.; Subramanian, R.; Apte, J.S.; Robinson, A.L.; Presto, A.A. Spatially dense air pollutant sampling: Implications of spatial variability on the representativeness of stationary air pollutant monitors. *Atmos. Environ. X* **2019**, *2*, 100012. [CrossRef]

24. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

25. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013; Volume 26.

26. Wright, M.N. *Random Forests: The First-Choice Method for Every Data Analysis?* Leibniz Institute for Prevention Research & Epidemiology: Bremen, Germany, 2019; [delivered 28 September 2019].

27. Freeman, E.A.; Moisen, G.G.; Coulston, J.W.; Wilson, B.T. Random forests and stochastic gradient boosting for predicting tree canopy cover: Comparing tuning processes and model performance. *Can. J. For. Res.* **2016**, *46*, 323–339. [CrossRef]

28. Meyer, H.; Reudenbach, C.; Hengl, T.; Katurji, M.; Nauss, T. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environ. Model. Softw.* **2018**, *101*, 1–9. [CrossRef]

29. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]

30. Eeftens, M.; Beelen, R.; de Hoogh, K.; Bellander, T.; Cesaroni, G.; Cirach, M.; Declercq, C.; Dedel˙ e, A.; Dons, E.; de Nazelle, A.;˙ et al. Development of land use regression models for $PM_{2.5}$, $PM_{2.5}$ absorbance, PM_{10} and PM_{coarse} in 20 European study areas; results of the ESCAPE project. *Environ. Sci. Technol.* **2012**, *46*, 11195–11205. [CrossRef] [PubMed]

31. Wang, M.; Beelen, R.; Basagana, X.; Becker, T.; Cesaroni, G.; de Hoogh, K.; Dedele, A.; Declercq, C.; Dimakopoulou, K.; Eeftens, M.; et al. Evaluation of land use regression models for NO_2 and particulate matter in 20 European study areas: The ESCAPE project. *Environ. Sci. Technol.* **2013**, *47*, 4357–4364. [CrossRe

7 Paper 2: Spatial and Temporal Variations in PM_{10} Concentrations between 2010–2017 in South Africa

Oluwaseyi Olalekan Arowosegbe [1,2,*], Martin Röösli [1,2] , Temitope Christina Adebayo-Ojo [1,2], Mohammed Aqiel Dalvie [3] and Kees de Hoogh[1,2]

[1] Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute, Socinstrasse 57, CH-4002 Basel, Switzerland; oluwaseyiolalekan.arowosegbe@swisstph.ch (O.O.A.); martin.roosli@swisstph.ch (M.R.); temitope.adebayo@swisstph.ch (T.C.A.-O.); c.dehoogh@swisstph.ch (K.d.H)

[2]Faculty of Science, University of Basel, CH-4003 Basel, Switzerland

[3]Centre for Environmental and Occupational Health Research, School of Public Health and Family Medicine, University of Cape Town, Rondebosch, 7700 Cape Town, South Africa; aqiel.dalvie@uct.ac.za (M.A.D.)

*Correspondence: oluwaseyiolalekan.arowosegbe@swisstph.ch

This paper was published in International Journal of Environmental Research and Public Health,
18, 24 (2021).

International Journal of
Environmental Research and Public Health

MDPI

Article

Spatial and Temporal Variations in PM$_{10}$ Concentrations between 2010–2017 in South Africa

Oluwaseyi Olalekan Arowosegbe [1,2,*], Martin Röösli [1,2], Temitope Christina Adebayo-Ojo [1,2], Mohammed Aqiel Dalvie [3] and Kees de Hoogh [1,2]

Citation: Arowosegbe, O.O.; Röösli, M.; Adebayo-Ojo, T.C.; Dalvie, M.A.; de Hoogh, K. Spatial and Temporal Variations in PM$_{10}$ Concentrations between 2010–2017 in South Africa. *Int. J. Environ. Res. Public Health* **2021**, *18*, 13348. https://doi.org/10.3390/ijerph182413348

Academic Editors: Christina H. Fuller and A. Kofi Amegah

Received: 11 November 2021
Accepted: 14 December 2021
Published: 18 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1] Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute, Socinstrasse 57, CH-4002 Basel, Switzerland; martin.roosli@swisstph.ch (M.R.); temitope.adebayo@swisstph.ch (T.C.A.-O.); c.dehoogh@swisstph.ch (K.d.H.)

[2] Faculty of Science, University of Basel, CH-4003 Basel, Switzerland

[3] Centre for Environmental and Occupational Health Research, School of Public Health and Family Medicine, University of Cape Town, Rondebosch, Cape Town 7700, South Africa; aqiel.dalvie@uct.ac.za

* Correspondence: oluwaseyiolalekan.arowosegbe@swisstph.ch

Abstract: Particulate matter less than or equal to 10 µm in aerodynamic diameter (PM$_{10}$ µg/m^3) is a priority air pollutant and one of the most widely monitored ambient air pollutants in South Africa. This study analyzed PM$_{10}$ from monitoring 44 sites across four provinces of South Africa (Gauteng, Mpumalanga, Western Cape and KwaZulu-Natal) and aimed to present spatial and temporal variation in the PM$_{10}$ concentration across the provinces. In addition, potential influencing factors of PM$_{10}$ variations around the three site categories (Residential, Industrial and Traffic) were explored. The spatial trend in daily PM$_{10}$ concentration variation shows PM$_{10}$ concentration can be 5.7 times higher than the revised 2021 World Health Organization annual PM$_{10}$ air quality guideline of 15 µg/m^3 in Gauteng province during the winter season. Temporally, the highest weekly PM$_{10}$ concentrations of 51.4 µg/m^3, 46.8 µg/m^3, 29.1 µg/m^3 and 25.1 µg/m^3 at Gauteng, Mpumalanga, KwaZulu-Natal and Western Cape Province were recorded during the weekdays.

The study results suggest a decrease in the change of annual PM$_{10}$ levels at sites in Gauteng and Mpumalanga Provinces. An increased change in annual PM$_{10}$ levels was reported at most sites in Western Cape and KwaZulu-Natal.

Keywords: particulate matter pollution; PM$_{10}$; South Africa; spatial; temporal

1 . Introduction

The levels of air pollution in sub-Saharan Africa (SSA) have remained high compared to other regions of the world that have witnessed notable improvements [1]. The deteriorating trend of air quality in SSA countries, such as South Africa, has been linked to rapid urbanization, industrialization and the resultant increase in population. South Africa relies significantly on fossil fuel for both industrial and domestic activities—over 80% of power generation is from fossil fuel. Other important sources of air pollution emission in South Africa include bush burning, land-fills, dust from construction sites and wind-blown dust from open land [2,3]. Exposure to ambient air pollution accounted for over four million deaths globally in 2019 [4]. Particulate matter less than or equal to 10 µm in aerodynamic diameter (PM$_{10}$ µg/m^3) is one of the most important pollutants of public health interest that is monitored in South Africa [5]. The revised 2015 National Air Quality standard of daily limit of 75 µg/m^3 and annual limit of 40 µg/m^3 are less stringent than the World's Health Organization's limit of 45 µg/m^3 and annual limit of 15 µg/m^3 [6,7].

The levels of PM$_{10}$ concentration can vary in space and time due to distinct meteorological conditions and anthropogenic sources, such as vehicular, domestic and industrial emissions, between the different provinces in South Africa [8,9]. Several air quality management policies and strategies have been introduced to address the worsening air quality in South Africa [6]. These include the identification and control of priority pollutants, the promulgation of regulations to reduce emissions from industries and the classification of air pollution prone areas as priority areas for efficient management of limited air quality management resources [6,10,11]. To this end, air quality management and monitoring resources are concentrated in four air pollution priority areas, including the Highveld, the Vaal triangle, the South Durban Basin and Waterberg, located in four different provinces (Gauteng, Mpumalanga, Western Cape and KwaZulu-Natal) of South Africa. These four areas were prioritized due to the propensity of the observed or outlook of air quality in these areas to exceed the national air quality standards [6,10,12]. Only a few previous studies of air pollutants have examined the spatial and temporal trends of PM_{10} from sites in these areas and SSA [10,12–15]. This is because of limited measurement data to explore the long-term spatial and temporal patterns of PM_{10} in these areas. A couple of studies have assessed the trend in PM_{10} mostly in air pollution priority areas of Gauteng and Mpumalanga province of South Africa [10,12,13,15,16]. In addition, Onyango et al. described the spatial and temporal variation in PM_{10} concentrations at three sites in Uganda [14].

Our previous study described the quality of ground-level PM_{10} measurements in four provinces of South Africa, Gauteng, Mpumalanga, Western Cape and KwaZulu-Natal, for the years 2010–2017 [17]. The earlier study explored methods to bridge the gap in daily PM_{10} data by imputing missing daily PM_{10} for some sites in these provinces for the study period. This study intends to build on the PM_{10} exposure data from the earlier study to characterize daily PM_{10} spatially and temporally for four provinces of South Africa. To investigate the pattern of change in PM_{10}, we assessed the change in annual PM_{10} average across the sites in these areas for the years 2010–2017. Additionally, we explored the characterization of potential influencing factors of PM_{10} emission around the sites. An improved understanding of the pattern of PM_{10} concentration between the four provinces can play a significant role in informing mitigation actions toward addressing the threat posed by air pollution, especially in low- and middle-income countries, such as South Africa, with limited ground-monitored data.

2. Materials and Methods

In this study, PM_{10} measurements from 44 monitoring sites across four provinces (Gauteng, Mpumalanga, Western Cape and KwaZulu-Natal) of South Africa were included. Hourly PM_{10} from the South African Air Quality Information System (SAAQIS). SAAQIS can be reached via their website (https://saaqis.environment.gov.za/, accessed on 22 October 2018). For our study, we selected, for each year between 2010 and 2017, all sites with more than or equal to 70% of total daily measurement data available during a year [17]. Missing data were imputed using a random forest machine learning method, including spatiotemporal predictors, like meteorological, land use and source-related variables, as described in detail in our previous paper [17]. The combined observed and imputed data were used for this study analysis. The distribution of the sites across the provinces differs substantially (Figure 1). The Vaal triangle airshed Priority Area monitoring network and the Highveld Priority Area air quality-monitoring network that cut across Mpumalanga and Gauteng Provinces were the earliest networks established to monitor ambient air quality in South Africa. The South Africa Weather Service classifies the majority of the sites (21) as industrial sites, 18 as residential sites and 5 as traffic sites. An overview of the state of annual PM_{10} availability is presented in supplementary material Table S1. Thirty-two of the forty-four sites (73%) have more than a year of PM_{10} measurement data.

