Distribution Agreement

Signature:

______________________________ _______________

Bryan N. Vu Date

Applications of Remote Sensing Data in Air Pollution Modeling and Utilization of Model-Derived Exposure Estimates in Epidemiological Studies

By

Bryan N. Vu
Doctor of Philosophy
Environmental Health Sciences

____
Yang Liu, PhD
Advisor

____
Kyle Steenland, PhD
Committee Member

____
Howard Chang, PhD
Committee Member

____
Matthew Strickland, PhD
Committee Member

____
Ana Rappold, PhD
Committee Member

Accepted:

____
Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

____
Date

Applications of Remote Sensing Data in Air Pollution Modeling and Utilization of Model-Derived Exposure Estimates in Epidemiological Studies

By

Bryan N. Vu
MSPH, Emory University, 2018
MPH, University of California, Irvine, 2016
BS, California State University, Long Beach, 2012

Advisor: Yang Liu, PhD

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Environmental Health Sciences
2021

Abstract

Applications of Remote Sensing Data in Air Pollution Modeling and Utilization of Model-Derived Exposure Estimates in Epidemiological Studies
By Bryan N. Vu

Air pollution models rely heavily on regions with sufficient ground monitors for calibration. Recent advances in remotes sensing techniques have been successfully implemented in air pollution modeling in regions with adequate monitoring networks. This dissertation aims to implement remote sensing techniques in a low- and middle-income (LMIC) setting where limited ground monitoring measurements exist. The first aim of this dissertation is to develop a satellite derived $PM_{2.5}$ (particulate matter with an aerodynamic diameter of 2.5 micrometer or less) exposure model to estimate $PM_{2.5}$ at 1 km resolution from 2010 to 2016 in Lima, Peru. Estimates from this model is subsequently used in a study to investigate the association between $PM_{2.5}$ and asthma in Lima to bridge the gaps in knowledge regarding air pollution studies in a LMIC setting where daily exposure often exceeds permissible standards. The next aim of this dissertation is to implementing remote sensing techniques in modeling a major wildfire event that requires finer spatial and temporal resolution data. The second aim of this dissertation is to build a machine learning model that incorporates low-cost sensors and the Synthetic Minority Over-sampling TEchnique (SMOTE) to artificially inflate extreme values in the training dataset to model the Camp Fire event in California in 2018. The methods and results from this aim will inform the necessary steps to improve model performance in modeling extreme events. Finally, the last aim of this dissertation is to utilize exposure estimates from a machine learning model to investigate the association between total $PM_{2.5}$, smoke $PM_{2.5}$, and non-smoke $PM_{2.5}$, and serval cardiovascular diseases (CVDs) including acute myocardial infarction, arrythmia, heart failure, ischemic heart disease, stroke, and total CVD. Results from this epidemiological study will provide more literature on the association between air pollution, both ambient and from wildland fire sources, and CVD outcomes.

Applications of Remote Sensing Data in Air Pollution Modeling and Utilization of Model-Derived Exposure Estimates in Epidemiological Studies

By

Bryan N. Vu
MSPH, Emory University, 2018
MPH, University of California, Irvine, 2016
BS, California State University, Long Beach, 2012

Advisor: Yang Liu, PhD

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Environmental Health Sciences
2021

Acknowledgement

My years at Emory have been a huge learning opportunity that I would not have obtained elsewhere. The knowledge I gained, the friends I made, the mentors I learned from, have made this an amazing journey.

First and foremost, I would like to thank my advisor, Dr. Yang Liu, for his unwavering support, insightful advice and guidance, and valuable mentorship. You have taught me how to grow as a researcher and I could not have accomplished this feat without you. I would like to also thank Dr. Kyle Steenland, Dr. Howard Chang, Dr. Matthew Strickland, and Dr. Ana Rappold for their guidance on all my aims. Each question that you pose always requires me to think outside the box and ensure that I capture the study from all angles. For that, I am extremely appreciative of my dissertation committee's time and effort in helping me build my skills in this doctoral program.

I also want to thank all my collaborators for the opportunities to work together on my aims. Thank you also to all the EH faculty and staff, including Robin, Angela, Ariadne, and Natalie, who have made my time here at Emory feel like a second home. To all the EHS students, and the best friends I have made, Donghai, Che-Jung, Jianzhao, Wenlu, and Jenn, for their constant encouragement and support.

Lastly, I want to thank my parents and my sisters for their endless support, encouragement, and love. I could not have gone very far on this journey without them. And to my nieces and nephews, your uncle always loves you.

Table of Contents

INTRODUCTION

Air pollution, particularly $PM_{2.5}$ (particulate matter with an aerodynamic diameter of 2.5 micrometer or less) is emitted from a variety of sources including engine combustion, biomass burning, power generation, and natural sources such as sea spray aerosols and wind-blown dust particles [1, 2]. $PM_{2.5}$ has also been linked to over 4 million global deaths each year as well as a plethora of adverse health outcomes including cardiovascular and respiratory diseases [3-5]. However, the composition of $PM_{2.5}$ may differ from region to region and from the source. Therefore, exposure measurements of $PM_{2.5}$ is crucial for researchers to investigate the association between air pollution and adverse health outcomes, especially in regions such as low- and middle-income countries with limited ground monitors to collect measurements. In recent years, applications of remote sensing techniques have been successfully implemented in air pollution modeling [6-9]. Satellite remote sensing data including aerosol optical depth (AOD) is a measure of extinction of the solar beam by dust and haze particles [10]. AOD can be used as a proxy for air pollution in models [11]. Using machine learning methods, AOD along with meteorological variables from chemical transport models and other ancillary parameters including vegetation index, population density, distance to road, and land use information can be used to calibrate existing ground measurements [12]. However, low- and middle-income regions including Lima, Peru has substantial air pollution issues and limited number of ground measurements to conduct epidemiological studies that require extensive ground measurements to effectively assign exposure. Only by implementing remote sensing techniques to model $PM_{2.5}$ in this region will researchers be able to effectively conduct epidemiological studies on various health outcomes to determine if Lima's air pollution needs mitigation.

Furthermore, climate change has made an impact on the Western United States. Specifically, California with its dry climate vegetation, the number wildland fires have been increasing in the past decades and each fire continues to burn more intensely and for a longer duration [13-15]. Nonetheless, even with a well-maintained network of monitors such as those maintained by the U.S. Environmental Protection Agency (U.S. EPA), the ground monitors may not accurate depict the extent and range in concentration of $PM_{2.5}$.

This is largely due to the monitors being located in more densely populated urban areas so the network of monitors lack uniform spatial distribution. Moreover, there is a lack of studies investigating the association between $PM_{2.5}$ and cardiovascular diseases (CVD). Although a limited number of studies regarding this relationship has been conducted elsewhere, mainly in China, Taiwan, and the eastern United States, few have been conducted in California to determine if $PM_{2.5}$ and more specific, smoke $PM_{2.5}$ leads to more emergency departmental visits for CVD outcomes [16-19].

This dissertation aims to address these gaps in knowledge. In the first aim, a $PM_{2.5}$ exposure model is built by introducing and implementing satellite remote sensing data and data from chemical transport models to estimate $PM_{2.5}$ at 1 km resolution between 2010 to 2016 in Lima, Peru. Estimates from this model can subsequently aid researchers in epidemiologic studies that pertain to air pollution in a quickly developing low- and middle-income country. In the second aim, a machine learning model is used to calibrate satellite remote sensing data, high resolution meteorological parameters and land use information, to a well-regulated monitoring network in California. The addition of low-cost sensor data and a Synthetic Minority Over-sampling TEchnique (SMOTE) is used to bolster the number of ground observations, allowing for an hourly model that can estimate wildland fire $PM_{2.5}$ at 3 km spatial resolution. Finally, the third aim applied satellite-derived total $PM_{2.5}$ estimated previously from a machine learning model to investigate the association between $PM_{2.5}$ and CVDs including acute myocardial infarction (AMI), arrythmia, heart failure (HF), ischemic heart disease (IHD), stroke, and total CVD. The implementation of a smoke $PM_{2.5}$ dataset from the U.S. EPA with a Hazard Mapping System that flags smoke plume pixels allows aim three to also assess smoke and non-smoke $PM_{2.5}$ with the CVD outcomes listed above.

DISSERTATION AIMS

Overarching Aim: To apply remote sensing data in air pollution modeling and utilize model-derived exposure estimates in epidemiological studies.

Aim 1: Build a $PM_{2.5}$ exposure model for Lima, Peru between 2010 to 2016 at 1 km spatial resolution using remotely sensed data such as AOD. Subsequently, use the exposure estimates derived from the model to conduct an epidemiological study that investigates the association between $PM_{2.5}$ and asthma emergency department visits.

Aim 2: Build an hourly $PM_{2.5}$ exposure for California, focusing on the Camp Fire episode in 2018, one of the biggest and most deadly wildfires in the state of California in recent years. The model will incorporate not only remotely sensed AOD, but also low-cost PM sensors and a technique to ensure that there are enough extreme high values in the training dataset to accurately predict wildland fire smoke $PM_{2.5}$.

Aim 3: Using model-derived $PM_{2.5}$ estimates, investigate the association between total $PM_{2.5}$, smoke PM2.5, and non-smoke $PM_{2.5}$, and the six cardiovascular outcomes.

REFERENCES

1. Prieto-Parra, L., et al., *Air pollution, PM2.5 composition, source factors, and respiratory symptoms in asthmatic and nonasthmatic children in Santiago, Chile.* Environment International, 2017. **101**: p. 190-200.
2. Anenberg Susan, C., et al., *Estimates of the Global Burden of Ambient PM2.5, Ozone, and NO2 on Asthma Incidence and Emergency Room Visits.* Environmental Health Perspectives, 2018. **126**(10): p. 107004.
3. Trasande, L. and G.D. Thurston, *The role of air pollution in asthma and other pediatric morbidities.* Journal of Allergy and Clinical Immunology, 2005. **115**(4): p. 689-699.
4. WHO (World Health Organization), *Burden of disease from the joint effects of Household and Ambient Air Pollution for 2012*. 2014.
5. Abrams, J.Y., et al., *Associations between Ambient Fine Particulate Oxidative Potential and Cardiorespiratory Emergency Department Visits.* Environmental Health Perspectives, 2017. **125**(10): p. 9.
6. Hu, X., et al., *Estimating ground-level PM2.5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model.* Remote Sensing of Environment, 2014. **140**: p. 220-232.
7. Kloog, I., et al., *Estimating daily PM2.5 and PM10 across the complex geo-climate region of Israel using MAIAC satellite-based AOD data.* Atmospheric Environment, 2015. **122**: p. 409-416.
8. Liu, Y., C.J. Paciorek, and P. Koutrakis, *Estimating Regional Spatial and Temporal Variability of PM(2.5) Concentrations Using Satellite Data, Meteorology, and Land Use Information.* Environmental Health Perspectives, 2009. **117**(6): p. 886-892.
9. Ma, Z., et al., *Estimating Ground-Level PM2.5 in China Using Satellite Remote Sensing.* Environmental Science & Technology, 2014. **48**(13): p. 7436-7444.
10. ESRL (Earth System Research Laboratory: Global Monitoring Division). *SURFRAD Aerosol Optical Depth*. August 25, 2017]; Available from: https://www.esrl.noaa.gov/gmd/grad/surfrad/aod/.

11. Kocifaj, M. and C.A. Gueymard, *Theoretical evaluation of errors in aerosol optical depth retrievals from ground-based direct-sun measurements due to circumsolar and related effects.* Atmospheric Environment, 2011. **45**(4): p. 1050-1058.

12. Xiao, Q., et al., *Full-coverage high-resolution daily PM2.5 estimation using MAIAC AOD in the Yangtze River Delta of China.* Remote Sensing of Environment, 2017. **199**: p. 437-446.

13. Liu, J.C., et al., *Wildfire-specific Fine Particulate Matter and Risk of Hospital Admissions in Urban and Rural Counties.* Epidemiology (Cambridge, Mass.), 2017. **28**(1): p. 77-85.

14. Xu, R., et al., *Wildfires, Global Climate Change, and Human Health.* New England Journal of Medicine, 2020. **383**(22): p. 2173-2181.

15. Li, S. and T. Banerjee, *Spatial and temporal pattern of wildfires in California from 2000 to 2019.* Scientific Reports, 2021. **11**(1): p. 8779.

16. Ban, J., et al., *Associations between short-term exposure to PM2.5 and stroke incidence and mortality in China: A case-crossover study and estimation of the burden.* Environmental Pollution, 2021. **268**: p. 115743.

17. Eichelberger, C., et al., *Emergency Department Visits and Subsequent Hospital Admission Trends for Patients with Chest Pain and a History of Coronary Artery Disease.* Cardiology and therapy, 2020. **9**(1): p. 153-165.

18. Li, M., et al., *Association Between PM2.5 and Daily Hospital Admissions for Heart Failure: A Time-Series Analysis in Beijing.* International Journal of Environmental Research and Public Health, 2018. **15**(10): p. 2217.

19. Madrigano, J., et al., *Long-term Exposure to PM2.5 and Incidence of Acute Myocardial Infarction.* Environmental Health Perspectives, 2013. **121**(2): p. 192-196.