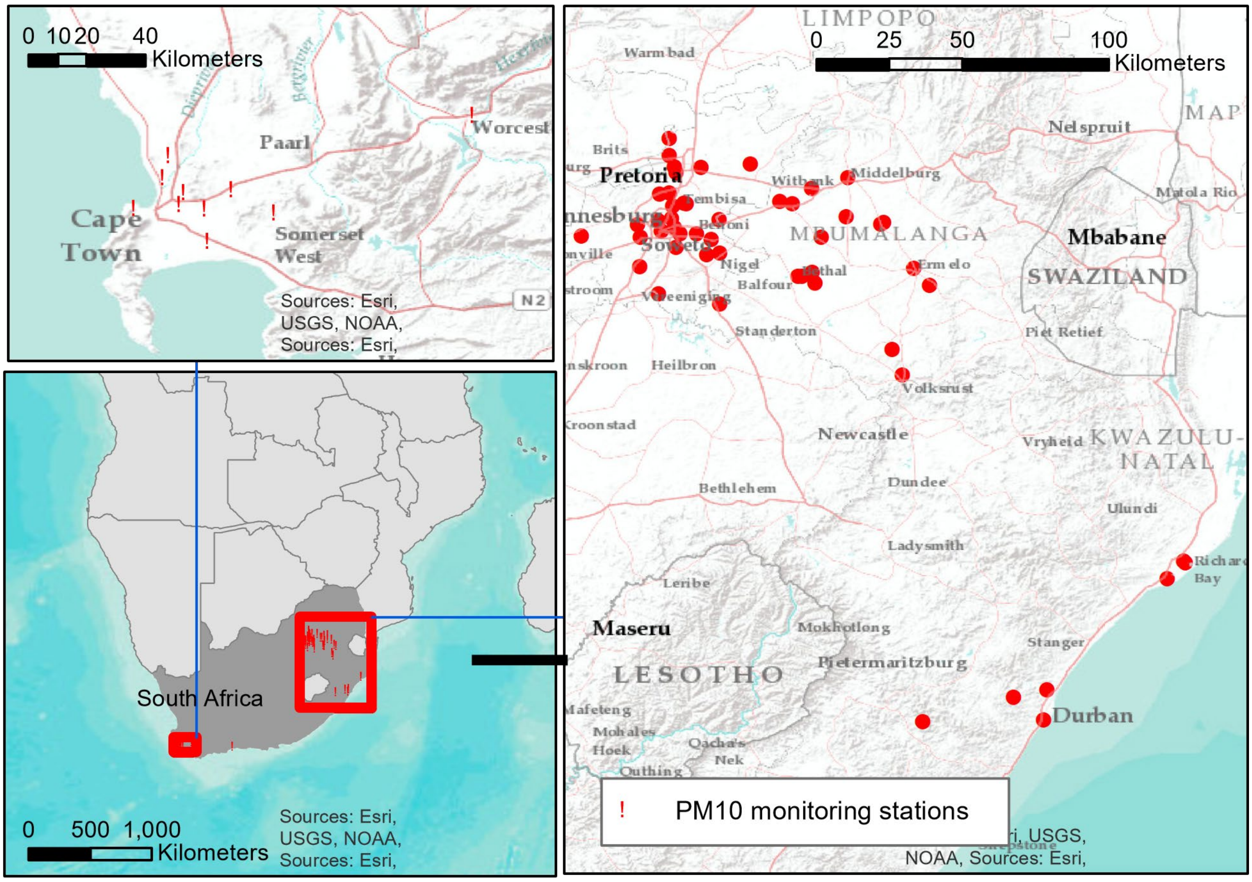


Figure 1. The spatial distribution of particulate matter (PM_{10}) monitoring stations included in this paper across the four provinces of South Africa operating at some point during 2010–2017.

To evaluate the potential influencing factors of PM_{10} around these monitoring sites, we explored multiple buffers (100, 300, 500, 1000, 10,000 m) of land use categories (Residential and Industrial), population density and road density around the monitoring sites. South Africa's road network was obtained from OpenStreetMap (OSM) and the sum of road length was calculated for two categories: (1) major roads defined as roads of OSM types of primary, secondary and tertiary roads and (2) all roads defined as roads of OSM types of residential, service, motorway and trunk. Population density was obtained from the Socioeconomic data and Application Center (SEDAC) dataset. Land use was classified based on the 2018 South Africa National Land cover dataset categories.

To evaluate changes in annual average PM_{10} concentrations for 2010–2017, we applied two formulas. For sites with two consecutive years with average PM_{10} data, the change was calculated by applying the formula:

$$\Delta = \left(\frac{Cx}{Cy} - 1\right) * 100 \quad (1)$$

where Δ is the change, Cx is the annual mean PM_{10} concentration in the current year and Cy is the annual mean PM_{10} concentration in the previous year.

For sites with missing data between successive years, the change in average PM_{10} for a year with average PM_{10} data was calculated by applying:

$$\Delta = \left(\frac{\frac{Cx}{Cy}}{y-x} - 1 \right) * 100 \quad (2)$$

where Δ is the change, Cx is the annual mean PM_{10} concentration in current year y, Cy is the annual mean PM_{10} concentration in the next previous year (year x) with an annual mean PM_{10} concentration and y - x is the number of year(s) between available measurements.

We also calculated annual changes of PM_{10} levels for the 32 sites with more than a year of PM_{10} sites using a linear regression analysis.

3. Results

3.1. Characterization of Sites

Figure 2 presents the level of variation in potential land use, road density and population variables that can provide information about the likely prominent influencing factors of PM_{10} around the site types as designated by the South Africa Weather Service. Buffers of different sizes ranging from 100, 300, 500, 1000, and 10,000 m radii around the sites were considered. Figure 3 presents the analysis for a 300-m buffer. The other buffers sizes did not show substantially different patterns in their distributions. Generally, the distribution of calculated land use, road density and population within a 300 m buffer are in agreement with the monitoring site classification, although industrial land use was actually lower around industrial sites compared to the other two site types. Residential land use and population density were highest for residential classified sites. Major road density within a 300 m buffer was highest for monitoring sites classified as traffic sites.

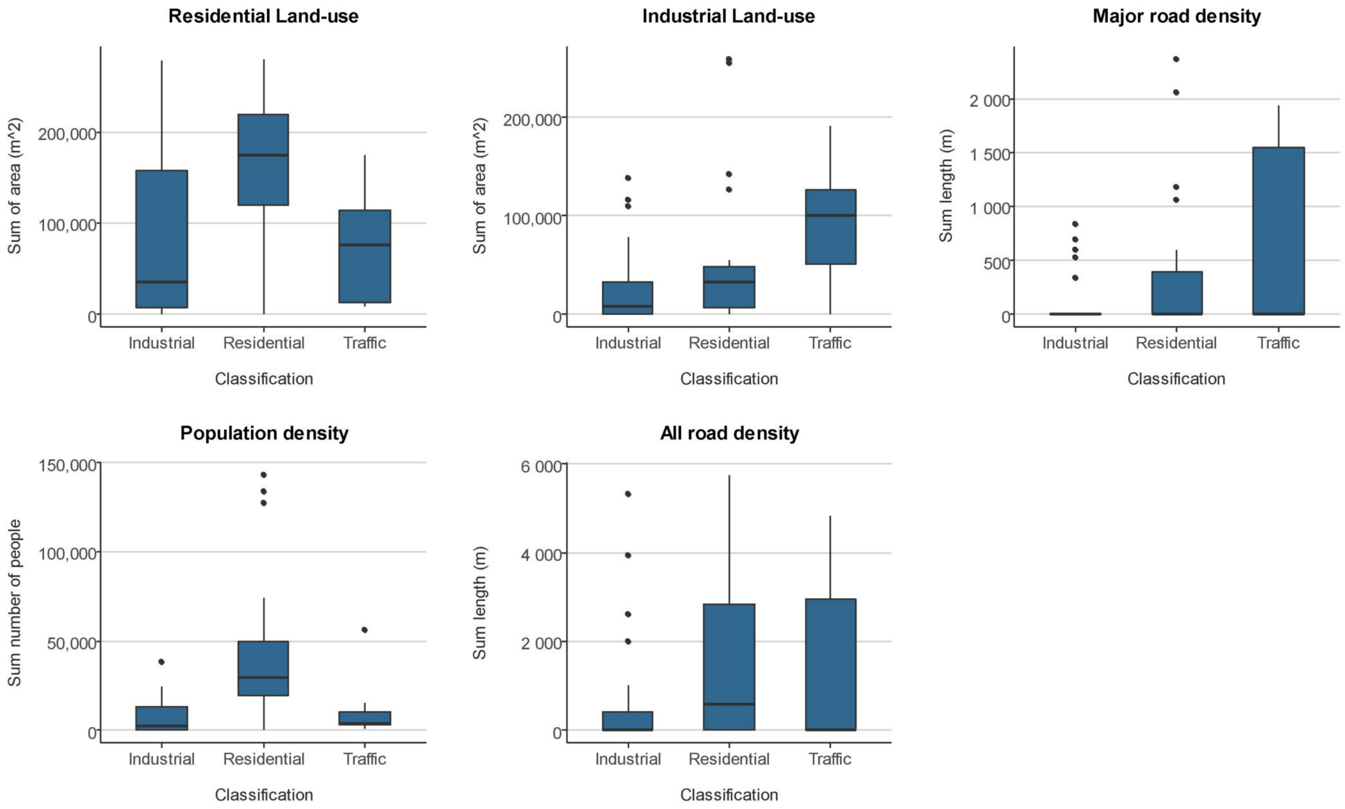


Figure 2. Distribution of indicators of PM_{10} emissions; land use (sum of area, m^2), road density (sum length, m) and population (sum number of people) within a 300 m buffer across the three site classifications.

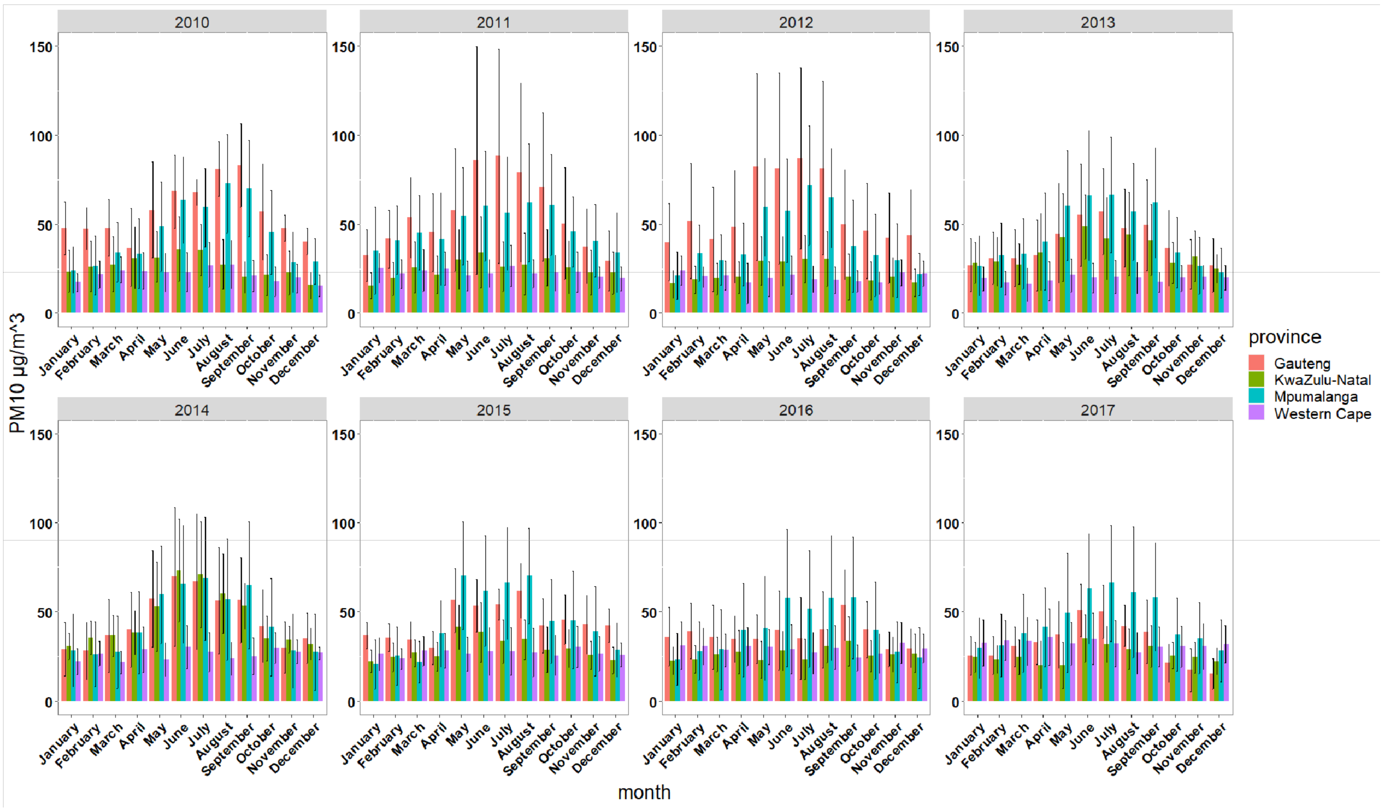


Figure 3. Variations in mean PM_{10} µg/m^3 concentration across the provinces and months of the year. Error bars represent one standard deviation of the mean.

3.2. Annual Change in Site's Average PM_{10} µg/m^3 Concentration over the Study Period

Table 1 shows how the levels of PM_{10} change across the sites for the years 2010–2017. In Gauteng province, the average change in annual PM_{10} concentration decreased in 5 of 8 sites (63%) (Table 1). In Mpumalanga province, the average change in annual PM_{10} concentration decreased in 9 of 12 sites (75%). The average change in annual PM_{10} concentration decreased in only 3 of 7 (43%) Western Cape Province sites. Similarly, a decrease in the average change in annual PM_{10} concentration was observed in only 2 of 5 (40%) KwaZulu-Natal province sites.

Table 1. Levels and changes in annual PM_{10} µg/m^3 concentrations across sites for the years 2010–2017. The first entry per site shows the annual PM_{10} concentration in µg/m^3. Subsequent entries depict the percentage changes compared to the previous entry. The last column shows the annual changes in µg/m^3 per year assuming a linear trend between the first and last available measurement per site.

Province	Site	2010	2011	2012	2013	2014	2015	2016	2017	Annual Change in PM_{10}
					Percentage increase/decrease					
Gauteng	Bodibeng			57.64	+2.54					+1.46
	Booysen			57.88		+7.42				+4.30
	Ekandustria				30.40	+70.51				+21.4
	Elandsfontein							29.45	− 1.64	− 0.48
	Leandra		22.14	− 28.07						− 6.22
	Orange Farm	57.25							− 5.42	− 3.10
	Randwater				47.19	− 4.40	− 0.67	+1.15	− 27.53	− 2.85
	Rosslyn			20.08	− 0.75	− 0.71				− 0.15

Table 1. *Cont.*

Province	Site	2010	2011	2012	2013	2014	2015	2016	2017	Annual Change in PM_{10}
Mpumalanga	Camden				53.95		−20.50		−3.41	−6.07
	Club			29.18	+14.16	+26.17		−6.38	−21.11	+0.02
	Embalenhle			37.08		+31.35			−4.38	+2.62
	Ermelo	40.50	+0.01	+15.93	−2.11	−37.80	+52.56	−10.68		−0.59
	Grootvlei		33.29		+2.36					+0.79
	Hendrina	38.23	−6.02	+11.23			−4.80	−29.07		−1.89
	Komati		64.56	+0.15		−7.33	+16.62	−15.79	+29.65	−0.06
	Middleburg	39.80	−20.34	+3.71	−19.66	−41.27	+50.82	−40.36		−3.84
	Phola				74.89	+1.90		−4.89	−6.57	−2.86
	Secunda	61.95	+1.09	−68.07	+122.87					−8.73
	Verykkop				24.39			−2.97	−21.07	−1.49
	Witbank	44.47			−1.12	−38.39	+119.33	−8.46		+1.58
Western Cape	Beliville			21.88	+10.84		+7.03	+0.05	+11.88	+1.65
	Foreshore		21.30	+3.40	−7.65		+15.22	−6.39	+15.28	+1.21
	George	21.95			−7.65					−1.68
	Goodwood		28.01	−8.44		+1.21	+12.64	+6.94		+0.84
	Stellenbosch			16.72	−0.09					−0.02
	Tableview		19.76	−8.11	−1.19					−0.91
	Wallacedene			16.91			+16.30	+31.45	+12.38	+4.05
KwaZulu-Natal	Brackenham		26.41	+13.66			+0.33	−6.90	+10.64	+0.45
	CBD		23.01	+13.91	+0.33		+5.00	−16.35	+6.32	+0.24
	Esikhaweni							27.22	−20.64	−5.62
	Gangles	34.61	+12.97		+7.07	+3.59				+2.88
	Ferndale	16.14	−19.44	−9.17						−2.16

Legend

1st annual PM_{10} average

Percentage decrease in annual PM_{10}

Percentage increase in annual PM_{10}

No Data

3.3. Monthly Differences in Daily PM_{10}

A summary of monthly mean PM_{10} concentrations across all sites per province and across the years 2010–2017 is presented in Figure 3. The pattern in the levels of PM_{10} across the four provinces (Gauteng, KwaZulu-Natal, Mpumalanga and Western Cape) of South Africa suggests seasonal variation in monthly PM_{10} levels. The monthly PM_{10} levels show a seasonal pattern across the provinces and are more prominent in Gauteng and Mpumalanga provinces. The monthly mean PM_{10} levels were highest in Gauteng Province and lowest in Western Cape Province. In Gauteng, the lowest monthly mean PM_{10} concentrations were recorded during the summer months (December–February), ranging from a monthly mean of 15.51 μg/m^3 recorded in December 2017 to 51.92 μg/m^3 recorded in February 2012. The monthly mean PM_{10} peaked during the winter months, ranging from 35.29 μg/m^3 recorded in July 2016 to 88.46 μg/m^3 recorded in July 2011. The highest mean PM_{10} recorded during the winter months is about 5.7 times higher than the revised 2021 WHO annual PM_{10} air quality guideline of 15 μg/m^3. In Western Cape Province, the lowest monthly mean during the summer months ranged from 15.53 μg/m^3 recorded in December 2010 to 34.17 μg/m^3 in February 2017. The monthly mean PM_{10} during the winter months ranged from 18.69 μg/m^3 recorded in August 2012 to 34.98 μg/m^3 recorded in June 2017. In general, all provinces recorded peak PM_{10} levels during the winter months in South Africa between June and August.

3.4. Week Day Differences in Daily PM_{10}

Figure 4 summarizes daily mean PM_{10} per province for the eight-year study period. Generally, marginal differences were found between different days of the week between 2010 and 2017. Average daily PM_{10} concentrations during the weekdays are slightly higher than during weekends in all four provinces. The highest PM_{10} concentrations of 51.4 µg/m^3, 46.8 µg/m^3, 29.1 µg/m^3 and 25.1 µg/m^3 at Gauteng, Mpumalanga, KwaZulu-Natal and Western Cape Province were recorded during the weekdays. Statistically significant differences in mean PM_{10} concentrations were observed between weekdays and weekends (F = 14.57 and value = 0.0009) and by province (F = 380.11 and *p* value =< 0.0001). The Pairwise Tukey's test comparisons suggest the difference between weekdays and weekends mean PM_{10} concentrations was statistically significant in all pairs of provinces but between KwaZulu-Natal and Western Cape (*p* value = 0.14).

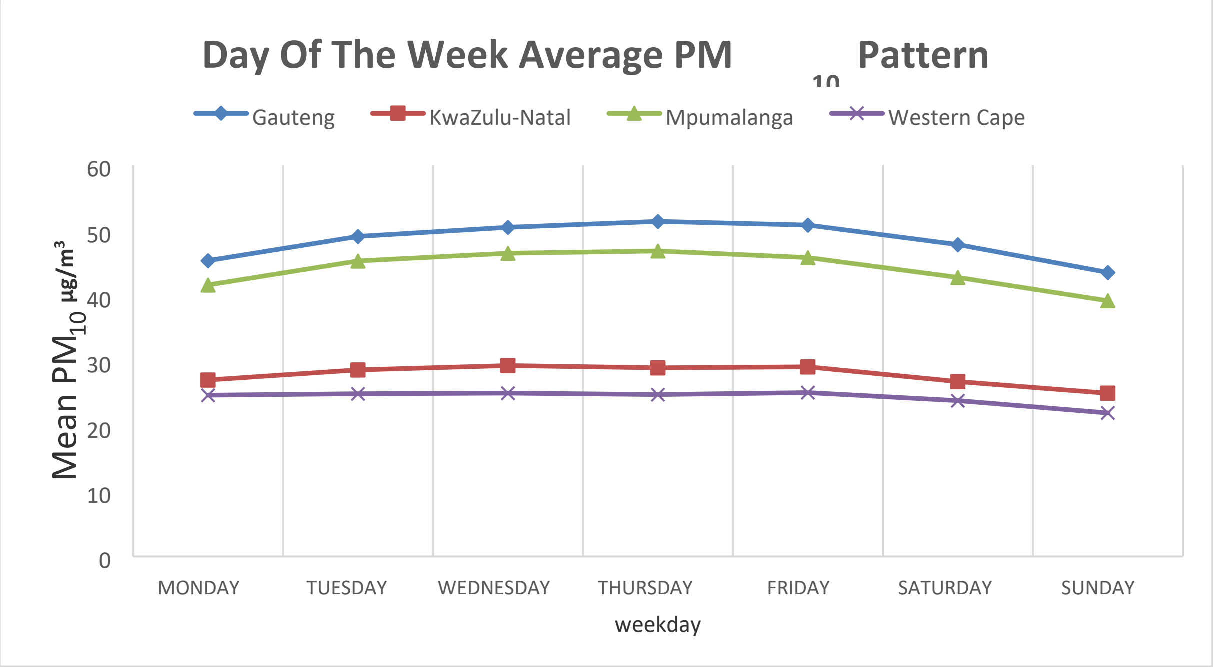


Figure 4. Weekdays variation in average daily PM_{10} µg/m^3 concentration across the provinces.

3.5. Spatial Variation in PM_{10}

A summary of descriptive statistics is presented in Table 2. There are 20 industrial sites, 18 residential sites and 5 traffic sites included in this analysis. These traffic sites are located in the three provinces of Gauteng (1 site), Western Cape (2 sites), KwaZulu-Natal (2 sites). The results from Table 2 show that the levels of PM_{10} concentration level is highest in Gauteng and for all provinces PM_{10} concentration is highest at the residential sites compared to industrial and traffic sites. In Gauteng, the concentration at one traffic site was similar to the concentration at the residential sites and substantially higher than the levels at the industrial sites.

Table 2. The distribution of daily PM_{10} concentration in µg/m^3 by province and site type.

	Site Classifications											
	Industrial				Residential				Traffic			
Province	N	Median	25–75% percentile	Min–Max	N	Median	25–75% percentile	Min-Max	N	Median	25–75% percentile	Min–Max
Gauteng	5114	29	17.5–43.9	7.03–139	4020	58.5	41.2–144	20.9–344	731	53.9	38.8–74.6	23.5–152
Western Cape	4750	22.3	16.5–29.7	9.68–82.5	2193	25.4	18.4–34.8	10.9–93.4	2922	21.1	16.4–27.7	10.7–74.0
Mpumalanga	18264	37.1	22.3–59.2	8.78–228	2921	37.7	21.9–61.4	9.1–216	NA	NA	NA	NA
KwaZulu-Natal	NA	NA	NA	NA	5114	26	16.5–37.2	7.27–130	23.6	23.6	18.3–32.2	13.1–79.9

NA: Not available.

The levels of PM_{10} concentration in Mpumalanga province are also high (Table 2). The industrial sites in Mpumalanga recorded the highest levels of PM_{10} compared to the industrial sites in other provinces. The PM_{10} concentration levels at both industrial and residential sites in Mpumalanga are comparable. The levels of PM_{10} concentration in Western Cape are the lowest compared to other provinces. Residential sites recorded the highest level of PM_{10} concentration in Western Cape. However, there are no substantial differences in the levels of PM_{10} concentration across the site types. Similarly, sites from KwaZulu-Natal province also recorded relatively low levels of PM_{10} concentration compared to Gauteng and Mpumalanga provinces. The levels at residential sites are also marginally higher than the levels at traffic sites (Table 2).

3.6. Attainment of PM_{10} Standards

Table 3 shows the percentage of days that PM_{10} concentration daily limits were exceeded using WHO and South Africa's National Air Quality Standard (NAAQS) for the sites in the four provinces from 2010–2017. Gauteng province reported the highest proportion of days exceeding the daily limits of WHO and NAAQS standards, with about 38% of the days exceeding the WHO daily standard and around 17% of days exceeding the NAAQS daily standard. In contrast, Western Cape Province reported the lowest percentage of days exceeding both WHO and NAAQS PM_{10} air quality standards (3% and 0.09%, respectively) between years 2010–2017 (Table 3).

Table 3. The percentage of PM_{10} (μg/m^3) concentration exceeding daily standards by province for the years 2010–2017.