CHAPTER 1A

Developing an Advanced $PM_{2.5}$ Exposure Model in Lima, Peru

Bryan N. Vu, Odon Sanchez, Jianzhao Bi, Qingyang Xiao, Nadia N. Hansel, William Checkley, Gustavo F. Gonzales, Kyle Steenland, Yang Liu

Vu, B.N.; Sánchez, O.; Bi, J.; Xiao, Q.; Hansel, N.N.; Checkley, W.; Gonzales, G.F.; Steenland, K.; Liu, Y. Developing an Advanced $PM_{2.5}$ Exposure Model in Lima, Peru. *Remote Sens.* **2019**, *11*, 641. https://doi.org/10.3390/rs11060641

ABSTRACT

It is well recognized that exposure to fine particulate matter ($PM_{2.5}$) affects health adversely, yet few studies from South America have documented such associations due to the sparsity of $PM_{2.5}$ measurements. Lima's topography and aging vehicular fleet results in severe air pollution with limited amounts of monitors to effectively quantify $PM_{2.5}$ levels for epidemiologic studies. We developed an advanced machine learning model to estimate daily $PM_{2.5}$ concentrations at a 1 km^2 spatial resolution in Lima, Peru from 2010 to 2016. We combined aerosol optical depth (AOD), meteorological fields from the European Centre for Medium-Range Weather Forecasts (ECMWF), parameters from the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem), and land use variables to fit a random forest model against ground measurements from 16 monitoring stations. Overall cross-validation R^2 (and root mean square prediction error, RMSE) for the random forest model was 0.70 (5.97 $\mu g/m^3$). Mean $PM_{2.5}$ for ground measurements was 24.7 $\mu g/m^3$ while mean estimated $PM_{2.5}$ was 24.9 $\mu g/m^3$ in the cross-validation dataset. The mean difference between ground and predicted measurements was −0.09 $\mu g/m^3$ (Std.Dev. = 5.97 $\mu g/m^3$), with 94.5% of observations falling within 2 standard deviations of the difference indicating good agreement between ground measurements and predicted estimates. Surface downwards solar radiation, temperature, relative humidity, and AOD were the most important predictors, while percent urbanization, albedo, and cloud fraction were the least important predictors. Comparison of monthly mean measurements between ground and predicted $PM_{2.5}$ shows good precision and accuracy from our model. Furthermore, mean annual maps of $PM_{2.5}$ show consistent lower concentrations in the coast and higher concentrations in the mountains, resulting from prevailing coastal winds blown from the Pacific Ocean in the west. Our model allows for construction of long-term historical daily $PM_{2.5}$ measurements at 1 km^2 spatial resolution to support future epidemiological studies.

KEYWORDS

$PM_{2.5}$; air pollution; MAIAC AOD; WRF-chem; random forest; machine learning; remote sensing; Lima; Peru

INTRODUCTION

$PM_{2.5}$ (fine particles with aerodynamic diameter of 2.5 µm or less), is emitted from a large variety of sources including industry, power generation, engine combustion, biomass burning, and natural sources such as sea spray aerosols and wind-blown dust particles [1, 2]. $PM_{2.5}$ contributes to 4.2 million global deaths in 2016, and studies have linked exposure to $PM_{2.5}$ with increased adverse health outcomes including respiratory and cardiovascular diseases among not only adults, but also children from North America, Europe, and Asia [3-6]. However, there is a limited number of air pollution studies in South America, where industrialization and continual urban growth may contribute to air pollution levels that far exceed those of Europe and North America [7, 8]. Current studies on air pollution in South America pertain mostly to PM_{10} (particles with aerodynamic diameter of 10 µm) or ozone, and are conducted in Brazil, Colombia, and Argentina [8-16]. To date, there has been little to no studies that investigate health outcomes with fine scale exposure measurements in South America.

Lima, Peru is the third most populous and the second most polluted major city in the Americas [4]. Lima's air pollution stems from an aging fleet of public transportation in urban areas and the widespread use of indoor biomass stoves in rural areas [4, 5]. A report by Banco Bilbao Vizcaya Argentaria (BBVA) Research indicates that the average age of Lima's vehicular fleet exceeds 15 years for private transport vehicles and 22 years for public transport vehicles [6]. Due to the densely populated urbanization of Lima, traffic congestion and exhaust from an aging motor fleet results in particulate matter levels that exceed the World Health Organization's (WHO) standards (25 µg/m^3, 24-h mean) [4, 17]. A study by Silva et al. found that for 6 of the 10 ground $PM_{2.5}$ monitors in Lima, 77% of the days between 2014 to 2015 exceeded the WHO's 24-h standards [18]. Moreover, while only 34% of the total population in Peru use solid fuel, 13% of the urban population and over 95% of the rural population rely on biomass fuel for cooking and heating, resulting in high levels of air pollution not only in urban areas but also in the mountainous rural areas [5]. Air pollution affects not only those living in Lima, but also the workers living in the rural communities in the outskirts of the city, who commute 90 to 180 minutes into the city for work [17]. Yet, there is a limited number of studies on the association between ambient air pollution and health risks in Lima. More studies are needed to assess the effects of $PM_{2.5}$, and potentially to curtail Lima's air pollution effects via new policies to improve air quality standards.

Many of the studies investigating air pollution in Lima have been cross-sectional in design, with childhood asthma as a popular health outcome [19, 20]. To date there have been no studies of air pollution and chronic disease. Limitations in directly utilizing ground-level air-monitoring data in epidemiologic studies include lack of monitoring stations and lack of daily measurements due to maintenance costs [21]. Recently, satellite remote sensing techniques have proven useful in estimating ground $PM_{2.5}$ concentrations [1]. Satellite remote sensing provide aerosol optical depth (AOD), a dimensionless measure of aerosol light extinction within a column of air on Earth's surface [22]. AOD can be used to estimate ground $PM_{2.5}$ concentrations with broad spatial coverage, expanding the ground monitoring networks into the rural areas where ground measurements are lacking [23]. Most commonly used AOD products are derive from the Moderate Resolution Imaging Spectroradiometer (MODIS) and Multiangle Imaging SpectroRadiometer (MISR) aboard the Earth observing System (EOS) satellites named Terra and Aqua launched by the National Aeronautics and Space Administration (NASA) in 1999 and 2002, respectively [24]. These products have also been widely used in recent studies to estimate $PM_{2.5}$ in southern California, China, and Pittsburgh, Pennsylvania [25-27]. A Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm, using time-series analysis and image-based processing techniques to make aerosol retrievals and atmospheric corrections over both dark vegetated land and brighter range of surfaces, can be used to retrieve AOD to achieve stronger correlations with $PM_{2.5}$ [28]. MAIAC AOD have been successfully implemented in estimating $PM_{2.5}$ in the United States, Middle East, and China [28-30].

Implementation of remote sensing techniques have proven successful in China and the United States [1, 23]. Using non-MAIAC AOD, Liu et al. compared model fit in a two-stage modeling technique to estimate $PM_{2.5}$ in Northeast U.S. with and without AOD, with results indicating that the AOD model (R^2 = 0.79) has higher predicting power compared to the non-AOD model (R^2 = 0.48) [23]. Xiao et al. conducted a study to estimate ground $PM_{2.5}$ concentrations over the Yangtze River Delta of China using MAIAC AOD and ground measurements from 2013 and 2014 with results showing good fit between ground measurements and prediction estimates (Cross Validation (CV) R^2 = 0.81 for 2013 and 0.73 for 2014) [1]. Additionally, Liang et al. implemented MAIAC AOD to estimate daily $PM_{2.5}$ concentrations in Beijing at 1 km^2 spatial resolution with high accuracy (mean annual R^2 from 0.79 to 0.86) [31]. Studies listed above found that the correlation between $PM_{2.5}$ and satellite MAIAC AOD, derived from statistical models including generalized linear regression and generalized additive modeling, are greatly improved when land use and meteorological parameters are included; nonetheless, results such as these suggests that MAIAC AOD by itself is a strong predictor of $PM_{2.5}$ concentrations [23, 28].

To date, remote sensing techniques have not been utilized in air pollution research in Lima, Peru due to insufficient ground monitoring data to correlate and validate model results. However, in recent years, the Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI) stations from the Ministry of Environment have begun collecting daily concentrations of $PM_{2.5}$ in Lima, Peru. This presents an opportunity to implement satellite remote sensing techniques in building a model to estimate ground-level $PM_{2.5}$ in a region with critically high levels of air pollution and limited number of epidemiological studies to assess its impact on health risks. In this analysis, we build a $PM_{2.5}$ exposure model to estimate daily $PM_{2.5}$ concentrations at 1 km^2 spatial resolution in Lima for years 2010 to 2016. This exposure model is derived from satellite MAIAC AOD, simulation data from chemical transport models (CTMs), meteorological fields from a forecast model, and land use parameters. The resulting daily estimates of $PM_{2.5}$ may be used in epidemiologic studies to assess its impact on both cardiovascular and respiratory health outcomes, and potentially support policies that will mitigate air pollution in Lima, Peru.

METHODS

Study area

Lima is the capital city of Peru, with over 10 million inhabitants. The city is nestled at 154 meters above sea level in the valleys of the Chillón, Rímac, and Lurín rivers, overlooking the Pacific Ocean in the west and the Andes Mountains lying about 3000 meters above sea level in the east. The study region spans from ~80 km north to south, and 40 km east to west, which includes the city of Lima and the seaport of Callao, together known as the Lima Metropolitan Area.

A grid of 2,970 1km^2 pixels was developed to cover the study region, and a 10 km buffer was added to ensure accuracy of any other parameters that need to be interpolated from coarser resolutions down to the modeling grid cells. The added buffer also allowed for better estimation of $PM_{2.5}$ concentrations near the outer boundaries of the study area. With the 10 km buffer, the total number of pixels increased to 5,959 during the model development and training period. In Figure 1, we show the study domain and location of ground monitors for the SENAMHI network and Johns Hopkins University (JHU) network as well as the mean $PM_{2.5}$ level at each monitor. The JHU network is part of the Genetic Asthma Susceptibility to Indoor Pollution in Peru, GASP study [32].

Ground $PM_{2.5}$ Data

There are ten SENAMHI stations that measure $PM_{2.5}$ and PM_{10} concentrations in Lima, Peru. These 10 monitoring stations are Thermo Beta 5014i monitors utilizing the beta ray attenuation method and are calibrated three times a year (February, June, and October, starting in October 2014) [33]. SENAMHI stations recorded daily mean measurements of PM_{10} starting in 2010 and $PM_{2.5}$ from 2014 to 2016 and its ten sites contributed 6,389 daily observations from 2014 to 2016 Additionally, data from 15 mobile air quality monitors located in Pampas de San Juan de Miraflores were provide by Johns Hopkins University (JHU stations) [34]. These monitors provided one mean estimate each week from November 2011 to March 2013, and were interpolated to the daily level by giving the six preceding days the same concentration as the measured value on the seventh day. One-km^2 grids that contained more than 1 JHU station were averaged, which reduced the number of stations from 15 to 6. The JHU sites provided 2,081 daily observations from six grid cells to the model fitting dataset. Table 1 shows the elevation and total number of measurements available at each monitor and their respective network.

Satellite Data

We Satellite aerosol optical depth (AOD) at 1 km^2 spatial resolution retrieved using the MAIAC (Multi-Angle Implementation of Atmospheric Correction) algorithm was obtained from the MAIAC science team at NASA's Goddard Space Flight Center. The MAIAC algorithm accomplishes atmospheric correction by first gridding the data to a fixed 1 km^2 grid and accumulating of up to 16 days of measurements [35]. Using a time series analysis, the pixels are grouped and the surface bidirectional reflectance distribution function (BRDF) and aerosol parameters over both dark vegetated surfaces and bright surfaces is derived [35].

AOD measurements from Arica (https://aeronet.gsfc.nasa.gov/cgi-bin/type_one_station_opera_v2_new?site=Arica&nachal=2&level=1&place_code=10) [36], the nearest Aerosol Robotic NETwork (AERONET) site located in Chile, was compared to an average of 5x5 km^2 box of MAIAC AOD centered at the Arica site to assess validity and accuracy from 2010 to 2015. AERONET is a ground-based remote sensing network that provides global observations of AOD [37]. AERONET L2 measurements within 15 minutes of the MAIAC measurements were used in the validation process to ensure accuracy; however, there may be some uncertainties in the validation results since Arica is located 1,017km northwest of Lima. Nonetheless, AERONET vs MAIAC AOD validation have been performed in the past showing good agreement [9, 38]. The highest annual correlation coefficients between MAIAC AOD and measurements from Arica ranged from 0.59 to 0.74 for Aqua and 0.60 to 0.79 for Terra. The highest correlation coefficient was observed in 2011 for Aqua and 2012 for Terra, with total number of observations ranging between 42 and 119. Subsequently, an average between Terra and Aqua MAIAC AOD was calculated and gap-filled through a random forest method discussed in Bi et al., which achieved a cross-validation R^2 of 0.82 [39]. Daily data for cloud fraction at 5 km^2 spatial resolution was downloaded from the Level-1 and Atmosphere Archive & Distribution System Distributed Active Archive Center (LAADS DAAC - https://ladsweb.modaps.eosdis.nasa.gov) [40] for 2010 to 2016 and processed through IDL. Processes of how cloud fraction data was used in gap-filling MAIAC AOD is described through Bi et al. [39].

Chemical Transport Model (CTM) data

SENAMHI produces Weather Research and Forecast model coupled with Chemistry (WRF-Chem) simulations for air quality forecasts in Lima at 5 km^2 spatial resolution [41]. WRF-Chem is a next generation atmospheric chemical transport model (CTM) developed by the National Oceanic and Atmospheric Administration (NOAA) and the National Center for Atmospheric Research (NCAR) [42]. CTMs simultaneously simulates the emissions, turbulent mixing, transport, transformation, and fate of trace gasses and aerosols using a combination of meteorological fields, topography data and emission modules based on measurements of emission factors and ambient concentrations [42]. SENAMHI WRF-Chem configuration has been previously described [41]. In brief, initial meteorological conditions were obtained from the National Centers for Environmental Prediction (NCEP) with emissions inventory derived mainly from anthropogenic vehicular emissions [41]. WRF-Chem data outputs were produced using emissions inventory based on vehicular traffic and packaged in monthly files with 26 vertical layers in the atmosphere every 6 hours (00:00, 06:00, 12:00 and 18:00 UTC); however, only the surface layer (vertical layer 0) was used and an average combining all four-time measurements were calculated. SENAMHI WRF-Chem parameters used in this study include cloud cover, albedo, surface pressure, temperature, u- and v- wind components, simulated $PM_{2.5}$, and planetary boundary layer height (PBL). There parameters were interpolated to the 1 km^2 modeling grid using an inverse distance weighting method.

Meteorological variables

Data at 6-hour increments for 28 parameters including dew point, temperature, wind and pressure was downloaded for January 2010 through December 2016 from the European Centre for Medium-Range Weather Forecasts (ECMWF) archive (http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/) [43] at 12.5 km^2 spatial resolution [44], and interpolated to the 1 km^2 modeling grid using inverse distance weighting. Subsequently, a daily average was calculated for each variable. As part of the cross-validation process, a correlation analysis was performed on temperature, wind, and pressure between WRF-Chem and ECMWF. Furthermore, temperature and dew point from ECMWF was used to calculate relative humidity (http://andrew.rsmas.miami.edu/bmcnoldy/Humidity.html) [45]. In addition, ground meteorological data was downloaded from the Weather Underground website for four individual-owned weather stations along with one airport station. These data were used to evaluate the quality of ECMWF and WRF-Chem meteorological parameters. In Figure S1 of the supplemental, we show a simple correlation matrix between Weather Underground temperature and relative humidity with WRF-Chem temperature and ECMWF relative humidity to investigate the relationship between measured ground observations and the quality of the forecasted data from ECMWF.

Land use variables

Elevation data from the Advanced Spaceborne Thermal Emission and Reflection Radiometer Global Digital Elevation Map (ASTER GDEM) was downloaded from EARTHDATA (https://search.earthdata.nasa.gov/search) [46]. Census population data for Lima was only available for 2012. To ensure completeness and consistency, LandScanTM yearly population data for 2010 through 2016 was used (https://landscan.ornl.gov/index.php/landscan-datasets) [47]. Land use parameters at 30-meter resolution (open shrubland, bare/sparse vegetation, water bodies, and artificial/urban areas) for 2010 were derived from the GlobeLand30 product produced by the National High Technology Research and Development Program of China [48]. The 30-meter spatial resolution raster was cut into 1 km^2 grids to match the MAIAC AOD grid cells, and a percent urbanization was calculated by dividing the area classified as urban in each 1 km^2 grid cell by the total area of that cell. Normalized difference vegetation index (NDVI) data at 500-meter spatial resolution (MYD13A1 Version 6) was downloaded from the LAADS DAAC for years 2010 to 2016 [49]. Since NDVI is produced at 16-day intervals, each 15 days preceding the day with measured NDVI was given the same NDVI values. Road Network Data was downloaded as an ArcGIS-ready shapefile from the OpenStreetMap project through Geofabrik (http://download.geofabrik.de/south-america/peru.html) [50], and processed in ArcGIS. The road network map was reclassified into three classes: motorways, primary and trunk roads, and secondary and tertiary roads, and a distance in meters was calculated between the centroid of each study domain grid cell to the nearest segment of road based on class.

Random forest model

A random forest (RF) model was used to fit 16 predictors to 8,470 ground measurements. The RF model's advantages include its accuracy in learning and classifying features, its ability to include a large number of input variables, and its output of variable importance. Random forest is a supervised machine learning model that works by averaging a set of decision trees that calculates the best predictions based on a subset of predictors [51]. The RF model selects a random subset of samples from all observations with replacement, and subsequently select the best set of predictors that provides the best split at each node [51]. The two main parameters in a random forest model are the number of predictors sampled for each node (m_{try}) of the tree and the number of trees or subset of samples to be averaged (n_{tree}). Comparison of results with different settings of m_{try} and n_{tree}, was conducted to achieve the best prediction accuracy. The 16 variables used in the random forest model training includes predicted MAIAC AOD from the gap-filling method, NDVI, percent urbanization, road category 3 distance, elevation, population density, interpolated WRF-Chem simulated $PM_{2.5}$, temperature, surface pressure, albedo, cloud fraction, PBL, and wind V and U components, and interpolated relative humidity and surface solar radiation downwards from ECMWF, with m_{try}, and n_{tree} set at 6, and 1000, respectively.

A 10-fold cross-validation (CV) process was carried out on the RF model to validate the prediction results. The model fitting dataset, consisting of 8,470 ground observations, were randomly divided into 10 segments with each segment containing 10% of the data. Nine of the segments were used as a training dataset set to fit the model and the remaining segment is used as a testing dataset to make predictions. This process is repeated 10 times, each time dividing the dataset at different intervals to ensure that the segments are not repeated. After the 10th repetition, the total number of predictions based on the testing dataset is combined into one dataset and is equal to the original number of ground observations. This CV technique is commonly used in similar studies estimating $PM_{2.5}$ and is better suited for a moderate to small sample size datasets.

RESULTS

Description of $PM_{2.5}$ Ground-Based Measurements

Analytic Daily predictions of $PM_{2.5}$ started on March 2, 2010 and ended on December 31, 2016. In total, 2,232 daily predictions were made between 2010 and 2016. In Figure S2 of the supplemental materials, we show histograms of all 16 predictors used in the modeling approach. Variables such as MAIAC AOD, surface solar radiation downwards, NDVI, temperature, and PBL were normally distributed. In contrast, variables that are temporally static such as road distance, elevation are non-normally distributed.

Figure 2 shows the time series of monthly mean ground measurements at each ground monitor from both the SENAMHI and the JHU networks. Mean (Std. Dev.) $PM_{2.5}$ for all JHU monitors from November 2011 to March 2013 is 18.9 (4.7) µg/m^3 with mean individual monitors ranging from 16.8 (4.0) (JHU Station 11) to 19.9 (5.8) µg/m^3 (JHU Station 9). The homogeneity of JHU measurements may be due to the spatial location of these monitors being clustered within 2-3 kilometers in the south region of the study domain. In general, ground JHU measurements peak to 29 µg/m^3 around April of 2012 and gradually decrease to 12.5 µg/m^3 in September of 2012 before increasing to a high of 30.5 µg/m^3 in March of 2013. It is unclear if $PM_{2.5}$ levels peak at this point or continues to increase as data beyond this period is unavailable for JHU monitors. All JHU monitors share this temporal trend; nonetheless, this similarity may again be due to the clustered location of the JHU monitors.

SENAMHI measurements show a slightly different temporal pattern. Mean (Std. Dev.) $PM_{2.5}$ for all SENAMHI monitors from April 2014 to December 2016 is 26.7 (11.6) µg/m^3 with mean individual monitors ranging from 15.2 (5.3) µg/m^3 (Station CDM) to 38.3 (12.2) µg/m^3 (Station ATE). SENAMHI $PM_{2.5}$ tend to peak at 52.1 µg/m^3 between July and August of 2014 (winter) and gradually decrease to 11.8 µg/m^3 around November and December (summer) before increasing again to a peak of 39.2 µg/m^3 from

March to April of 2015. Temporal trends also indicate $PM_{2.5}$ decreases from May of 2015 to a low of 13.3 μg/m^3 in February of 2016 before increasing to a peak of 63.6 μg/m^3 in June of 2016. Although most monitors within the SENAMHI network share this temporal trend, there is spatial variation coinciding with the location of the monitors. The three monitors closest to the shore (Stations CDM, SBJ, and SMP) all have the lowest mean $PM_{2.5}$ measurements (15.2, 18.2, and 17.2 μg/m^3, respectively), while the three monitors with the highest measurements (ATE: 38.3 μg/m^3, PPD: 32.8 μg/m^3, and SJL: 31.1 μg/m^3) are located further inland closer to the Andes Mountains. The differences in trends between JHU and SENAMHI networks may be a result of the JHU monitors being located in the southern part of Lima, where trends in temperature, winds, and other predictors of $PM_{2.5}$ may be different compared to the SENAMHI stations. Furthermore, SENAMHI stations are distributed across a larger area of the study domain and may have the potential to detect more spatial variability compared to JHU monitors. Although there is variability in the range of $PM_{2.5}$ levels between the two monitoring networks, both networks suggests that $PM_{2.5}$ levels are highest during the Summer Although JHU and SENAMHI stations share peaks in common during the months of March through May, ground measurements are only available for JHU sites from November of 2011 to March of 2013 and from April of 2014 to December of 2015, with no spatial or temporal similarities to the SENAMHI network. Therefore, a continuous and fair comparison of the two networks is not possible.

Random Forest Model Performance and Cross-Validation

With A linear mixed effects model (LME) was original conducted (cross-validation (CV) R^2 and root mean square error (RMSE) was 0.60 (6.85 μg/m^3)); however, the RF model was found to outperform the traditional LME model. The RF R^2 (RMSE) was 0.70 (5.95 μg/m^3), and the CV R^2 (RMSE) was 0.70 (5.97 μg/m^3), indicating that the model is stable and that there is good fit between the predictors and the ground measurements. Figure 3 panel A shows the density plot of CV predicted vs. measured $PM_{2.5}$ concentrations. The slope and intercept from the RF model CV are 1.05 and -1.04 μg/m^3, respectively, indicating a good fit (optimal, slope=1, intercept=0). Results from our CV indicates that our model slightly overestimates lower $PM_{2.5}$ measurements and underestimates higher $PM_{2.5}$ measurements. Furthermore, in Figure 3 panel

B, we show good agreement between the ground measurements and our daily estimate measurements through a Bland Altman plot. In the Bland Altman plot, the difference between ground and predicted $PM_{2.5}$ measurements are plotted against the mean of each pair. The mean difference between observations in the CV dataset was -0.09 μg/m^3 with a standard deviation of 5.97 μg/m^3. The Bland Altman plot indicates that there is good agreement between the ground and predicted measurements with 94.5% of the observations falling within 2 standard deviations of the mean differences. Figure 4, shows the importance rankings of each predictor in the RF model, which is a measure of parameter predictive power based on a permutation test. Under the null hypothesis in a random forest model, each predictor variable is not important; the permutation test rearranges the values of that variable to detect any improvement in prediction accuracy [51, 52]. The RF model suggests that surface downward solar radiation, temperature, relative humidity, PBL and AOD are the most important predictors of $PM_{2.5}$.

Figure 5 shows a time series of monthly mean ground measurements and predictions from the RF model for each ground monitor. The RF model is able to track well the temporal variability of the ground monitors, but tends to underestimate higher peaks and overestimate the low points. This trend is observed in both the SENAMHI and JHU networks. We show the predicted annual mean $PM_{2.5}$ concentrations across our study region in μg/m^3 in Figure 6. Mean annual $PM_{2.5}$ concentrations start at 14.6 μg/m^3 along the coastline and gradually increases up to 48.5 μg/m^3 against the Andes Mountains on the east. Monitors with the lower mean $PM_{2.5}$ measurements are also those that are located closer to the coast line, and are at a lower elevation. Temporally, $PM_{2.5}$ levels are highest during 2010 and dipping lower during 2011 to 2014 before increasing back up in 2015 through 2016. Although ground measurements are not available for 2010, the increase in predicted mean annual $PM_{2.5}$ from 2015 to 2016 can be observed in the monthly mean measurements from the SENAMHI monitors (Figure 2), which show a spike in $PM_{2.5}$ during the months of April and May of 2016 compared to relatively lower levels in 2015. Month to month variation can be seen in supplemental Figure S3. $PM_{2.5}$ is highest starting from April through October (highest in May-June, winter) before decreasing during the months of November to March (lowest in February, summer). Although this monthly trend is different from those observed in the JHU ground measurements, they are consistent with monthly mean SENAMHI ground measurements. This may be due to a smaller number of ground measurements for JHU compared to SENAMHI in the model fitting dataset. Furthermore, JHU monitors produced weekly measurements, which had to be interpolated to daily estimates for model fitting; therefore, monthly trends may not be meaningful for JHU measurements.