Province	WHO Standard		NAAQS Standard	
	Number of Days Exceeding Daily Limit	% of Days Exceeding Daily Limit [a]	Number of Days Exceeding Daily Limit	% of Days Exceeding Daily Limit [a]
Gauteng	3820/9865	38.7	1605/9865	16.3
Mpumalanga	7139/21185	33.7	3104/21185	14.7
KwaZulu-Natal	549/7671	7.2	108/7671	1.4
Western Cape	272/9865	2.8	8/9865	0.1

[a] The percentage of days PM_{10} concentration daily limits were exceeded based on the number of days with PM_{10} data for 2010–2017 divided by the total number of days with valid PM_{10} data. WHO standard; World Health Organization 2021 daily standard of 45 μg/m^3. NAAQS; South Africa's National Air Quality 2006 daily standard of 75 μg/m^3.

4. Discussion

This study adds to existing evidence on levels of PM_{10} in South Africa. The eight-year trend in PM_{10} level suggest that PM_{10} is still high in the earliest high pollution designated priority areas around Gauteng and Mpumalanga provinces. However, there is evidence of decreasing PM_{10} levels at most sites in both Gauteng and Mpumalanga Provinces. While the level of PM_{10} of most sites in KwaZulu-Natal and Western Cape Provinces suggest an increase in PM_{10} levels during the study period. The presented analysis identified trends in ambient PM_{10} concentrations in four South African Provinces for the years 2010–2017.

4.1. Spatial and Temporal Trends in Daily PM_{10}

The Vaal Triangle Airshed Priority Area (VTAPA) and Highveld Priority Area around Gauteng and Mpumalanga Provinces were the first areas designated as air pollution priority areas in South Africa due to the observed or expected level of air pollution in these areas [2,10]. Both provinces share similar emissions profiles; they are home to the majority of coal-powered plants, coal mining, gold mining, mine tailing, petrochemical and ferroalloy industries in South Africa. To address the level of air pollution in these areas, air quality management plans were developed to guide actions towards improving the air quality in these priority areas. Some of the actions implemented to reduce the air pollution in these areas include the closure of a high polluting industry, large-scale domestic electrification program in these areas to reduce domestic emissions [2,10]. Thus, the decrease in the levels of PM_{10} reported in this study at most sites around these areas in Mpumalanga and Gauteng Provinces could be because of the changing emission profiles in these priority areas due to these mitigation actions. Despite the lower levels of PM_{10} reported in KwaZulu-Natal and Western Cape Provinces compared to Gauteng and Mpumalanga

Provinces, the average change in annual PM_{10} increased at most sites in these Provinces. This trend signals a deteriorating air quality in these areas that is likely due to changes in emissions profiles in these areas. The Southern Basin Industrial areas in KwaZulu-Natal have been identified as air pollution hotspots due to the high density of industrial activities in this area [18]. There are also concerns about the air quality in Western Cape provinces, especially around the increasing informal settings in Western Cape Province [19].

The monthly variation in PM_{10} across the provinces during the study period shows that PM_{10} concentrations are highest during the winter months between June and August. This is consistent with results of a study in Gauteng assessing the characteristics of ground-monitored $PM_{2.5}$ and PM_{10} between years 2010 and 2014 [13] and a study conducted in eMbalenhle—a low socio-economic in Mpumalanga province [15]. Similarly, the marginal seasonal difference in PM_{10} reported in Western Cape Province follows the pattern reported in a Western Cape study that reported the seasonal difference in PM_{10} in 2016 using data from one monitoring site [16]. A Ugandan study, however, reported higher PM_{10} concentration during the dry seasons compared to wet seasons [14]. The difference in seasonal weather patterns and sources of PM_{10} between South Africa and Ugandan could explain the seasonal difference in PM_{10} concentration. The cold season in most of South Africa’s provinces is characterized by cold weather and an increase in solid biomass use as a source of energy. The reliance of South African’s on solid biomass as a source of energy for cooking and heating system during the winter has been reported in other studies [2,20–22]. Residential fuel consumption in South Africa includes kerosene, residential fuel oil, LPG, sub-bituminous coal, wood/wood waste, other primary solid biomass and charcoal. Overall, residential fuel consumption dropped from 2010 to 2017, but domestic coal consumption increased slightly [22]. This study also highlights the fact that domestic sources of PM_{10} contribute substantially to the variability of PM_{10} in South Africa.

The trend in weekday PM_{10} level follows a similar pattern across the four provinces. PM_{10} concentration increased through the weekdays, reaching its peak between Wednesday and Friday. This study suggests that there is a difference between PM_{10} levels between weekends and weekdays, with lower PM_{10} levels reported during the weekends. Although there are only four traffic sites in this analysis, the decreased level in traffic-related activities during the weekend might be responsible for the observed lower PM_{10} levels during the weekend.

4.2. PM_{10} Level across Site Types

The trends in PM_{10} across the primary environment types in South Africa used for classifying the PM_{10} monitoring sites by the South Africa Weather Service are the Industrial, Residential and Traffic areas. The majority of the monitoring sites included in this analysis are industrial and residential sites. Table 2 shows that average PM_{10} levels were highest in residential sites compared to other categories of sites during the study period in all four provinces. This result is not unusual; similar results were reported in Gauteng areas of South Africa [13]. Our result also show that PM_{10} levels are generally higher at residential sites in the other three provinces. A possible explanation for the high levels of PM_{10} concentration levels at residential areas across the provinces is the high level of domestic burning in residential areas in South Africa. Previous studies have highlighted domestic emissions as the predominant source of particulate matter in South Africa [2,13]. It has also been argued that because the industrial emissions are released into a stable atmosphere in stacks above the generally shallow boundary layer height in South Africa could have affected the dispersion of the emissions to the ground level [2,13]. We also explored the variation in residential and industrial land use and road and population density around the monitoring sites. The residential radii have the highest level of variation from multiple influencing factors of PM_{10} emission. The high variability in the multiple sources of PM_{10} emission suggests that the high density of PM_{10} emissions around residential areas could explain the highest concentration of PM_{10} recorded in residential sites in our analysis. The high level of PM_{10} concentration and high variability of potential PM_{10} influencing factors around residential areas have implications on the population's health outcomes [23].

4.3. Strengths and Limitations

There are some limitations in this study worth nothing. First, the pattern of missingness of PM_{10} exposure data during the study period poses a challenge to understanding the time-series trend in PM_{10} exposure data across the sites. The results presented are for four out of nine provinces in South Africa. Thus, these results cannot be extrapolated beyond the provinces that contributed data to our analysis. In addition, the representativeness of the site types is also a limitation of this study; the majority of the sites included in this study are industrial and residential sites. To address the challenge of missing daily PM_{10} data, this study combined observed and imputed PM_{10} exposure data from 44 monitoring sites across four provinces in South Africa to investigate the trends in PM_{10} concentrations. Despite the limitations, our results provide some insights on trends of PM_{10} concentrations in the four provinces during the study period.

5. Conclusions

It has been over a decade since the promulgation of South Africa's National Environmental Management Air Quality Act in 2004. There have been concerns over the progress made so far [11]. We found that PM_{10} levels are higher than the WHO limits standard across the four provinces. The provincial differences in PM_{10} concentration show that PM_{10} levels are higher around air pollution priority areas, while the temporal variability of PM_{10} suggest that emissions during the winter months contribute markedly to the high level of PM_{10} recorded during the winter seasons.

An interesting result for future epidemiological studies in South Africa is the high level of PM_{10} and high variability of potential influencing factors of PM_{10} emission around where people live and work. Taken together, these results have implications for addressing the trends of PM_{10} pollution in South Africa.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ijerph182413348/s1, Table S1: The pattern of PM_{10} data availability across the sites for years 2010–2017.

Author Contributions: Conceptualization, O.O.A., K.d.H. and M.R.; methodology, O.O.A. and K.d.H.; formal analysis, O.O.A.; data curation, O.O.A. and K.d.H. writing—original draft preparation, O.O.A. and K.d.H.; writing—review and editing, O.O.A., K.d.H., M.R., M.A.D. and T.C.A.-O.; supervision, K.d.H., M.R. and M.A.D.; funding acquisition, M.R., M.A.D. and K.d.H. All authors have read and agreed to the published version of the manuscript.

Funding: 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.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: We would like to thank the staff of the different monitoring stations and the staff at South Weather Services who provided data for this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Health Effects Institute. *The State of Global Air*; Health Effects Institute: Boston, MA, USA, 2020.
2. Pretorius, I.; Piketh, S.; Burger, R.; Neomagus, H. A perspective on South African coal fired power station emissions. *J. Energy S. Afr.* **2015**, *26*, 27–40. [CrossRef]
3. Altieri, K.E.; Keen, S.L. Public health benefits of reducing exposure to ambient fine particulate matter in South Africa. *Sci. Total. Environ.* **2019**, *684*, 610–620. [CrossRef] [PubMed]
4. GBD 2019 Risk Factors Collaborators. 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* **2019**, *396*, 1223–1249.
5. Khumalo, T.N. *2017 State of Air Report and National Air Quality Indicator*; Department of Environmental Affairs: Pretoria, South Africa, 2017.
6. Department of Environmental Affairs. *2nd South Africa Environment Outlook*; Department of Environmental Affairs: Pretoria, South Africa, 2016.
7. World Health Organization. *WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*; World Health Organization: Geneva, Switzerland, 2021.
8. Mkoma, S.L.; Mjemah, I.C. Influence of meteorology on ambient air quality in Morogoro, Tanzania. *Int. J. Environ. Sci.* **2011**, *1*, 1107.
9. Czernecki, B.; Półrolniczak, M.; Kolendowicz, L.; Marosz, M.; Kendzierski, S.; Pilguj, N. Influence of the atmospheric conditions on PM10 concentrations in Poznan´, Poland. *J. Atmos. Chem.* **2017**, *74*, 115–139. [CrossRef]
10. Feig, G.; Garland, R.M.; Naidoo, S.; Maluleke, A.; Van Der Merwe, M. Assessment of changes in concentrations of selected criteria pollutants in the Vaal and Highveld Priority Areas. *Clean Air J.* **2019**, *29*. [CrossRef]
11. Tshehla, C.; Wright, C.Y. 15 Years after the National Environmental Management Air Quality Act: Is legislation failing to reduce air pollution in South Africa? *S. Afr. J. Sci.* **2019**, *115*, 1–4. [CrossRef]
12. Feig, G.; Naidoo, S.; Ncgukana, N. Assessment of ambient air pollution in the Waterberg Priority Area 2012-2015. *Clean Air J.* **2016**, *26*, 21–28. [CrossRef]
13. Hersey, S.P.; Garland, R.M.; Crosbie, E.; Shingler, T.; Sorooshian, A.; Piketh, S.; Burger, R. An overview of regional and local characteristics of aerosols in South Africa using satellite, ground, and modeling data. *Atmos. Chem. Phys.* **2015**, *15*, 4259–4278. [CrossRef] [PubMed]
14. Onyango, S.; Parks, B.; Anguma, S.; Meng, Q. Spatio-Temporal Variation in the Concentration of Inhalable Particulate Matter (PM10) in Uganda. *Int. J. Environ. Res. Public Health* **2019**, *16*, 1752. [CrossRef] [PubMed]
15. Lina, N.D.; Engelbrecht, J.C.; Wright, C.Y.; Oosthuizen, M.A.; Thabethe, N.D.L. Human health risks posed by exposure to PM10 for four life stages in a low socio-economic community in South Africa. *Pan Afr. Med. J.* **2014**, *18*, 206. [CrossRef] [PubMed]
16. Olaniyan, T.; Dalvie, M.A.; Röösli, M.; Naidoo, R.N.; Künzli, N.; de Hoogh, K.; Berman, D.; Parker, B.; Leaner, J.; Jeebhay, M.F. Short term seasonal effects of airborne fungal spores on lung function in a panel study of schoolchildren residing in informal settlements of the Western Cape of South Africa. *Environ. Pollut.* **2020**, *260*, 114023. [CrossRef] [PubMed]
17. Arowosegbe, O.; Röösli, M.; Künzli, N.; Saucy, A.; Adebayo-Ojo, T.; Jeebhay, M.; Dalvie, M.; de Hoogh, K. Comparing Methods to Impute Missing Daily Ground-Level PM_{10} Concentrations between 2010–2017 in South Africa. *Int. J. Environ. Res. Public Health* **2021**, *18*, 3374. [CrossRef] [PubMed]
18. Department of Environmental Affairs. South Durban Basin Multi-Point Plan Case Study Report. In *Air Quality Act Implementation: Air Quality Management Planning*; Department of Environmental Affairs: Pretoria, South Africa, 2007.
19. Department of Environmental Affairs. *Status of Air Quality in South Africa & Roadmap for Asbestos Waste Disposal: Department of Environmental Affairs Briefing*; Department of Environmental Affairs: Pretoria, South Africa, 2019.
20. Friedl, A.; Holm, D.; John, J.; Kornelius, G.; Pauw, C.J.; Oosthuizen, R.; van Niekerk, A.S. Air pollution in dense, low-income settlements in South Africa. In Proceedings of the National Association for Clean Air (NACA), Mpumalanga, South Africa, 1–3 October 2008.
21. Scorgie, Y.; Burger, L.; Annegarn, H. Socio-economic Impact of Air Pollution Reduction Measures–Task 2: Establishment of Source Inventories, and Task 3: Identification and Prioritisation of Technology Options. In *Report Compiled on Behalf of NEDLAC*; National Economic Development and Labour Council: Pretoria, South Africa, 2003; Volume 25.
22. Department of Environmental Affairs. *Greenhouse Emissions National Inventory Report 2017*; Department of Environmental Affairs: Pretoria, South Africa, 2017.
23. Pope, C., 3rd. Epidemiology of fine particulate air pollution and human health: Biologic mechanisms and who's at risk? *Environ. Health Perspect.* **2000**, *108*, 713–723. [CrossRef] [PubMed]

8 paper 3: 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.

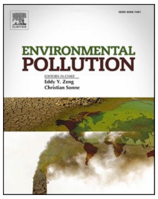
e Centre for Environmental and Occupational Health Research, School of
Public Health and 12 Family Medicine, University of Cape Town, Cape Town, South Africa.
* Correspondence: c.dehoogh@swisstph.ch

This paper was published in Environmental Pollution, 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[1]

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

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 epidemiological transition of

[1] 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

diseases from communicable diseases to non-communicable diseases in sub-Saharan African (SSA) countries (Adebayo-Ojo et al., 2022; Gouda et al., 2019; Kone 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 μg/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 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.

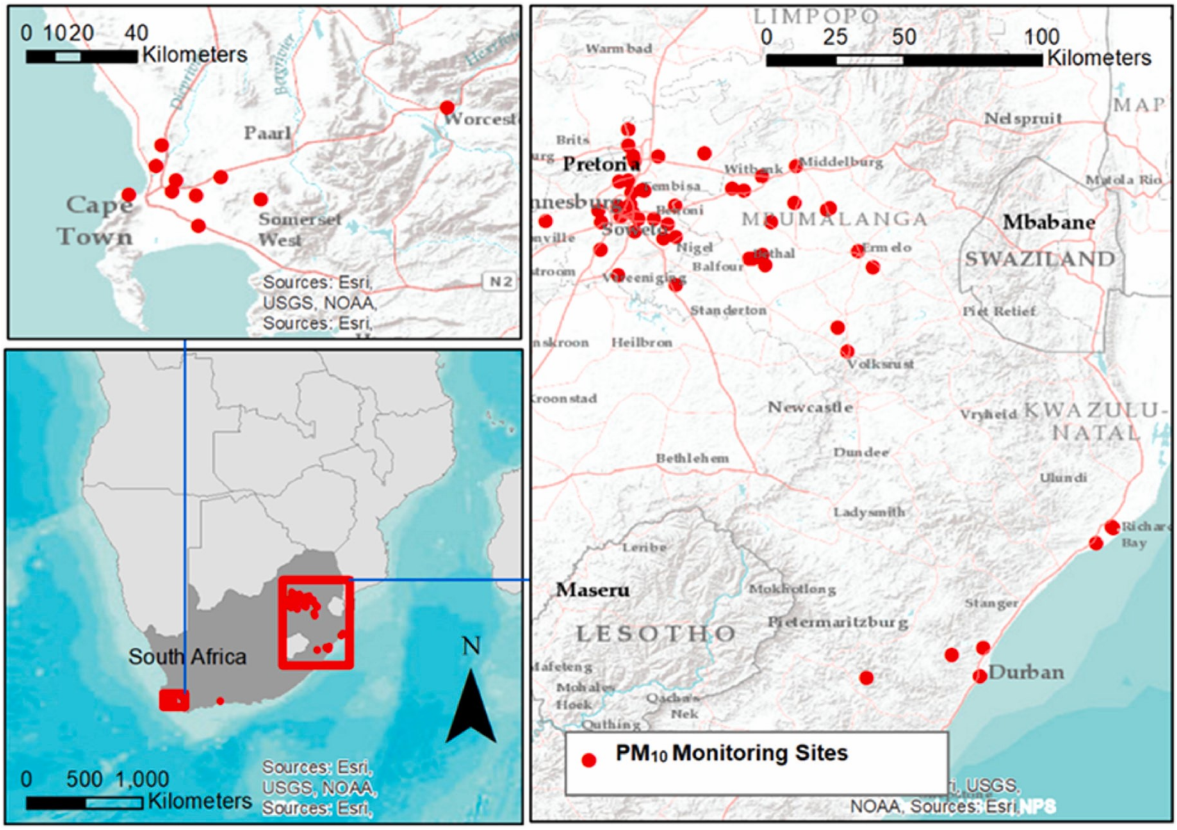


Fig. 1. The spatial distribution of PM$_{10}$ Monitoring Sites.

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/produ cts/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.

2.3.1. Spatial and temporal predictors

Meteorological variables play an important role in the dispersion of air pollutants (De Visscher, 2013; Lana 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 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.

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)).

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:

$$PredMAIAC.AOD_{i,\ t} = MAIAC.AOD_{i,\ t}$$

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

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:

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

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*. SPT1-10$_{i,t}$ 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:

$$PredPM_{10i,t} = PM_{10i,t}\text{⍰}\, f(RF_{predPM10i,t} + XGBoost_{predPM10i,t} + SVR_{predPM10i,t} + latitude_i + longitude_i) \quad (3)$$

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

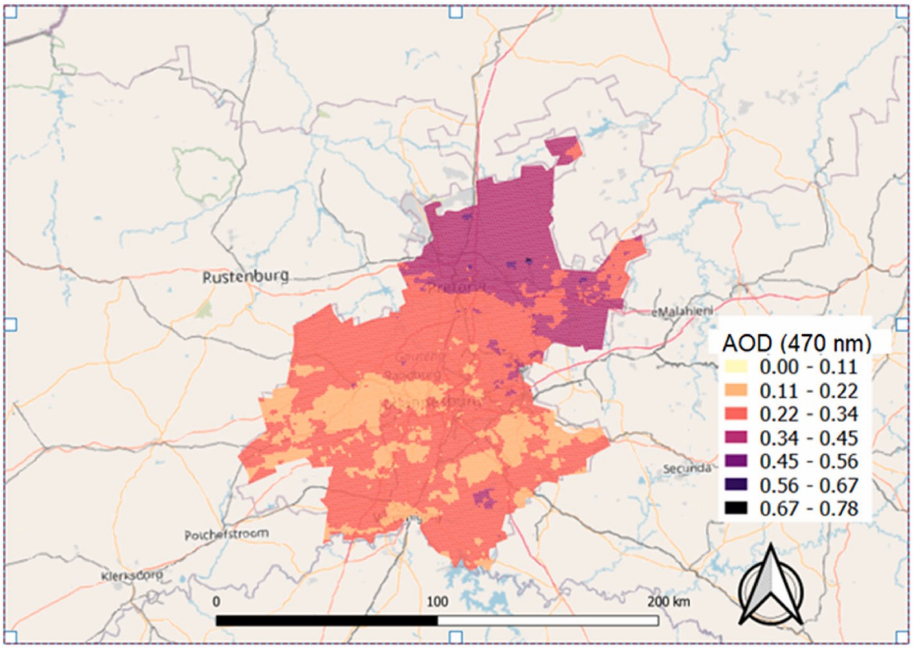


Fig. 2. Gauteng prediction map of AOD 470 nm for June 6, 2016.

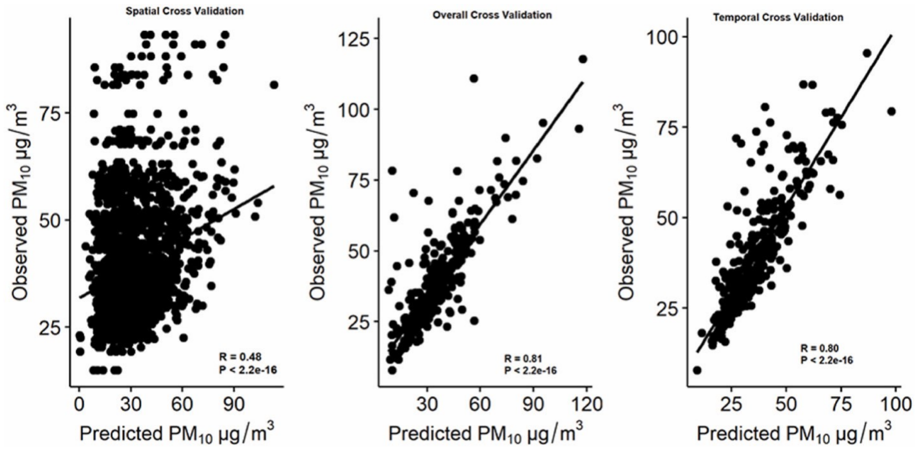


Fig. 3. Scatter plots between predicted and observed PM$_{10}$ concentrations of the spatial, temporal and overall cross validation of the ensemble model.

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 µg/m^3), spatial and temporal R^2 and RMSE are reported for the ensemble averaged model.

Model	CV	R2	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.

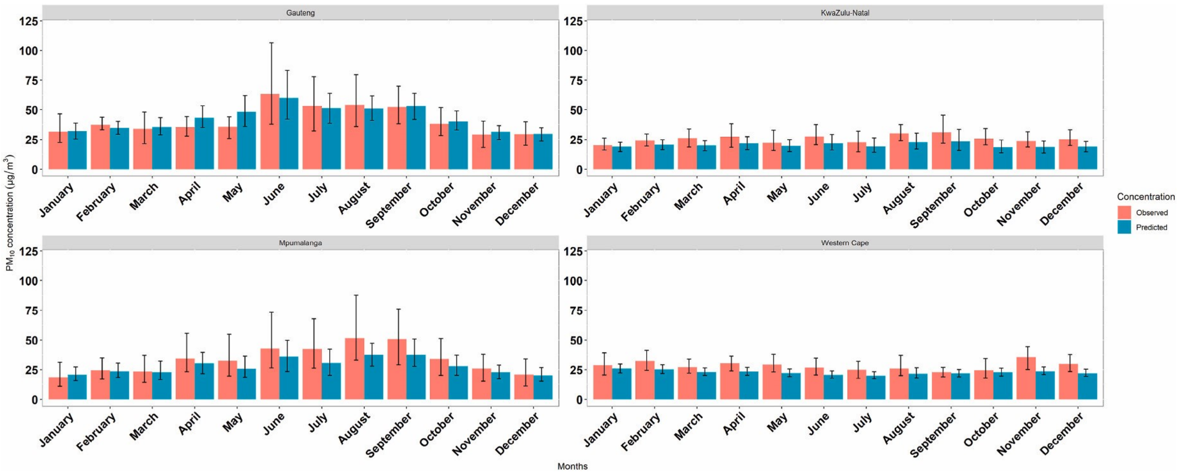


Fig. 4. Monthly observed versus predicted PM$_{10}$ means. Error bars represent standard deviation of the mean.