DISCUSSION

Until recently, studies to model the concentration of $PM_{2.5}$ have been limited in South America due to lack of ground monitoring data. Previous studies have estimated historical ambient $PM_{2.5}$ concentrations globally from a combination of satellite remote sensing data and chemical transport models; however, these studies were conducted at coarse resolution (e.g., 10 x 10 km^2) and were evaluated by ground $PM_{2.5}$ measurements from the literature. Furthermore, results from these studies do not provide daily measurements to aid in epidemiological health studies [53]. Brazil, Chile, Colombia, Ecuador, and Peru are the few countries with existing $PM_{2.5}$ monitors in South America prior to February of 2016; yet, Chile is the only country with known spatio-temporal and forecast models of $PM_{2.5}$ [54, 55]. The Chilean $PM_{2.5}$ model was constructed using three winter months of hourly $PM_{2.5}$ measurements from 11 monitors and incorporated CTMs; however, their model did not incorporate satellite remote sensing techniques to enhance prediction capabilities, and their model could only forecast $PM_{2.5}$ levels in the proceeding 48 hours [54]. The only current existing model of $PM_{2.5}$ in Peru is constructed through kriging techniques using ArcGIS for the province of Cusco [56]. The Cusco model was derived from a singular fixed monitor that recorded 24-h time-integrated samples for only 12 days during July 2005, and measured $PM_{2.5}$ at “subjectively chosen hot spots” using stand-alone laser photometers to augment ground measurements [56]. Although this study may provide support for short-term acute exposure of $PM_{2.5}$ health studies, it does not provide daily historical measurements for epidemiologic studies that investigate population health effects due to acute exposure to $PM_{2.5}$, especially outside of Cusco, like Lima, where pollution levels are much higher.

Our $PM_{2.5}$ model is the first advanced model in Peru to incorporate both satellite remote sensing data and CTM outputs to provide daily ground measurements at 1 km^2 resolution in Lima, the most populated and polluted region of Peru, to aid in epidemiologic studies. A major strength of this study is the ability to estimate $PM_{2.5}$ in Lima at a high resolution through the implementation of MAIAC gap-filled AOD. Our finer-scale model is able to capture local spatiotemporal trends compared to coarser resolution products, and are better suited for use in epidemiological health studies that require daily measurements of exposure at fine-resolution. Additionally, predictions from our model correspond well at each ground monitor station (as seen in Figure 5). Maximum concentrations are typically observed between May and September (winter months) with minimum concentrations generally observed between October and April (summer months); however, these trends vary from year to year and between each monitoring site. Furthermore, monthly variation in $PM_{2.5}$ concentrations is also affected by meteorological conditions present in Lima. In the summer months, Lima is subjected to smaller and less permanent marine thermal inversion due to the Humboldt oceanic current in the west. The result is a decrease in stratiform clouds and an increase in solar irradiation in conjunction with lower relative humidity and higher temperatures, which leads to re-suspension of course PM and the prevention of secondary PM formation, decreasing the levels of $PM_{2.5}$ [18]. During the winter; however, there is an increase in stratiform clouds along with an increase in relative humidity and light precipitation, resulting in wet deposition of PM_{10} and a subsequent increase in $PM_{2.5}$ due to secondary formation via converted gas-particulate [18].

Nonetheless, our study uses an emerging ensemble classifier, the random forest model, to generate our estimates which comes with limitations and uncertainties. Currently, annual predictions from the RF model show that concentrations of $PM_{2.5}$ are lowest near the coast, and in and around the urban centers of Lima, while gradually rising with elevation up to the Andes Mountains. This may be driven by the fact that all ground $PM_{2.5}$ monitors are located below 500 meters above sea level, and monitors located at lower elevation have lower $PM_{2.5}$ levels. As a result, when $PM_{2.5}$ levels are extrapolated beyond the existing ground data, their levels continue to increase with elevation up to the mountains and predictions made at elevation above 1000 meters may contain more uncertainty. Furthermore, the average height of JHU monitors is located at 132.7 (Std. Dev. = 43.6) meters above ground, while the mean height for SENAMHI monitors is 213.4 (Std. Dev. = 90.9) meters, indicating that SENAMHI monitors have a wider range of elevation height compared to JHU monitors. Additionally, JHU monitors also have a more homogeneous level of $PM_{2.5}$ since their daily values were interpolated from weekly measurements and comprised of 25% of the total ground measurements, which may add to the explanation of why elevation had relatively lower importance in the RF model. To counter the effects of elevation in the model, distance from shoreline was added to the model as a predictor. Although distance from coast should have explained much of the variation in $PM_{2.5}$ as the annual maps suggests, this variable did not improve the "out of bag" R^2 in the RF model and also did not change the resulting predictions maps and was subsequently discarded from the final model. A possible reason for why distance from coast did not improve model performance may be due to the cluster of JHU monitors all residing close to the coast. Because of their proximity to each other, as well as to the coast, the JHU monitors do not exhibit enough spatial variability both in terms of $PM_{2.5}$ levels to impact model performance. Furthermore, Lima's distinct topography and geographic location also lends to the spatial distribution of $PM_{2.5}$ concentrations. As discussed previously, much of Lima's production of $PM_{2.5}$ stems from an aging vehicular fleet located mostly in the densely populated urban areas in and around the metropolitan cities. Additionally, $PM_{2.5}$ is also being produced in rural areas from biomass burning as fuel. The spatial pattern of $PM_{2.5}$ seen in the annual prediction maps may be a result of persistent and prevailing coastal winds from the south and southwest pushing pollutants from the coastal cities and trapping them against the Andes Mountains in the east and northeast [18]. This phenomenon is similar to that seen in the Los Angeles Basin, where the topography is nearly identical to that of Lima with prevailing coastal winds blowing pollutants against the Transverse Ranges [57]. Nonetheless, census data indicate that the number of residents living above 1000 meters above sea level is relatively small and may not impact future epidemiologic studies.

Consequently, a limitation of this study is the lack of monitors located at higher altitudes to validate our results. All monitors are located centrally in the urbanized metropolitan area of Lima, with no monitors in the far corners of the North, East, and South in our study domain. Furthermore, all JHU monitors are clustered within a few kilometers of each other in the mid-southern region of Lima, covering 6 of the 2,970 grid cells in the study domain which may affect their predictive capabilities on the rest of the study domain leading to the lack of spatial variability from north to south in the study domain. Additionally, JHU ground measurements were collected from late 2011 to early 2013, while the SENAMHI measurements were collected from mid-2014 through 2016, which impact model predictive abilities across the years (i.e. borrowing prediction capabilities of JHU measurements to estimate $PM_{2.5}$ in the entire study domain for 2014 to 2016 and conversely borrowing prediction capabilities of SENAMHI measurements to estimate $PM_{2.5}$ in the entire study domain for 2011 to 2013). Nonetheless, JHU measurements served the purpose of increasing our sample size and help to make our model more stable and robust. When JHU measurements were not included in our RF model, the CV R^2 was 0.67 (RMSE=6.68 μg/m^3), and were subsequently kept in the model fitting dataset to enhance not only sample size but also to provide additional spatial and temporal quality to the ground measurements. Finally, before utilizing and applying the model-derived dataset in epidemiological studies, future research will focus on evaluating model forecasting capacity on a daily basis. Furthermore, the SENAMHI ground monitors have longer periods of PM_{10} measurements. Future study will also explore converting PM_{10} measurements to $PM_{2.5}$ to maximize ground observations in the model fitting process [58]. Silva et al. have studied the relationship between $PM_{2.5}$ and PM_{10} concentrations at each of the 10 SENAMHI stations with Pearson correlation coefficients ranging from 0.49 to 0.72, and that the annual $PM_{2.5}$/PM_{10} for the stations range from 0.21 to 0.44, indicating that $PM_{2.5}$ concentrations represent 21% to 44% of the total PM_{10} in Lima [18].

CONCLUSIONS

Our satellite-driven $PM_{2.5}$ exposure model is the first of its kind in both Lima and South America, incorporating satellite remote sensing data, meteorological fields from chemical transport models, and land use parameters to estimate daily $PM_{2.5}$ measurements at 1 km resolution, with greater spatial and temporal coverage than previous studies conducted in Peru. Predicted daily $PM_{2.5}$ levels by our model allow for construction of consistent long-term historical measurements that bridges the data gaps created by sparse data quality from both the SENAMHI and JHU monitor networks, and would provide strong data support for epidemiologic studies that focus on both cardiovascular and respiratory outcomes in Lima. Our future research will focus on converting PM_{10} to $PM_{2.5}$ from the SENAMHI monitors to maximize ground observations across years prior to 2014, and improve model stability and precision, and further improve on the accuracy of our predictions for use in urgently needed epidemiologic studies to assess the impact of air pollution in Lima, Peru.

ACKNOWLEDGEMENT

Research reported in this publication was supported by the NIH Fogarty International Center, National Institutes of Environmental Health Sciences (NIEHS) R01ES018845, R01ES018845-S1, National Cancer Institute, National Institute for Occupational Safety and Health, and the NIH under Award Number U01 TW0101 07. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research was also support by the HERCULES Center Pilot Project Program. We also express great thanks to Gustavo Gonzales for his support in this project

REFERENCE

1. Ma, Z., et al., Estimating Ground-Level PM2.5 in China Using Satellite Remote Sensing. Environmental Science & Technology, 2014. 48(13): p. 7436-7444.

2. Prieto-Parra, L., et al., Air pollution, PM2.5 composition, source factors, and respiratory symptoms in asthmatic and nonasthmatic children in Santiago, Chile. Environment International, 2017. 101: p. 190-200.

3. Liu, Q., et al., Effect of exposure to ambient PM(2.5) pollution on the risk of respiratory tract diseases: a meta-analysis of cohort studies. Journal of Biomedical Research, 2017. 31(2): p. 130-142.

4. WHO (World health Organization). WHO Global Urban Ambient Air Pollution Database. 2016 August 25, 2017]; Available from: http://www.who.int/phe/health_topics/outdoorair/databases/cities/en/.

5. WHO (World Health Organization), Climate and Health Country Profile - 2015: Peru. 2015.

6. Research, B., Peru Automobile Market Outlook. 2010.

7. Mead, N.V. Pant by numbers: the cities with the most dangerous air-listed. 2017 February 13 2017 [cited 2019 March 4 2019]; Available from: https://www.theguardian.com/cities/datablog/2017/feb/13/most-polluted-cities-world-listed-region.

8. González, C.M., et al., High-resolution air quality modeling in a medium-sized city in the tropical Andes: Assessment of local and global emissions in understanding ozone and PM10 dynamics. Atmospheric Pollution Research, 2018. 9(5): p. 934-948.

9. Della Ceca, L.S., et al., Satellite-based view of the aerosol spatial and temporal variability in the Córdoba region (Argentina) using over ten years of high-resolution data. ISPRS Journal of Photogrammetry and Remote Sensing, 2018. 145: p. 250-267.

10. Gómez, C.D., et al., Spatial and temporal disaggregation of the on-road vehicle emission inventory in a medium-sized Andean city. Comparison of GIS-based top-down methodologies. Atmospheric Environment, 2018. 179: p. 142-155.

11. Martins, L.D., et al., Extreme value analysis of air pollution data and their comparison between two large urban regions of South America. Weather and Climate Extremes, 2017. 18: p. 44-54.

12. Zalakeviciute, R., et al., Quantifying decade-long effects of fuel and traffic regulations on urban ambient PM2.5 pollution in a mid-size South American city. Atmospheric Pollution Research, 2018. 9(1): p. 66-75.

13. Lin, et al., Air pollution and respiratory illness of children in São Paulo, Brazil. Paediatric and Perinatal Epidemiology, 1999. 13(4): p. 475-488.

14. Ribeiro, A.G., et al., Incidence and mortality for respiratory cancer and traffic-related air pollution in São Paulo, Brazil. Environmental Research, 2019. 170: p. 243-251.

15. Amarillo, A.C. and H.A. Carreras, The effect of airborne particles and weather conditions on pediatric respiratory infections in Cordoba, Argentine. Environmental Pollution, 2012. 170: p. 217-221.

16. de Miranda, R.M., et al., Urban air pollution: a representative survey of PM2.5 mass concentrations in six Brazilian cities. Air Quality, Atmosphere & Health, 2012. 5(1): p. 63-77.

17. Scholl, L., et al., Comparative Case Studies of Three IDB-Supported Urban Transport Projects. Inter-American Development Bank. 2015.

18. Silva, J., et al., Particulate matter levels in a South American megacity: the metropolitan area of Lima-Callao, Peru. Environmental Monitoring and Assessment, 2017. 189(12): p. 635.

19. Baumann, L.M., et al., Effects of distance from a heavily transited avenue on asthma and atopy in a periurban shantytown in Lima, Peru. Journal of Allergy and Clinical Immunology, 2011. 127(4): p. 875-882.

20. Carbajal-Arroyo, L., et al., Impact of Traffic Flow on the Asthma Prevalence Among School Children in Lima, Peru. Journal of Asthma, 2007. 44(3): p. 197-202.

21. CEHTP (California Environmental Health Tracking Program). Air Quality: Measures and Limitations. April 11, 2018]; Available from: http://www.cehtp.org/faq/air/air_quality_measures_and_limitations.

22. ESRL (Earth System Research Laboratory: Global Monitoring Division). SURFRAD Aerosol Optical Depth. August 25, 2017]; Available from: https://www.esrl.noaa.gov/gmd/grad/surfrad/aod/.