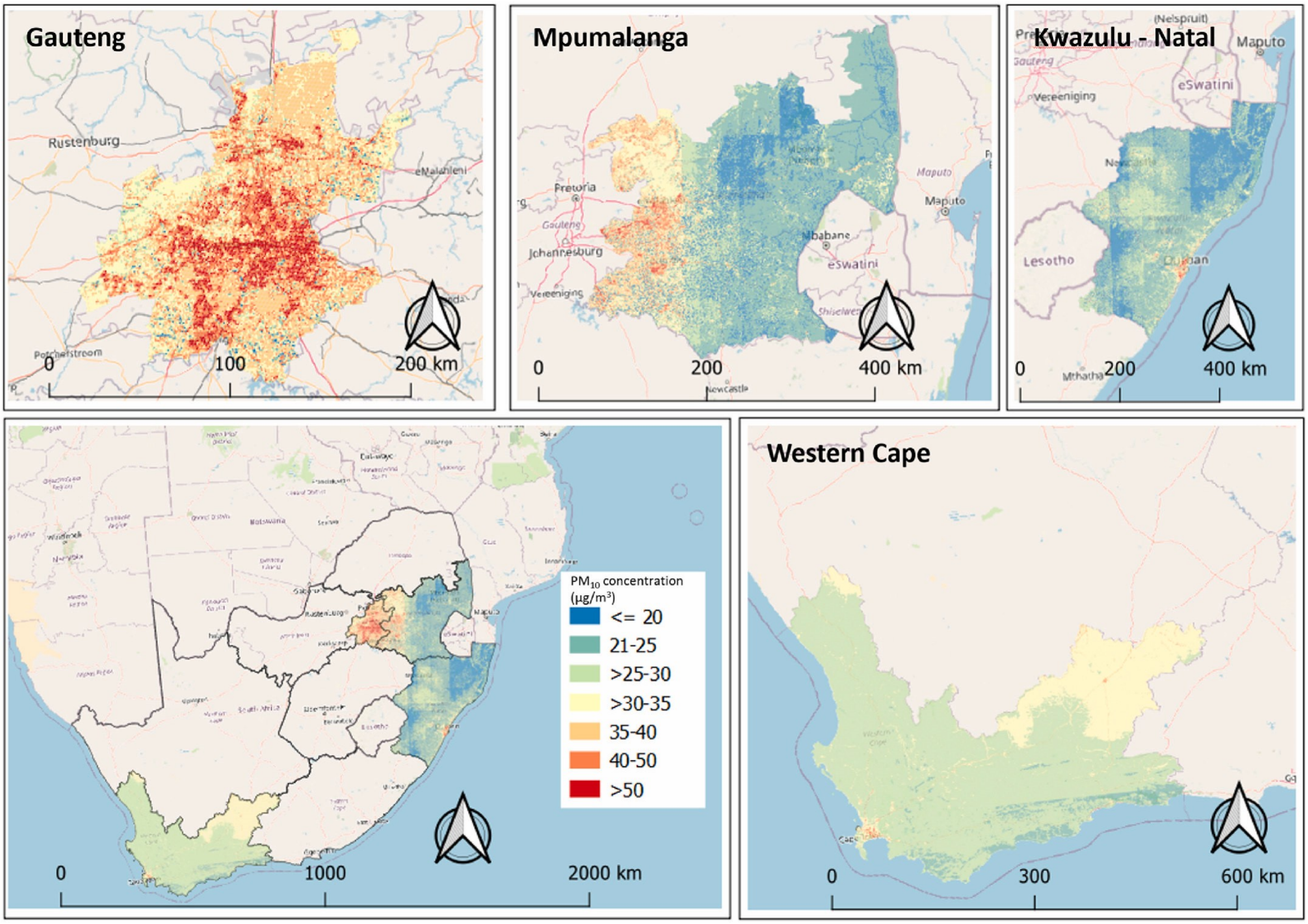


Fig. 5. Annual mean PM_{10} concentrations (µg/m^3) for 2016 at 1 km × 1 km grid cells aggregated from daily estimates.

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

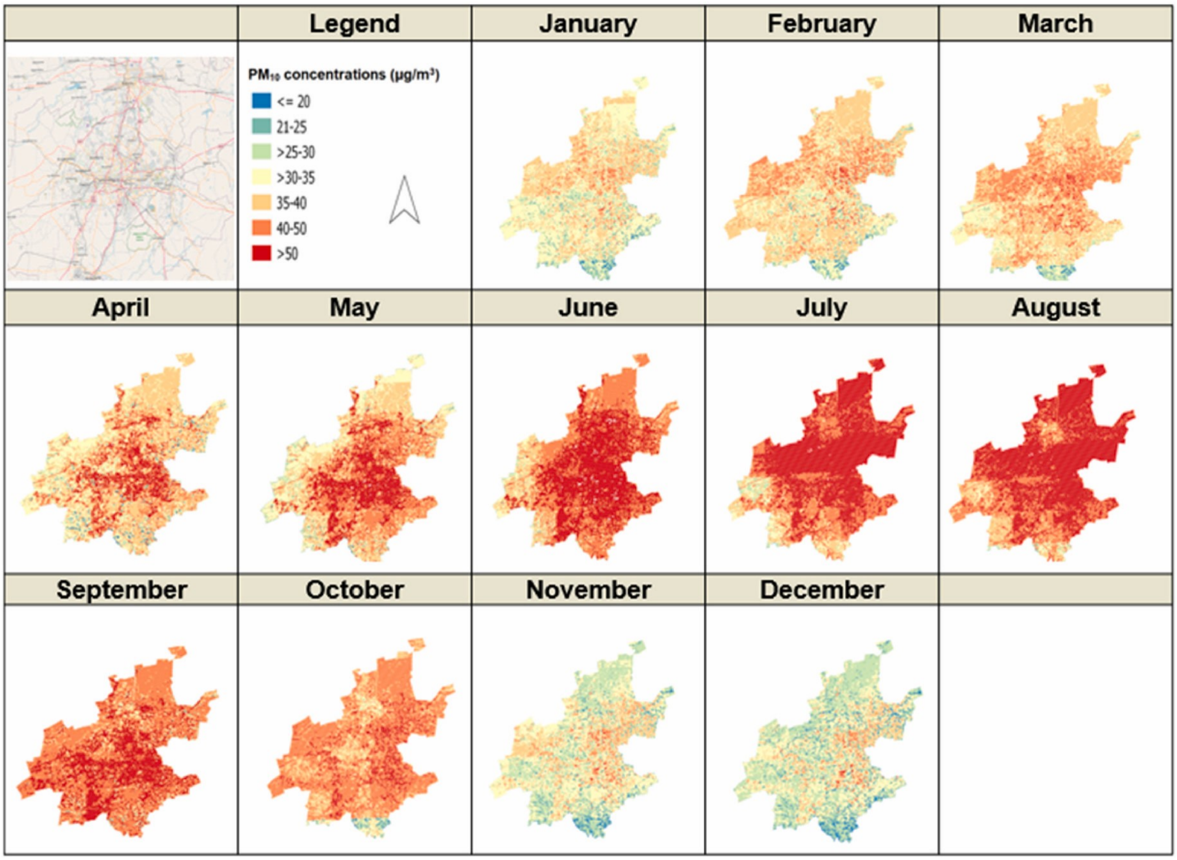


Fig. 6. Gauteng Province estimated monthly mean PM$_{10}$ concentrations (µg/m^3) for 2016 at 1 km × 1 km grid cells aggregated from daily estimates.

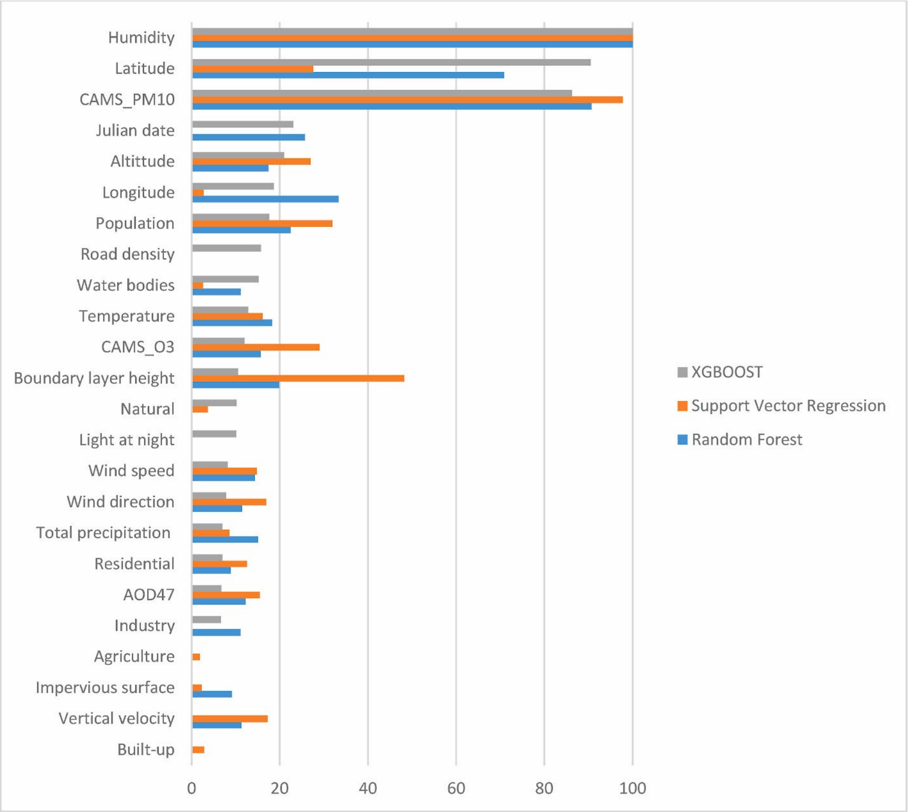


Fig. 7. Relative importance (scale from 0 to 100%) of the top 20 predictors from the individual models in Stage 2.

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 μg/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 Ro osli** : 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., Ro osli, 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., Ro osli, 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., Dedelė, 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., Horanyi, 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.

Kone, 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.

Lana, 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.

9 Discussion

This thesis presents new approaches to fill the gap in ground-monitored PM_{10} in South Africa. In tandem with the overall aim of this thesis, three individual studies presented the approach and results of the PM_{10} exposure assessment in South Africa. This chapter will discuss the challenges of air pollution exposure in South Africa and offer some suggestions on how to improve air quality monitoring in South Africa.

9.1 The state of Air Pollution Measurement.

The availability of good and reliable ground-monitored PM_{10} is central to these thesis objectives. The National Environmental Management Air Quality Act (AQA) of 2004 introduced air quality monitoring as a tool for effective air pollution management in South Africa (Government of South Africa 2005). The introduction of this Act also gave birth to ambient air quality standards in South Africa set in 2009 where PM_{10} was identified as a criteria pollutant of interest (South African Department of Environmental Affairs (DEA) 2009). $PM_{2.5}$ only became a priority air pollutant in South Africa in 2012 (The Law Library of Congress 2018). The revised 2015 National Air Quality standards for PM_{10} are a 75 µg/m^3 daily limit and an annual limit of 45 µg/m^3 (Department of Environmental Affairs 2016). An important feature of the 2004 AQA is the framework for the management of air quality management. The management of air pollution was devolved to local municipalities while the national Department of Environmental Affairs provides legislation. In addition, areas with recorded bad air quality and areas prone to increased air pollution levels were designated as air pollution priority areas (Government of South Africa 2005). This was done to mobilize available limited resources to areas where they are needed most. The prioritization of air quality management resources has an impact on the distribution of air pollution monitoring sites across South Africa. To date, there are four air pollution priority areas in South Africa; the Highveld, the Vaal Triangle, the South Durban Basin and Waterberg located in four different provinces (Gauteng, Mpumalanga, Western Cape and KwaZulu-Natal) of South Africa (Department of Environmental Affairs 2016, Feig, Naidoo et al. 2016, Feig, Garland et al. 2019). The density of air pollution monitoring sites in these areas informed our decision to focus on the four provinces. The first two studies (studies I and II) of this thesis presented the availability and completeness of PM_{10} at the sites distributed across these provinces. The huge gap in the continuous availability and completeness of PM_{10} was a trend that is consistent across sites from the four provinces (Arowosegbe, Röösli et al. 2021). This trend coupled with the limited number of monitoring sites may affect the ability of the spatiotemporal model ability in study III to capture variability in PM_{10} and the generalizability of model estimates to areas contributing little or no ground-level data information. We, however, reduced these biases by focusing on the provinces with some ground-level information. Study I also explored methods to increase the daily availability of PM_{10} in sites with relatively better PM_{10} annual data completeness. Our decision to only provide a model for 2016 in study III was also informed by the availability and completeness of ground-level PM_{10} data.