23. Liu, Y., C.J. Paciorek, and P. Koutrakis, Estimating Regional Spatial and Temporal Variability of PM(2.5) Concentrations Using Satellite Data, Meteorology, and Land Use Information. Environmental Health Perspectives, 2009. 117(6): p. 886-892.

24. Remer, L.A., et al., The MODIS Aerosol Algorithm, Products, and Validation. Journal of the Atmospheric Sciences, 2005. 62(4): p. 947-973.

25. Meng, X., et al., Estimating PM2.5 speciation concentrations using prototype 4.4 km-resolution MISR aerosol properties over Southern California. Atmospheric Environment, 2018. 181: p. 70-81.

26. Russell, M.C., J.H. Belle, and Y. Liu, The impact of three recent coal-fired power plant closings on Pittsburgh air quality: A natural experiment. Journal of the Air & Waste Management Association, 2017. 67(1): p. 3-16.

27. Zheng, Y., et al., Estimating ground-level PM2.5 concentrations over three megalopolises in China using satellite-derived aerosol optical depth measurements. Atmospheric Environment, 2016. 124: p. 232-242.

28. Hu, X., et al., Estimating ground-level PM2.5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. Remote Sensing of Environment, 2014. 140: p. 220-232.

29. Kloog, I., et al., Estimating daily PM2.5 and PM10 across the complex geo-climate region of Israel using MAIAC satellite-based AOD data. Atmospheric Environment, 2015. 122: p. 409-416.

30. Xiao, Q., et al., Full-coverage high-resolution daily PM2.5 estimation using MAIAC AOD in the Yangtze River Delta of China. Remote Sensing of Environment, 2017. 199: p. 437-446.

31. Liang, F., et al., MAIAC-based long-term spatiotemporal trends of PM2.5 in Beijing, China. Science of The Total Environment, 2018. 616-617: p. 1589-1598.

32. Bose, S., et al., Association of traffic air pollution and rhinitis quality of life in Peruvian children with asthma. PloS one, 2018. 13(3): p. e0193910-e0193910.

33. Scientific, T.F. 5014i Beta Continuous Ambient Particulate Monitor. March 4 2019]; Available from: https://www.thermofisher.com/order/catalog/product/5014I.

34. Underhill, J.L., et al., Association of Roadway Proximity with Indoor Air Pollution in a Peri-Urban Community in Lima, Peru. International Journal of Environmental Research and Public Health, 2015. 12(10).

35. Lyapustin, A., et al., MODIS Collection 6 MAIAC algorithm. Vol. 11. 2018. 5741-5765.

36. Gile, D.M. AERONET: AEROSOL ROBOTIC NETWORK: Site: Arica. 2019 March 6 2019]; Available from: https://aeronet.gsfc.nasa.gov/cgi-bin/type_one_station_opera_v2_new?site=Arica&nachal=2&level=1&place_code=10.

37. Giles, D.M. AERONET: AEROSOL ROBOTIC NETWORK. 2018 April 11, 2018]; Available from: https://aeronet.gsfc.nasa.gov/.

38. Martins, V.S., et al., Seasonal and interannual assessment of cloud cover and atmospheric constituents across the Amazon (2000–2015): Insights for remote sensing and climate analysis. ISPRS Journal of Photogrammetry and Remote Sensing, 2018. 145: p. 309-327.

39. Bi, J., et al., Incorporating Snow and Cloud Fractions in Random Forest to Estimate High Resolution PM2.5 Exposures in New York State. 2018, Emory University.

40. Ederer, G. EARTHDATA: LAADS DAAC. March 6 2019]; Available from: https://ladsweb.modaps.eosdis.nasa.gov.

41. Sánchez-Ccoyllo, O., et al., Modeling study of the particulate matter in lima with the WRF-Chem model: Case study of april 2016. International Journal of Applied Engineering Research, 2018. 13: p. 10129-10141.

42. Grell, G.A., et al., Fully coupled “online” chemistry within the WRF model. Atmospheric Environment, 2005. 39(37): p. 6957-6975.

43. Forecasts, E.C.f.M.-R.W. ERA Interim, Daily. March 6, 2019]; Available from: http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/.

44. ECMWF (European Centre for Medium-Range Weather Forecasts). About. 2018 April 12, 2018]; Available from: https://www.ecmwf.int/en/about.

45. McNoldy, B. Calculate Temperature, Dewpoint, or Relative Humidity. March 6, 2019]; Available from: http://andrew.rsmas.miami.edu/bmcnoldy/Humidity.html.

46. Berrick, S. EARTHDATA: EARTHDATA Search. March 6, 2019]; Available from: https://search.earthdata.nasa.gov/search.

47. Energy, U.-B.f.t.D.o. Oak Ridge National Laboratory: LandScan Datasets. March 6, 2019]; Available from: https://landscan.ornl.gov/index.php/landscan-datasets.

48. Chen, J., et al., Global land cover mapping at 30m resolution: A POK-based operational approach. ISPRS Journal of Photogrammetry and Remote Sensing, 2015. 103: p. 7-27.

49. Didan, K., MYD13A1 MODIS/Aqua Vegetation Indices 16-day L3 Global 500m SIN Grid V006, N.E.L. DAAC, Editor. 2015.

50. GmbH, G. GEOFABRIK downloads: Peru. Available from: http://download.geofabrik.de/south-america/peru.html.

51. Breiman, L., Random Forests. Machine Learning, 2001. 45(1): p. 5-32.

52. Liaw, A. and M. Wiener, Classification and Regression by RandomForest. Vol. 23. 2001.

53. van Donkelaar, A., et al., Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. Environ Health Perspect, 2010. 118(6): p. 847-55.

54. Nicolis, O., et al., Spatio-temporal modelling for assessing air pollution in Santiago de Chile. AIP Conference Proceedings, 2017. 1798(1): p. 020113.

55. Riojas-Rodriguez, H., et al., Air pollution management and control in Latin America and the Caribbean: implications for climate change. Rev Panam Salud Publica, 2016. 40(3): p. 150-159.

56. Pearce, J.L., et al., Characterizing the spatiotemporal variability of PM2.5 in Cusco, Peru using kriging with external drift. Atmospheric Environment, 2009. 43(12): p. 2060-2069.

57. Kim, S., et al., Size Distribution and Diurnal and Seasonal Trends of Ultrafine Particles in Source and Receptor Sites of the Los Angeles Basin. Journal of the Air & Waste Management Association, 2002. 52(3): p. 297-307.

58. Yuval and D.M. Broday, Enhancement of PM2.5 exposure estimation using PM10 observations. Environmental Science: Processes & Impacts, 2014. 16(5): p. 1094-1102.

CHAPTER 1A TABLES AND FIGURES

Table 1. $PM_{2.5}$ ground monitor information, elevation, and total number of observations at each monitor and their respective network.

Network	Station	Elevation (m.)	# of Measurements
JHU	Station 02	94.6	339
JHU	Station 07	123.6	417
JHU	Station 08	74.2	288
JHU	Station 09	186.0	443
JHU	Station 10	192.1	287
JHU	Station 11	109.2	307
SENAMHI	ATE	372.7	528
SENAMHI	CDM	124.5	544
SENAMHI	CRB	219.5	737
SENAMHI	HCH	301.2	696
SENAMHI	PPD	186.0	778
SENAMHI	SBJ	131.3	581
SENAMHI	SJL	237.5	757
SENAMHI	SMP	58.5	775
SENAMHI	STA	254.3	598
SENAMHI	VMT	328.3	395

Note: SENAMHI Station is abbreviated from the name of the location. JHU stations collected measurements from November 2011 to March 2013 and SENAMHI stations collected measurements from April 2014 to December 2016.

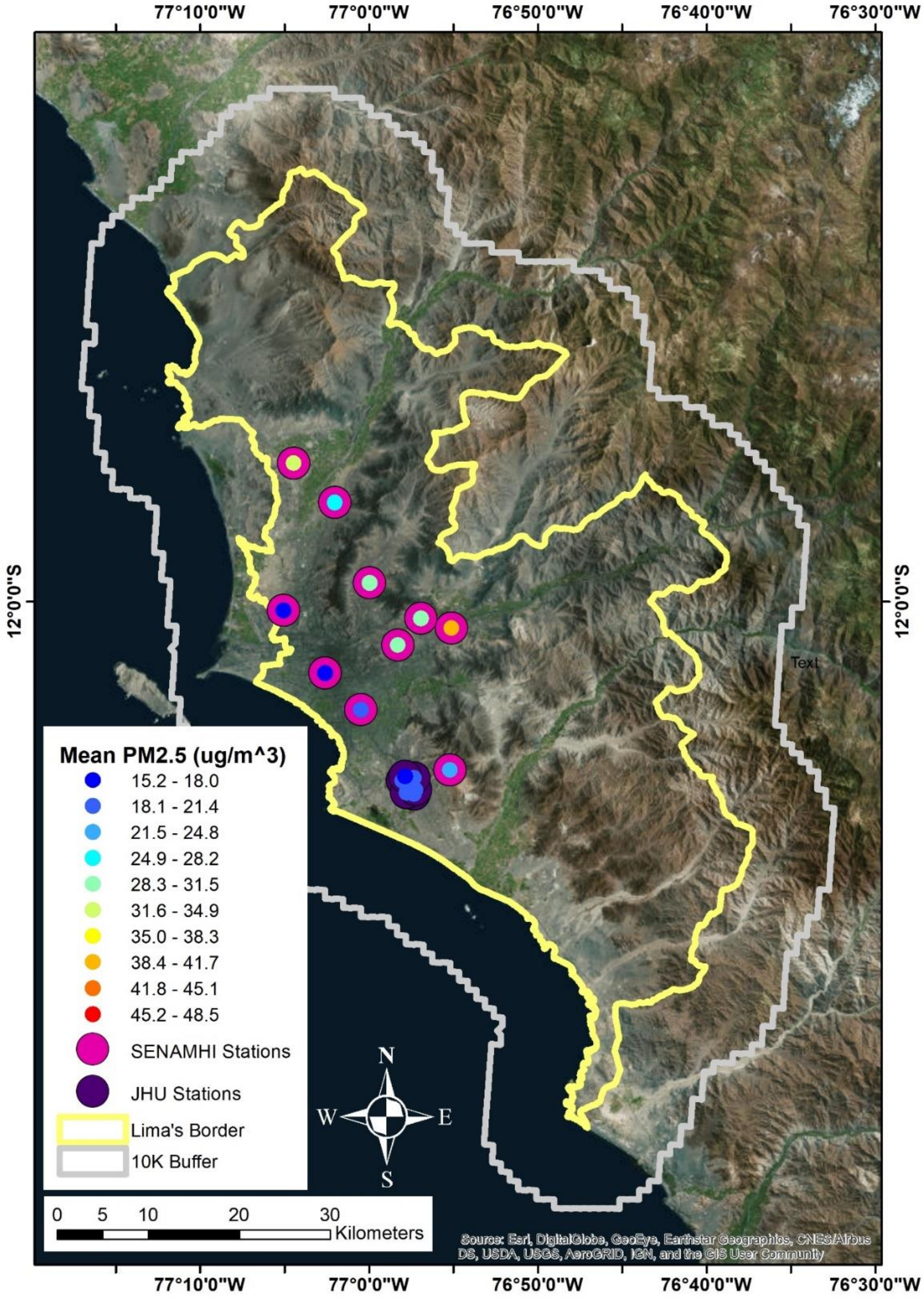


Figure 1. Study domain and location of air monitors. The yellow line details the Lima political border while the grey line details the 10km buffer. The magenta circles denote the location, distribution, and overall mean $PM_{2.5}$ concentrations in μg/m^3 of the SENAMHI monitor network while the purple circles denote the same information for the JHU monitor network.