9.2 Spatial predictors

Important groups of predictor variables included in studies I-III to characterize PM_{10} are spatial predictors such as road traffic, land use, population, and impervious surface data. Spatial variables are used as proxies for possible sources of air pollution emissions, especially in areas like South Africa where there are no emission data. These variables vary across different locations but their values in time are mostly unavailable and are assumed to be constant in time. We used Geography Information Systems (GIS) methods to resample these variables to 1 x 1 km grid cells across our study areas. The spatial variables values were subsequently used in our models to capture the spatial variability in PM_{10}. The ability of the spatial variables to capture spatial variability was further explored in studies III and I. The variable of importance ranking of the spatial variables suggests they performed relatively poor when compared with temporal variables in explaining the variation in PM_{10} across the study areas. The poor performance of the spatial variables may be due; i) they are not good proxies for air pollution emission sources which might be due to the spatial resolution of these variables. ii) spatial variables' values do not vary in time i.e their values are constant over time. Thus, the constant pattern of these variables adds little explanatory information when used to characterize air pollution exposure.

9.3 Spatiotemporal predictors

In contrast to spatial predictors, spatiotemporal predictors are time and location varying variables. In studies I and III we used spatiotemporal predictors to construct our models. The majority of the spatio-temporal predictors are freely available on open source repositories. Reanalysis meteorological and atmospheric composition variables of PM_{10}, NO_2 and O_3 were obtained to achieve continuous spatial and temporal completeness of their values for our analysis. In study III, CAMS AOD estimates of AOD were used to impute missing MAIAC AOD across our study area. The complete time series of AOD data were subsequently used to reconstruct PM_{10} concentration levels for the year 2016 across the four South Africa provinces. The variable of importance of studies I and III indicates that spatio-temporal predictors are relatively better in characterizing ground-level PM_{10} across South Africa. An important variable that emerged as a good predictor for explaining ground-level PM_{10} measurements is the CAMS ensemble estimate of PM_{10}. This suggests that the CAMS ensemble PM_{10} estimates have a good correlation with the ground-monitored PM_{10}. The completeness of the CAMS ensemble PM_{10} estimates makes it a good variable to be considered in hybrid models, especially in areas with sparse ground monitored air quality data.

9.4 AOD as a predictor of ground-monitored PM_{10}

The satellite product of AOD was conceptually considered a major predictor to characterize PM_{10} levels in South Africa because it is assumed that since AOD captures the columnar distribution of suspended particles; it would be associated with the distribution of ground-level particles. The availability of AOD at 1×1 km spatial and daily temporal resolution corresponds with the resolutions of target PM_{10} estimation spatial and temporal domain PM_{10} of study III. The result of study III however suggests that AOD did not contribute substantially to explaining the variation in ground-level PM_{10}. A poor association between AOD and ground-monitored PM have also been reported in other studies (Hersey, Garland et al. 2015, Alvarado, McVey et al. 2019). An important explanation for the poor predictive performance of AOD in South Africa despite relatively better data completeness of 68% for the year 2016 is the vulnerability of South Africa's vertical column to plumes from biomass burnings. The persistent transboundary transportation of these plumes at stable layers close to ground level between 3 – 5 km above South Africa might impact the ability of AOD to capture the characteristics of ground level PM (Hersey, Garland et al. 2015). This, however, suggest that the columnar properties of AOD should not be considered as the primary variable for explaining ground-level PM concentration in South Africa.

10 Methodological considerations for exposure assessment

The methodological approaches to reconstruct PM_{10} exposure for 2016 across the four provinces of South Africa were informed by the quality of ground-monitored PM_{10} and the spatial and temporal predictors available to construct these models. Importantly, we considered statistical models that can capture the complex underlying relationship between ground-monitored PM_{10} and the spatio-temporal predictors. In addition, the generalizability of the models across the four provinces was also an important factor that was considered.

First, as earlier emphasized, the distribution of the monitoring networks in South Africa is sparse and data completeness is a challenge. In study I, we tried to improve the completeness of the ground-monitored PM_{10} data. To this effect, we compared three models (National, Provincial and Site-specific) using Random Forest (RF) machine learning model to explore the statistical relationship between ground-monitored PM_{10} and spatial and temporal variables to increase daily PM_{10} data at some sites with at least 70% of annual PM_{10} data. The site-specific model predictions were closest to the observed PM_{10} (Arowosegbe, Röösli et al. 2021). The results of study I suggest both National and Provincial models are less generalizable possibly due to the limited number of monitoring sites to capture the variation in ground-level PM_{10} across the study areas. Study II shows PM_{10} levels vary across the four provinces, site classifications and land use categories (Arowosegbe, Röösli et al. 2021). Given the variability of PM_{10} across the four provinces and the level of data completeness of ground-level PM_{10} for the years 2010 – 2017 together, we decided to model PM_{10} exposure for only 2016. The year 2016 had the highest number of monitoring sites with the most complete PM_{10} daily data. In addition, we decided to use the ensembling of three machine-learning methods to achieve average PM_{10} predictions across the study areas. Machine learning methods are increasingly used for modelling environmental exposure because they do not assume the underlying relationship that exists between environmental exposure variables i.e. they learn the relationship from the data (Masih 2019). Extreme gradient boosting (XGBoost), RF and Support Vector Regression (SVR) were the individual learners selected to calibrate ground-level PM_{10} and a RF model was used to average the daily predictions from the individual models. For study III analysis, we combined auxiliary markers of air pollution emissions and distribution such as satellite-derived AOD, meteorological variables, road network, population, and ensemble estimates of PM_{10}, O_3 and NO_2 with daily ground-level PM_{10} concentration in an ensemble machine learning framework. The ensemble machine learning framework has three main advantage over individual models; (i) It improves the individual learners' prediction performance by averaging the predictions performance, (ii) The averaged predictions from the ensemble framework is more robust to overfitting and (iii) The averaged predictions are less bias compared to individual learners prediction (Zhang and Ma 2012). The applicability of exposure models for epidemiological studies depends on the reliability and generalizability of the models in space and time. Thus, study III models were cross-validated spatially and temporally. Because the ground-level PM_{10} data are not representative of the study areas, we cannot assume the ground-level PM_{10} used for our exposure modelling is independent and evenly distributed across the study areas. This informed the spatial leave location out validation method used to assess the performance in space. The spatial leave location out validation allowed us to leave out data points of some sites and predict their PM_{10} concentration from the remaining sites, not held-out. In conclusion, the ensemble machine learning framework improved the individual learners' performance and predictive capacity.

11 The implication of our exposure modelling results for similar studies in LMICs

This thesis was conceived to explore satellite information as an opportunity to bridge the gap in ground-monitored air quality data in South Africa. South Africa has one of the most extensive air quality monitoring networks in sub Saharan Africa (SSA) countries. There are officially more than 180 reference-monitoring stations in South Africa but as earlier alluded to the majority of these stations are not operational all year round because of insufficient technical capacity, vandalism and financial constraints. The challenges of air quality in South Africa have left a gap in air quality management and air pollution health impact studies. Study III, to the best of our knowledge, is the first study integrating satellite information and spatio-temporal predictors with groundmonitored air pollution data to fill the gaps in ground-monitored air quality measurements in SSA. The key results from our studies should be highlighted as we believe more studies from SSA will consider spatio-temporal models using satellite information to address the challenges of air quality measurement in the region.

Continuous air quality measurement is still a big challenge for air pollution exposure assessment. Study I detailed the gap in ground monitored PM_{10} data. This gap in ground-level monitored PM_{10} influenced the selection of the year 2016 for PM_{10} exposure modelling. This implies that reliable and consistent ground monitored data is necessary for similar spatio-temporal models in areas with sparse ground monitored data. This further compound the challenges to bridging the gap in PM_{10}. Nonetheless, the good predictive capacity of ensemble estimates of air pollutants is a piece of important information from our studies that future air pollution exposure assessment studies from sub-Saharan Africa countries should explore. The expected United States National Aeronautics and Space Administration (NASA) Multi-Angle Imager (MAIA) for Aerosol satellite instrument revolving around the earth at 740 kilometres is equipped with a specialized camera designed to differentiate different sizes of aerosols (particulate matter) based on how the particles reflect or absorb sunlight is an improvement on the satellite-derived columnar measurement of AOD (Diner, Boland et al. 2018). This could potentially resolve the problem of frequent dilution of South African vertical columns with plumes from burning emissions. In addition, the proposed calibration of the MAIA satellite instrument data with ground-monitored data and computer models will further enhance MAIA's product data for particulate matter exposure modelling in LMICs. However, the coverage of the under-development MAIA satellite is currently limited to a few selected primary and secondary target areas including two SSA cities of Johannesburg, Addis Abba designated as primary target areas and six SSA cities of Dakar, Accra, Lagos, Harar, Cape Town, and Nairobi selected as secondary target areas.

12 The case for low-cost monitoring sensors

The cost of establishing and maintaining reference-monitoring sites is one of the major challenges of air quality monitoring in LMICs including countries in SSA. Indeed, the estimated $100,000 per year to install and maintain a reference monitor might be responsible for just one ground-level monitor per 15.9 million people in SSA (World Health Organization 2018, Pinder, Klopp et al. 2019). Low-cost sensors are becoming popular for addressing the challenges of access and affordability of reference monitors, especially in LMICs. However, because the performance of these sensors can vary based on the technology and local meteorological conditions, low-cost sensors were initially used as indicative measures of air quality and were mostly used by citizen scientists to drive awareness about local air quality challenges (Amegah 2018, Levy Zamora, Xiong et al. 2018, Malings, Tanzer et al. 2020).

Nonetheless, low-cost sensors offer settings with limited reference networks an opportunity to collect air pollution data. The co-location of these sensors with reference monitors to derive calibration factors is been used to improve the quality of data from these sensors (McFarlane, Isevulambire et al. 2021). This approach offers an opportunity to interpret and use the data from these sensors to drive policy actions and research purposes. Studies from SSA are exploring the capacity of low-cost sensors with some cities in SSA now on the map of cities with air quality measurements due to the deployment and continuous development of these sensors based on the understanding of local conditions (Awokola, Okello et al. 2020, McFarlane, Isevulambire et al. 2021, Sewor and Obeng 2021).

13 Conclusion and Recommendation

The primary objective of this thesis is to bridge the gap in ground-monitor PM_{10} data using satellite information in four provinces of South Africa. This thesis addressed the objectives as follows. Firstly, we assessed the availability and methods to increase daily ground – monitor PM_{10} across the four provinces for different years between 2010 – 2017. Secondly, the spatial and temporal pattern of daily ground–monitor PM_{10} across the four provinces was assessed. Finally, we presented PM_{10} exposure maps across the four provinces for the year 2016. This thesis provides insight into the feasibility of using remote sensing data to bridge the gap in ground-monitor PM_{10} in South Africa. Our findings emphasized the gaps in ground-monitor data in South Africa and its impact on the success of PM_{10} exposure assessment using a spatio-temporal statistical modelling approach. In light of our findings, there is no one-size-fits-all approach to addressing air quality data gaps in South Africa and we believe this applies to all LMICs with limited air quality data.