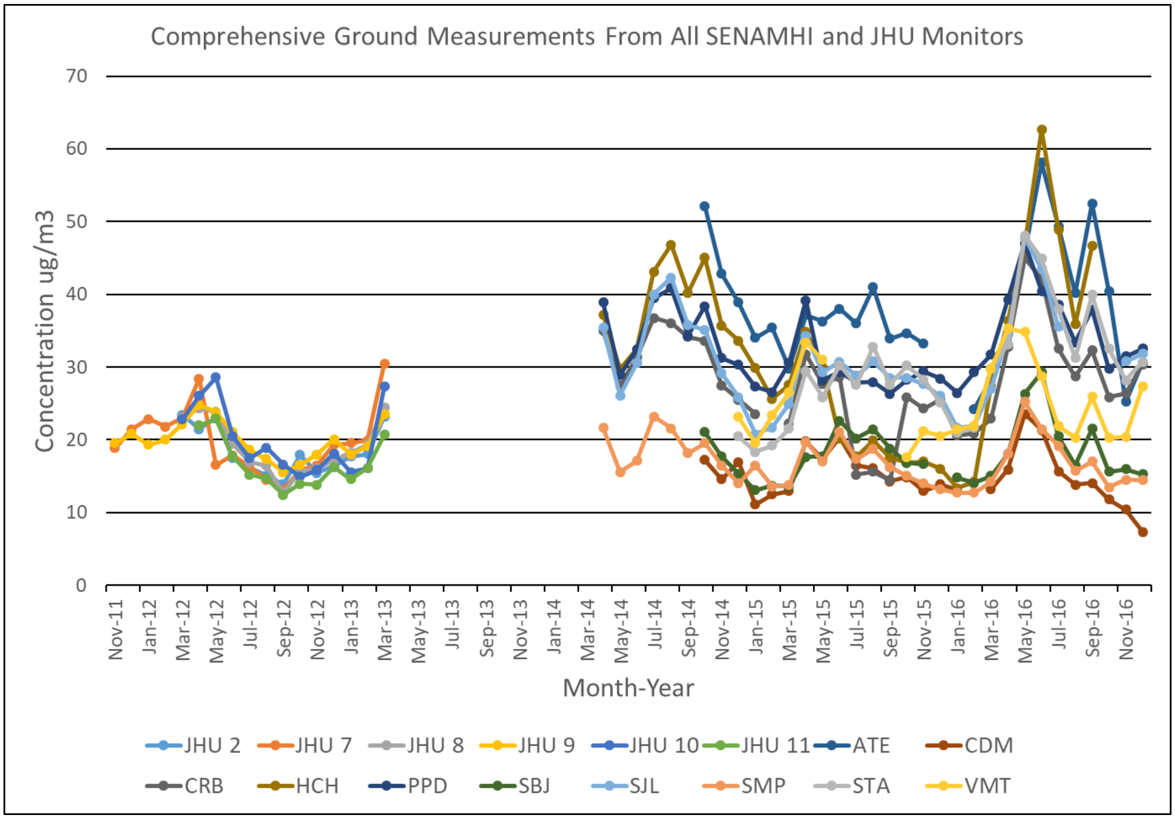


Figure 2. Time series of monthly mean ground $PM_{2.5}$ measurements in µg/m^3 at each monitor station for both SENAMHI and JHU network from November 2011 through December 2016. SENAMHI Station names are abbreviated from the name of the location.

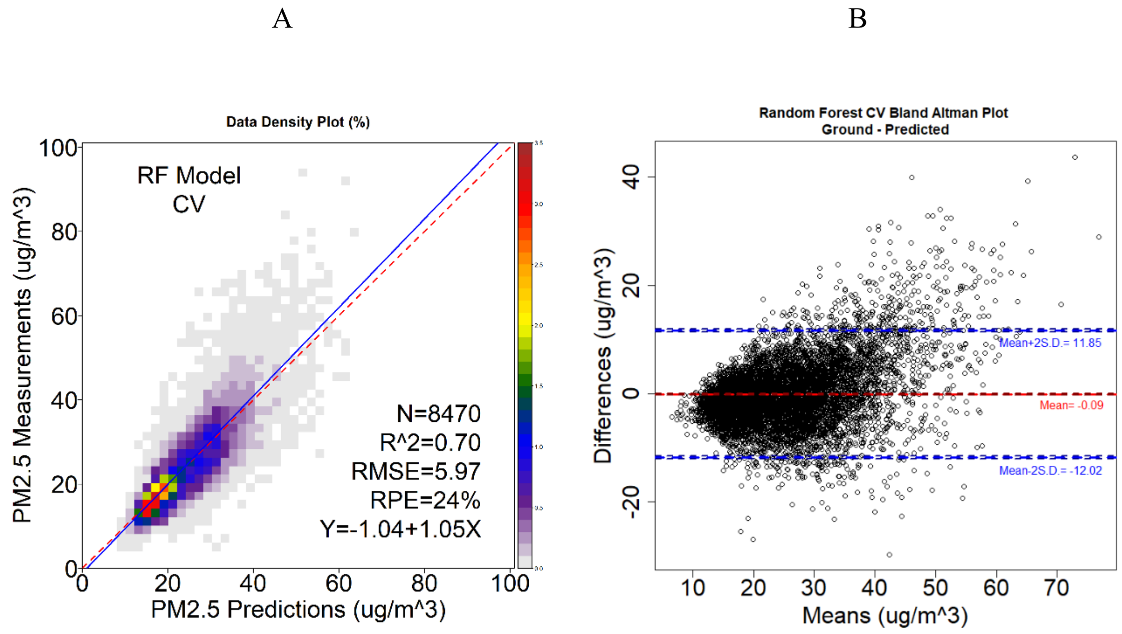


Figure 3. (A) Density plot of ground and predicted $PM_{2.5}$ measurements in μg/m^3 based on the cross-validation of the Random Forest model. (B) Bland-Altman plot of differences between ground and predicted $PM_{2.5}$ in μg/m^3 against the means of each pair. This plot shows good agreement as 94.5% of observation pairs fall within 2 standard deviations of the mean difference.

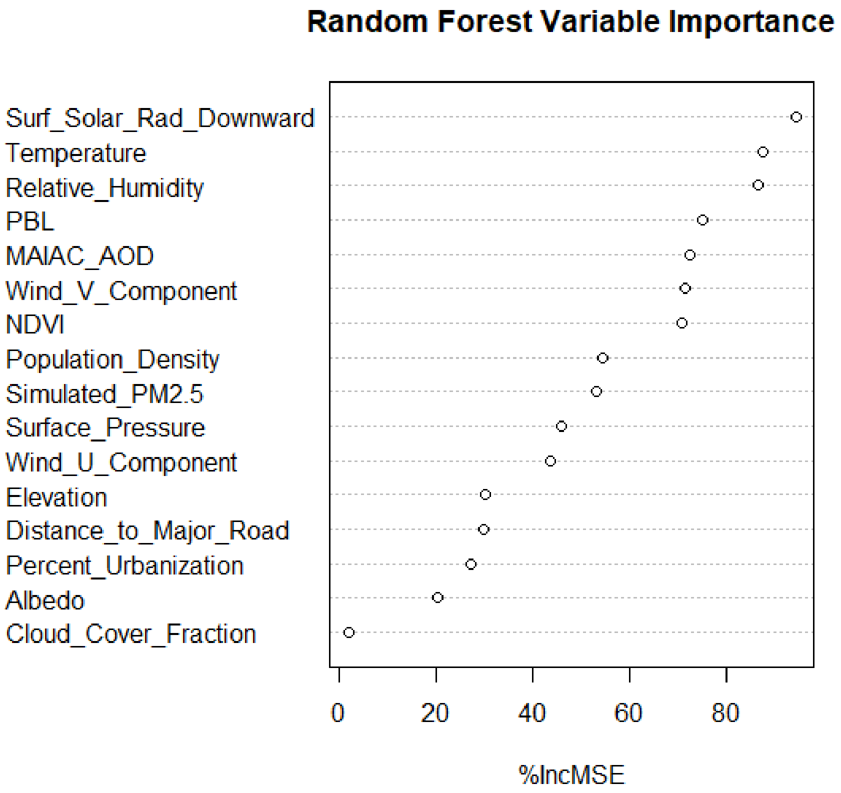
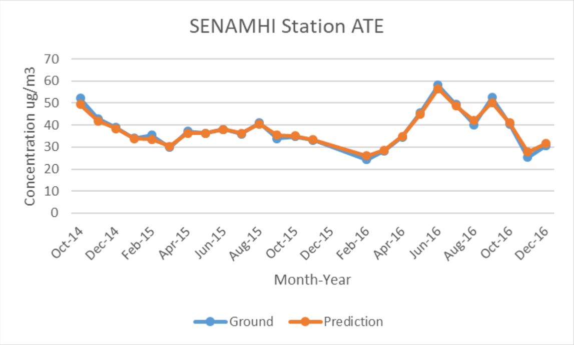
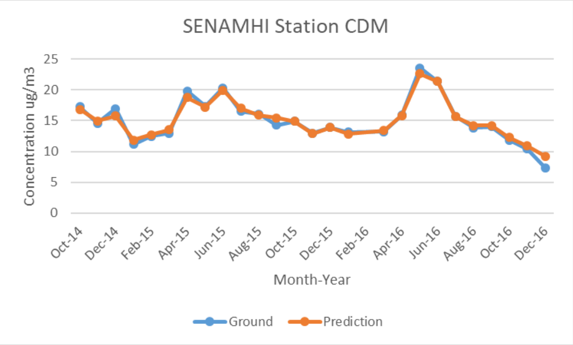
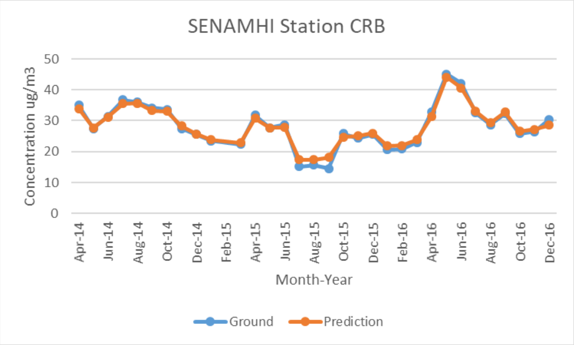
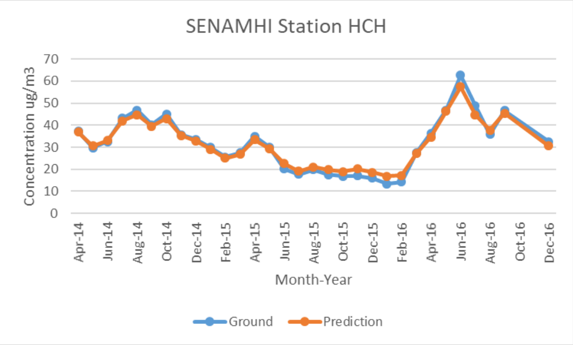


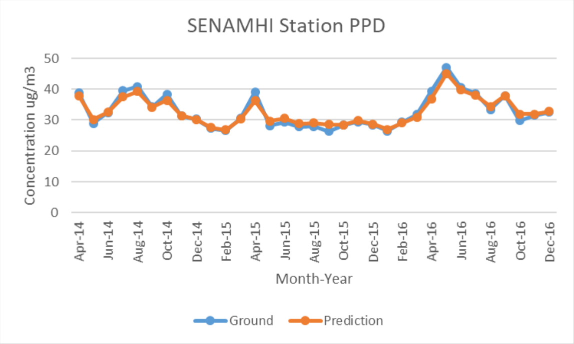
Figure 4. Importance of each variable in the Random Forest model by percent increase MSE.