The continuous advancement in air quality measurement tools and the use of remote sensing data offers South Africa excellent opportunities to bridge the gap in air quality data. However, this can only be achieved within an efficient air quality management system. An efficient air quality management system in South Africa will recognize the distribution of air quality monitoring and the pattern of air pollution in South Africa to design a complementary and integrative air quality measurement system that delineates the role of low-cost sensors and remote sensing data for air quality monitoring in South Africa. The proliferation of low-cost sensors has made it difficult to identify the best sensors. However, a science-driven low-cost sensor deployment protocol informed by results from field calibration exercises and the suitability of these sensors base on South Africa's environment, power and internet connection situation will help guide the adoption and use of these sensors in South Africa. In addition, satellite-derived estimates of air pollution can play an important role in optimally understanding the dynamics of air pollution as demonstrated by our study III. It is, therefore, important that South Africa's air quality measurement systems recognizes the growing number of satellite estimates measuring different properties of air pollutants and assess their applicability to characterizing air pollution in South Africa. While the primary focus of air quality stakeholders in South Africa should be on improving air quality monitoring, efforts should also be placed on the following:

Data transparency and data use: Access to data for research and policy actions in South Africa is still a challenge. In South Africa, there is an online repository for air quality data managed by South African Air Quality Information Systems (SAAQIS). However, the availability and quality of air quality data on this platform are dependent on SAAQIS's access to data from the different monitoring networks across South Africa. Local, provincial and private authorities manage a substantial number of sites in South Africa and exercise some level of discretion on the availability of their data on the SAAQIS platform. Thus, incorporating data transparency into SAAQIS architecture will improve local and international air pollution community access to air quality data for scientific and public awareness purposes.

Air quality management capacity building: South Africa has the most extensive air quality monitoring in SSA. However, a pertinent gap noticed in South Africa's air quality data is the lack of continuous air quality measurements. Stations are reportedly shut down for several reasons bordering the financial and technical capacities of these stations to function optimally. An evaluation of both the individual (technical officers) and organization capacities needed to ensure the effective functioning of these sites is necessary. To ensure continuous monitoring of air quality data in South Africa, the South Africa Department of Environmental Affairs coordinates air quality monitoring activities in South Africa and should collaborate with the different monitoring networks and other stakeholders to design and implement continuous training activities for the stations' technical staff and create an ecosystem for knowledge sharing and colearning on air quality measurements best practices. Adequate funding for continuous monitoring could also be achieved through cooperate social responsibility and through funding from institutions with interest in air quality monitoring.

14 References

Alvarado, M. J., A. E. McVey, J. D. Hegarty, E. S. Cross, C. A. Hasenkopf, R. Lynch, E. J. Kennelly, T. B. Onasch, Y. Awe and E. Sanchez-Triana (2019). "Evaluating the use of satellite observations to supplement ground-level air quality data in selected cities in low-and middle-income countries." Atmospheric Environment **218**: 117016.

Amegah, A. K. (2018). "Proliferation of low-cost sensors. What prospects for air pollution epidemiologic research in Sub-Saharan Africa?" Environmental pollution **241**: 1132-1137.

Arowosegbe, O. O., M. Röösli, T. C. Adebayo-Ojo, M. A. Dalvie and K. de Hoogh (2021). "Spatial and Temporal Variations in PM10 Concentrations between 2010–2017 in South Africa." International journal of environmental research and public health **18**(24): 13348.

Arowosegbe, O. O., M. Röösli, N. Künzli, A. Saucy, T. C. Adebayo-Ojo, M. F. Jeebhay, M. A. Dalvie and K. de Hoogh (2021). "Comparing Methods to Impute Missing Daily Ground-Level PM10 Concentrations between 2010–2017 in South Africa." International journal of environmental research and public health **18**(7): 3374.

Awokola, B. I., G. Okello, K. J. Mortimer, C. P. Jewell, A. Erhart and S. Semple (2020). "Measuring air quality for advocacy in Africa (MA3): Feasibility and practicality of longitudinal ambient PM2. 5 measurement using low-cost sensors." International journal of environmental research and public health **17**(19): 7243.

Department of Environmental Affairs (2016). 2nd South Africa Environment Outlook. D. o. E. Affairs. Pretoria.

Diner, D. J., S. W. Boland, M. Brauer, C. Bruegge, K. A. Burke, R. Chipman, L. Di Girolamo, M. J. Garay, S. Hasheminassab and E. Hyer (2018). "Advances in multiangle satellite remote sensing of speciated airborne particulate matter and association with adverse health effects: from MISR to MAIA." Journal of Applied Remote Sensing **12**(4): 042603.

Feig, G., R. M. Garland, S. Naidoo, A. Maluleke and V. d. M. Marna (2019). "Assessment of changes in concentrations of selected criteria pollutants in the Vaal and Highveld priority areas." Clean Air Journal **29**(2).

Feig, G. T., S. Naidoo and N. Ncgukana (2016). "Assessment of ambient air pollution in the Waterberg Priority Area 2012-2015." Clean Air Journal= Tydskrif vir Skoon Lug **26**(1): 21-28.

Government of South Africa (2005). National Environment Management: Air Quality Act 39 of 2004. Cape Town, Government of South Africa.

Hersey, S. P., R. M. Garland, E. Crosbie, T. Shingler, A. Sorooshian, S. Piketh and R. Burger (2015). "An overview of regional and local characteristics of aerosols in South Africa using satellite, ground, and modeling data." Atmospheric Chemistry and Physics **15**(8): 4259-4278.

Levy Zamora, M., F. Xiong, D. Gentner, B. Kerkez, J. Kohrman-Glaser and K. Koehler (2018). "Field and laboratory evaluations of the low-cost plantower particulate matter sensor." Environmental science & technology **53**(2): 838-849.

Malings, C., R. Tanzer, A. Hauryliuk, P. K. Saha, A. L. Robinson, A. A. Presto and R. Subramanian (2020). "Fine particle mass monitoring with low-cost sensors: Corrections and long-term performance evaluation." Aerosol Science and Technology **54**(2): 160-174.

Masih, A. (2019). "Machine learning algorithms in air quality modeling." Global Journal of Environmental Science and Management **5**(4): 515-534.

McFarlane, C., P. K. Isevulambire, R. S. Lumbuenamo, A. M. E. Ndinga, R. Dhammapala, X. Jin, V. F. McNeill, C. Malings, R. Subramanian and D. M. Westervelt (2021). "First measurements of ambient PM2. 5 in kinshasa, democratic republic of Congo and Brazzaville, republic of Congo using fieldcalibrated low-cost sensors." Aerosol and Air Quality Research **21**(7).

Pinder, R. W., J. M. Klopp, G. Kleiman, G. S. Hagler, Y. Awe and S. Terry (2019). "Opportunities and challenges for filling the air quality data gap in low-and middle-income countries." Atmospheric Environment **215**: 116794.

Sewor, C. and A. A. Obeng (2021). "The Ghana Urban Air Quality Project (GHAir): Bridging air pollution data gaps in Ghana." Clean Air Journal **31**(1): 1-2.

South African Department of Environmental Affairs (DEA) (2009). National ambient air quality standards. D. o. E. Affairs.

The Law Library of Congress (2018). Regulation of Air Pollution. G. L. R. Center. Washington, D.C. World Health Organization (2018). "WHO global ambient air quality database (update 2018)." _World Health Organization: Geneva, Switzerland_.

Zhang, C. and Y. Ma (2012). _Ensemble machine learning: methods and applications_, Springer.

15 Curriculum Vitae

Oluwaseyi Olalekan Arowosegbe, MPH

Switerland: Hofackerstrasse 61, 4132 Basel, Tel. +41779298371; aros.sheyi@gmail.com

Nigeria: No 51 Alex Ekwueme way, Jabi Lake Abuja; Tel. +2348063473854

EDUCATION

PhD Epidemiology August 2018 – July 2022
University of Basel, Swiss Tropical and Public Health Institute, Switzerland

MPH Epidemiology (Distinction in dissertation) February 2014 – July 2017 University of Cape Town, South Africa

Bachelor of Science and Honours in Public Health September 2007 – June 2011
Babcock University, Ilisan-Remo, Ogun State, Nigeria.

WORK EXPERIENCE

Swiss Tropical and Public Health Institute/University of Basel September 2018 – Present

Role: PhD candidate

- Develop grant proposal
- Design research methods
- Collect, manage, analyze environmental and health outcomes data
- Develop manuscripts and reports
- Communicate research results
- Teaching assistant for climate change and health masters course

APIN Public Health Initiatives, Lagos, Nigeria January 2018 – September 2018

Role: Program Associate (Monitoring and Evaluation)

- Collected, validated and reported programs data
- Evaluated program data
- Communicated programs data
- Designed and managed Microsoft Excel based monitoring tool

Project PINK BLUE, Abuja, Nigeria August 2017 – January 2018

Role: Research Assistant

- Developed research proposals
- Designed data collection tools
- Collected qualitative and quantitative data
- Data management and analysis
- Presented research results

Centre for Environmental and Occupational Health Research,
University of Cape Town, South Africa. April 2015 – September 2016

Role: Research Assistant

- Data management
- Data analysis
- Presented research results

STATISTICAL PROGRAMMES AND GEOGRAPHY INFORMATION SYSTEM TOOLS PROFICIENCY

R	Intermediate
STATA	Intermediate
SAS	Basic
ArcGIS & QGIS	Intermediate

RELEVANT TRAINING, WORKSHOPS AND SEMINAR

2021 Observational Epidemiology Workshop: Advanced Methods for Data and Exposure-Response Analyses, Swiss School of Public Health.

2019 International Summer School on Geospatial Data Science with R, Jena, Germany.

2016 Advanced Survival Analysis and Prognostic Modelling Workshop, Centre for Infectious Disease and Epidemiology, University of Cape Town, South Africa.

MICROSOFT OFFICE SKILLS

Microsoft Word	Intermediate
Microsoft Excel	Intermediate
Microsoft PowerPoint	Intermediate

LANGUAGES

English	Advanced
Pigin English	Advanced
Yoruba	Native

APPROVED RESEARCH PROJECTS

2018 – Swiss Government Excellence Scholarship (PhD Fellowship) (CHF 84,600)

Title of the project: Spatial-Temporal approach for bridging the gap in South Africa's historical ground level PM_{10} concentration **Role**: Main Applicant

2021 – 2024 Air Quality Management Community of Practice in West Africa funded by Bureau of Oceans and International Environmental Quality at the U.S. Department of State. **Role:** Co-Investigator

PUBLICATIONS

1. **Arowosegbe, O.O**., Röösli, M., Künzli, N., Saucy, A., Adebayo-Ojo, T.C., Schwartz, J., Kebalepile, M., Jeebhay, M.F., Dalvie, M.A. and de Hoogh, K., 2022. Ensemble averaging using remote sensing data to model spatiotemporal PM10 concentrations in sparsely monitored South Africa. Environmental Pollution, 119883.
2. Adebayo-Ojo TC, Wichmann J, **Arowosegbe OO**, Probst-Hensch N, Schindler C, Künzli N. Short-Term Joint Effects of PM10, NO2 and SO2 on Cardio-Respiratory Disease Hospital Admissions in Cape Town, South Africa. International Journal of Environmental Research and Public Health. 2022 Jan;19(1):495.
3. **Arowosegbe O.O**, Röösli M, Adebayo-Ojo TC, Dalvie MA, de Hoogh K. Spatial and Temporal Variations in PM_{10} Concentrations between 2010 – 2017 in South Africa. International journal of environmental research and public health. 18(24), p.13348.
4. **Arowosegbe O.O**, Röösli M, Künzli N, Saucy A, Adebayo-Ojo TC, Jeebhay MF, Dalvie MA, de Hoogh K. Comparing Methods to Impute Missing Daily Ground-Level PM_{10} Concentrations between 2010–2017 in South Africa. International journal of environmental research and public health. 2021 Jan;18(7): 3374.
5. Murray, Christopher JL, et al. "Five insights from the global burden of disease study 2019." *The Lancet* 396.10258 (2020): 1135-1159.
6. R.C.W. Chidebe, T.C. Orjiako, D.K. Atakere, **O.O. Arowosegbe**, D. Onu, N. Okoro, S.A. Dantsoho, E.J. Nwagboso, P. Emezue, and J. Abdulazeez. Determinants of Early Cancer Screening Behaviour in Nigeria. Journal of Global Oncology 2018 4:Supplement 2, 54s-54s