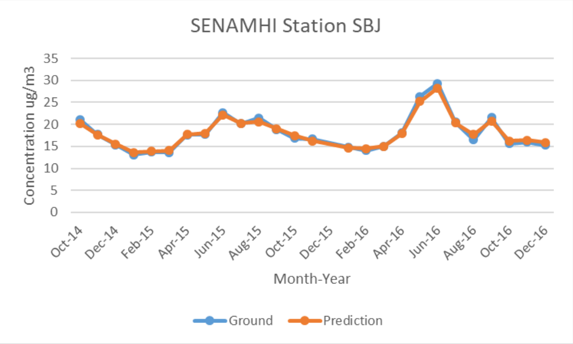

SENAMHI Station ATE

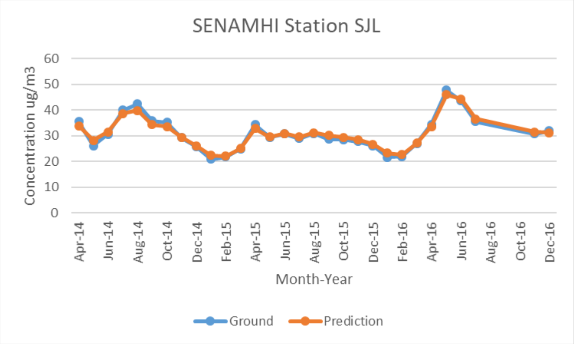

SENAMHI Station CDM

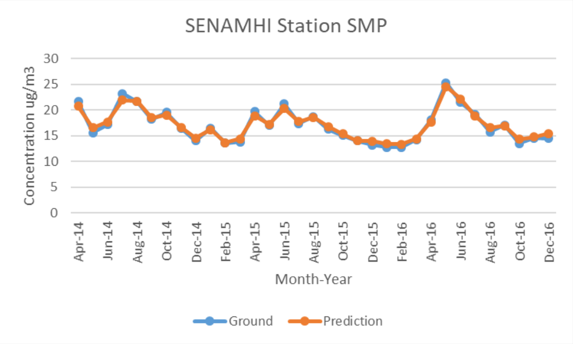

SENAMHI Station CRB


SENAMHI Station HCH


SENAMHI Station PPD


SENAMHI Station SBJ


SENAMHI Station SJL


SENAMHI Station SMP

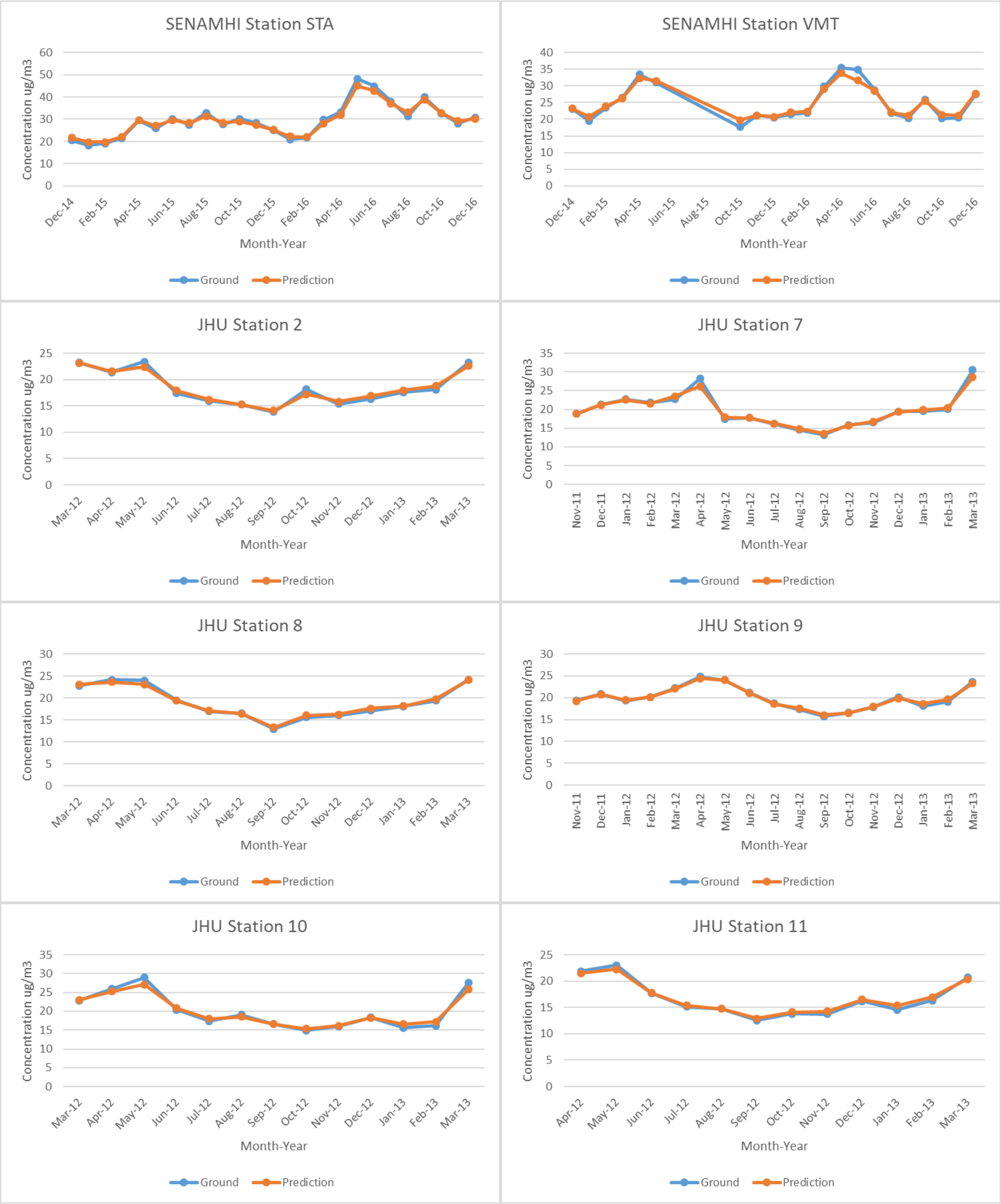


Figure 5. Time series of monthly mean ground measurements and predicted $PM_{2.5}$ in µg/m^3 based on Random Forest model at each monitor station. SENAMHI Station names are abbreviated from the name of the location.

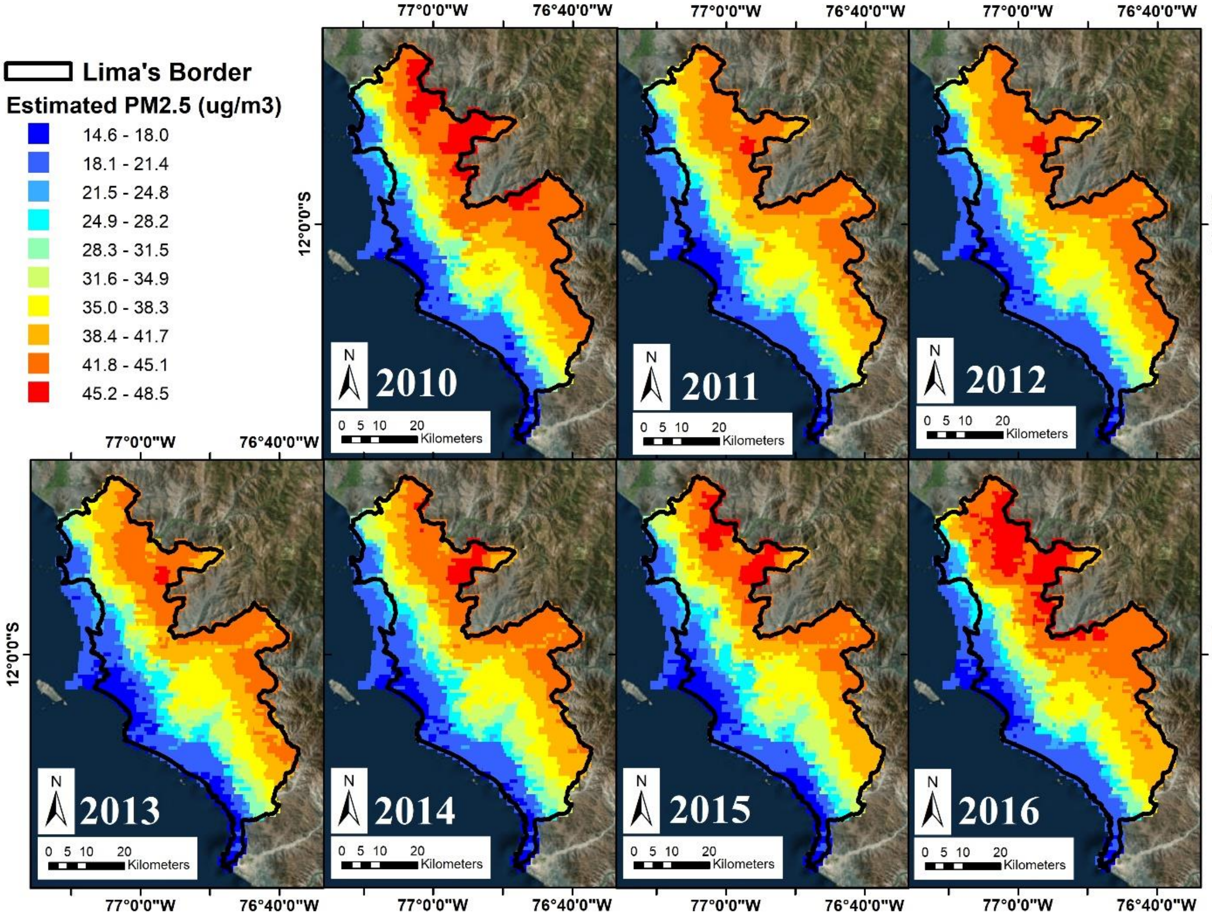


Figure 6. Annual mean prediction maps of $PM_{2.5}$ in µg/m^3 from the Random Forest model in Lima, Peru from 2010 to 2016.

SUPPLEMENTAL CHAPTER 1A

1. Supplemental Methods

Measurements for daily average temperature and relative humidity from Weather Underground were correlated with temperature from WRF-Chem and relative humidity from ECMWF. In Figure S1, we show the simple correlation matrix between these variables. Correlation coefficients between Weather Underground temperature with WRF-Chem temperature and ECMWF relative humidity were 0.69 and 0.59, respectively. Correlation coefficients between Weather Underground relative humidity with WRF-Chem temperature and ECMWF relative humidity were both 0.05, respectively. The correlation coefficient between Weather Underground temperature and relative humidity was 0.17.

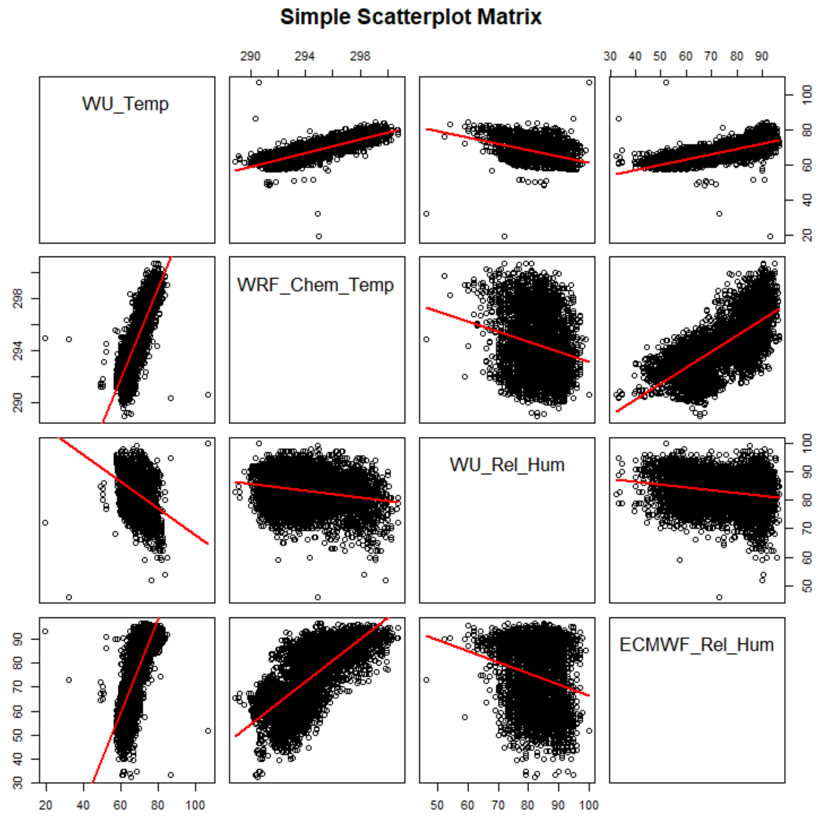
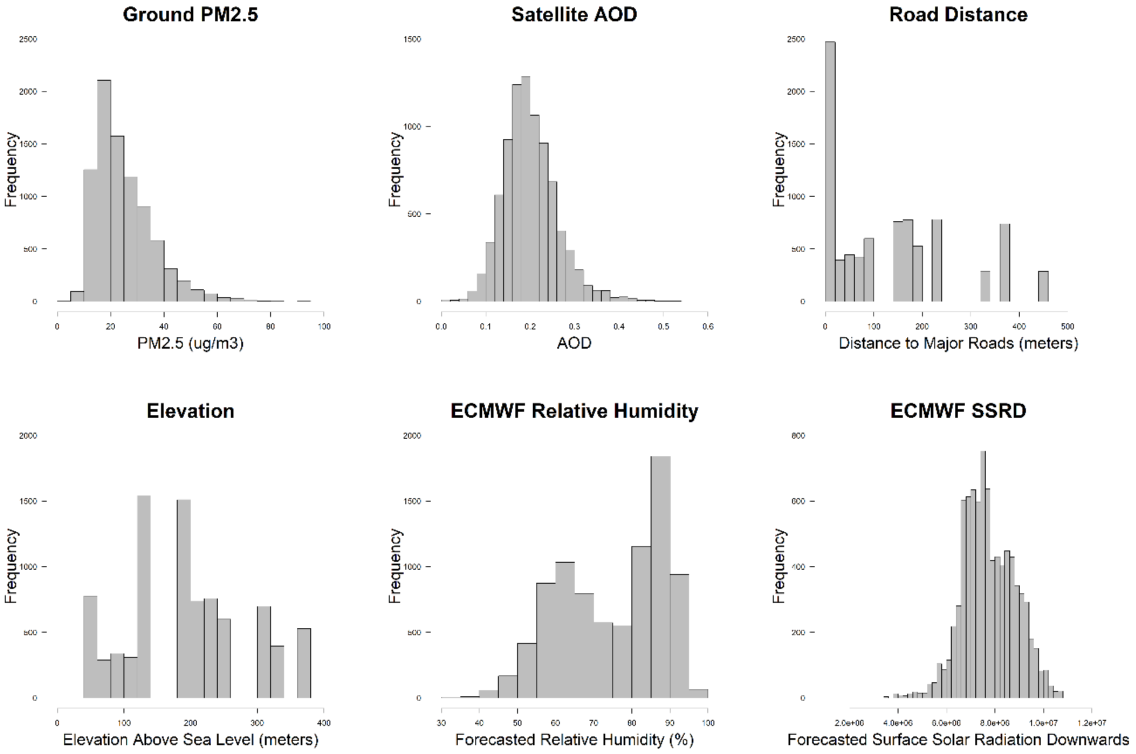


Figure S1. Simple Correlation Matrix between Weather Underground temperature and relative humidity with WRF-Chem temperature and ECMWF relative humidity.

2. Supplemental Results

Histograms of all the predictors used in the modeling approach can be seen in Figure S2. Variables including AOD, surface solar radiation downwards (SSRD, or the solar radiation in the downward direction at the surface), NDVI, temperature, PBL, and wind U component were considered normally distributed. Distance from monitor to nearest major road, elevation, percent urbanization, and population are assumed to be mostly static between years and non-normally distributed due to limited number of ground monitors with 6 of the 10 monitors (JHU sites) densely clustered within a region. WRF-Chem $PM_{2.5}$ was right skewed while wind V component was left skewed. WRF-Chem albedo, cloud fraction, and pressure are also non-normally distribution, which may also be a result of the location and distribution of the monitor stations.



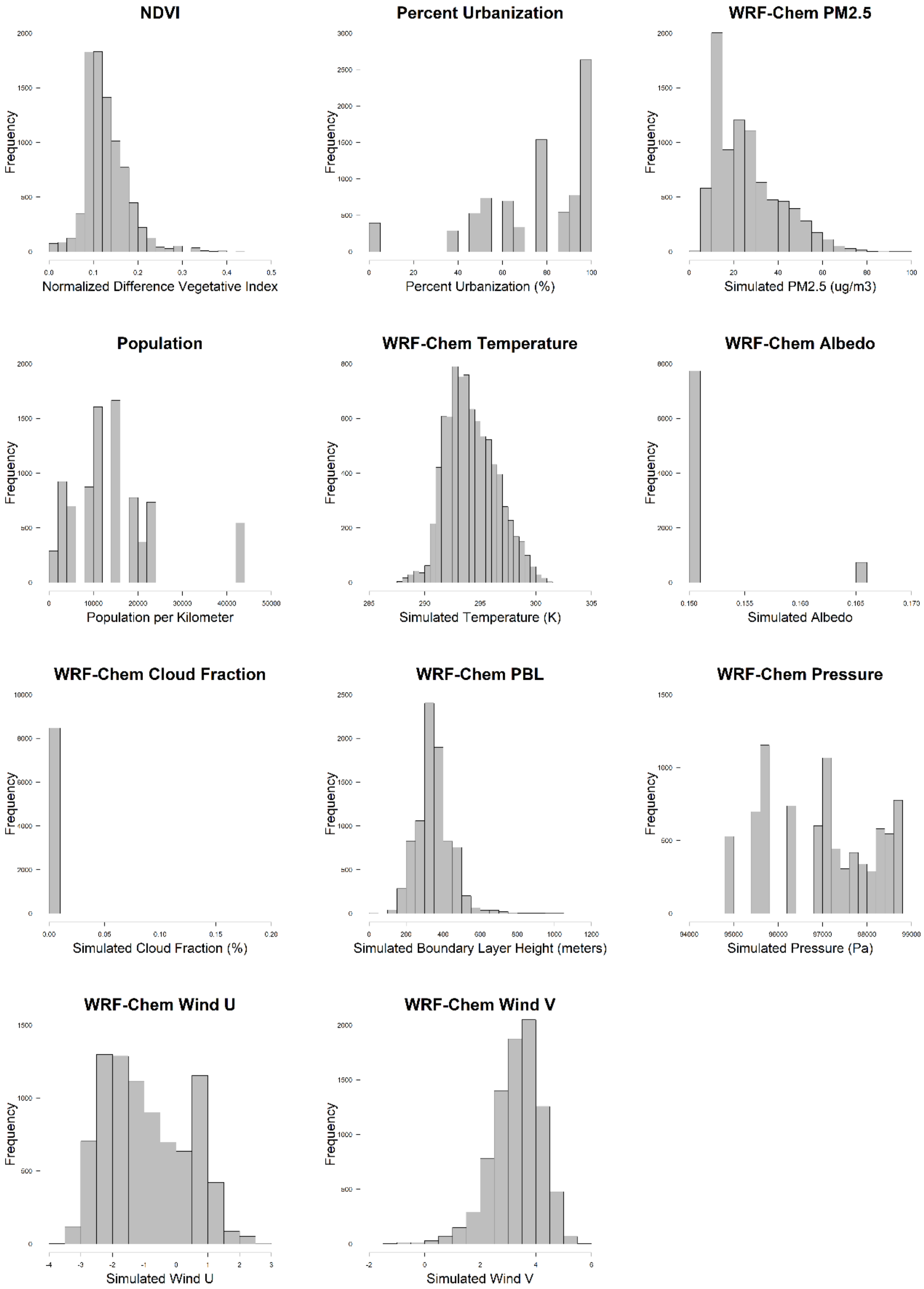
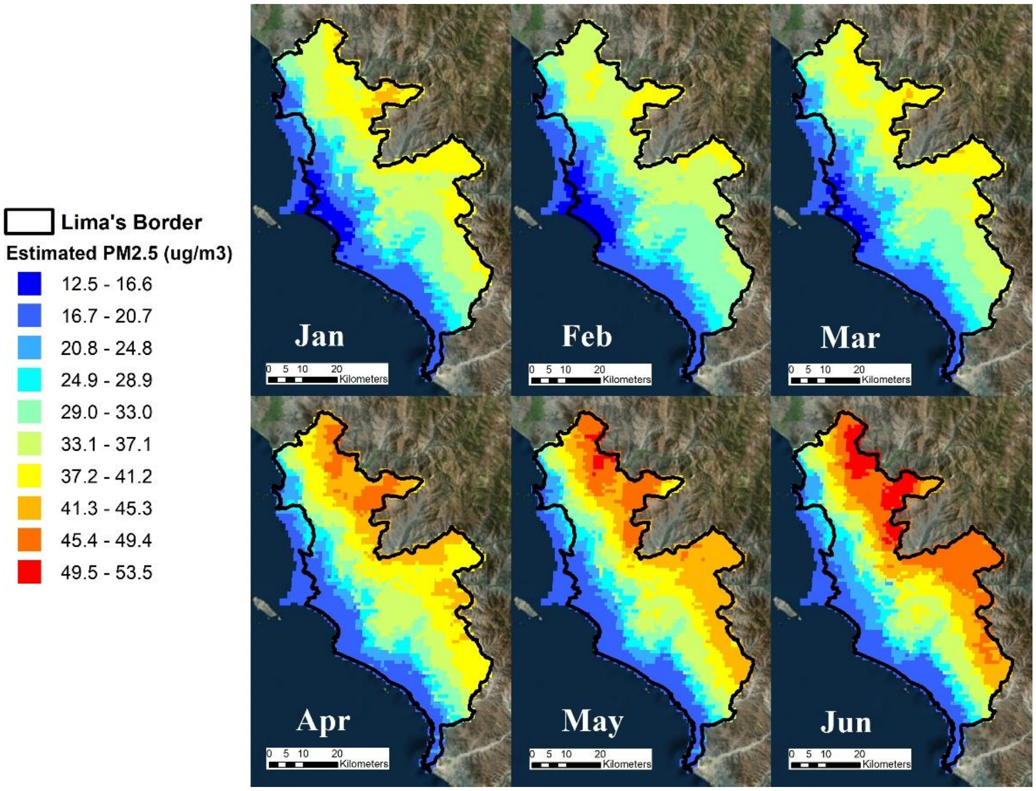


Figure S2. Histograms of each predictor variable.

In Figure S3, we show the monthly mean prediction maps of $PM_{2.5}$ in μg/m^3 for 2015. Overall, concentrations of $PM_{2.5}$ spatially increases from the month of April and peaks during June before decreasing to December, aggreeing with the monthly mean estimates from the ground monitors.



CHAPTER 1B

The association between asthma emergency department visits and satellite-derived $PM_{2.5}$ in Lima, Peru

Bryan N. Vu, Vilma Tapia, Stefanie Ebelt, Gustavo F. Gonzales, Yang Liu, Kyle Steenland

Vu BN, Tapia V, Ebelt S, Gonzales GF, Liu Y, Steenland K. The association between asthma emergency department visits and satellite-derived $PM_{2.5}$ in Lima, Peru. *Environ Res*. 2021 Aug;199:111226. doi: 10.1016/j.envres.2021.111226. Epub 2021 May 4. PMID: 33957138; PMCID: PMC8195863.

ABSTRACT

Background: Asthma affects millions of people worldwide. Lima, Peru is one of the most polluted cities in the Americas but has insufficient ground $PM_{2.5}$ (particulate matter that are 2.5 microns or less in diameter) measurements to conduct epidemiologic studies regarding air pollution. $PM_{2.5}$ estimates from a satellite-driven model have recently been made, enabling a study between asthma and $PM_{2.5}$.

Objective: We conducted a daily time-series analysis to determine the association between asthma emergency department (ED) visits and estimated ambient $PM_{2.5}$ levels in Lima, Peru from 2010 to 2016.

Methods: We used Poisson generalized linear models to regress aggregated counts of asthma on district-level population weighted $PM_{2.5}$. Indicator variables for hospitals, districts, and day of week were included to account for spatial and temporal autocorrelation while assessing same day, previous day, day before previous and average across all 3-day exposures. We also included temperature and humidity to account for meteorology and used dichotomous percent poverty and gender variables to assess effect modification.

Results: There were 103,974 cases of asthma ED visits during the study period across 39 districts in Lima. We found a 3.7% (95% CI: 1.7%-5.8%) increase in ED visits for every interquartile range (IQR, 6.02 $\mu g/m^3$) increase in $PM_{2.5}$ same day exposure with no age stratification. For the 0 to 18 years age group, we found a 4.5% (95% CI: 2.2%-6.8%) increase in ED visits for every IQR increase in $PM_{2.5}$ same day exposure. For the 19 to 64 years age group, we found a 6.0% (95% CI: 1.0%-11.0%) increase in ED visits for every IQR in average 3-day exposure. For the 65 years and up age group, we found a 16.0% (95% CI: 7.0%-24.0%) decrease in ED visits for every IQR increase in $PM_{2.5}$ average 3-day exposure, although the number of visits in this age group was low (4,488). We found no effect modification by SES or gender.

Discussion: Results from this study provide additional literature on use of satellite-driven exposure estimates in time-series analyses and evidence for the association between $PM_{2.5}$ and asthma in a low- and middle-income (LMIC) country.

KEYWORDS

$PM_{2.5}$; asthma; time-series; ED visits; remote sensing

INTRODUCTION

According to the 2018 Global Asthma Report (GBA), approximately 339 million people are living with asthma worldwide [1]. Asthma results in not only premature deaths but also reduced quality of life in people of all ages [2]. Asthma is a chronic disease that affects the airways, especially those of children and the elderly due to early-life development of lung function and subsequent decline in function as age increases. Since the comparison of prevalence and trends of asthma between countries requires large-scale surveys that have not been implemented since the early 2000s, much of the statistics on the burden of asthma are provided by the GBA report. The GBA report also notes that asthma prevalence has been increasing in the past few decades by as much as 50% per decade [3, 4]. Emergency department (ED) visits due to asthma exacerbation incur both direct and indirect economic costs including diagnostic tests, medication, and work and school days lost [5]. Although estimates are not available for many of the developing countries, the economic burden of asthma in the United States is between \$150 to \$3,000 in direct costs per patient, totaling more than \$56 billion annually [6]. Indirect costs, including lost pay from sickness and lost work output from missed school and work days, total \$3 billion between 2008-2013 in the U.S. [6]. Air pollution, including $PM_{2.5}$ (particulate matter with an aerodynamic diameter of 2.5 μm or less), has been shown to be associated with many cardiovascular and respiratory diseases. Furthermore, epidemiologic studies have indicated that exposure to $PM_{2.5}$ may exacerbate asthmatic symptoms [7-9]. Moreover, past studies on the association between air pollution and asthma have relied on ground monitor measurements, and more recently, on modeled estimates that are more spatially and temporally resolved [10-12].

To date, studies on the association between $PM_{2.5}$ and asthma were largely conducted in developed countries with sufficient numbers of daily ground monitoring measurements of $PM_{2.5}$ [12-15]. However, results from these studies may not provide sufficient guidance for low- and middle-income countries (LMIC) where ground monitors are scarce and the composition and concentrations of $PM_{2.5}$ may differ. Lima, Peru is the third most populous and one of the most polluted cities in the Americas. Lima’s air pollution is largely driven by an aging vehicular fleet in the urban center [16, 17]. Furthermore, particulate matter in Lima may be comprised mainly of black or elemental carbon, nitrogen oxide (NOx), and carbon dioxide from biomass burning, and diesel and gasoline combustion [18]. Conversely, composition of $PM_{2.5}$ in developed countries may contain more sulfur due to power generation and industrial sources. $PM_{2.5}$ in developed countries may also have reduced NOx and carbon dioxide due to newer vehicular engines that burn fossil fuels more cleanly [18]. As such, $PM_{2.5}$ concentrations in developed countries tend to be lower than in LMICs. A study by Silva et al. reported an annual average $PM_{2.5}$ of 26 μg/m^3 in Lima from 2010 to 2015, which exceeds the World Health Organization's (WHO) annual guidelines of 10 μg/m^3 [19]. Sparse daily monitoring data in Lima between March 2014 through December 2016 indicate that at least one in ten $PM_{2.5}$ monitors exceeds WHO daily guidelines of 25 μg/m^3 on 93% of those days. Furthermore, Lima consistently ranks among the top five most polluted urban centers in South America [16]. The spatial distribution of $PM_{2.5}$ in Lima is impacted by local wind conditions. Air pollution generated in more urban districts near the coast is pushed and trapped against the Andes Mountains in the east by winds blowing from the west. These high $PM_{2.5}$ levels in conjunction with the meteorological and topographical characteristics of Lima pose a major public health threat and warrant further investigations on air pollution and adverse health outcomes in Lima, Peru.

As one of the most rapidly developing urban centers in South America, one-third of the population of Peru resides in Lima. Although Peru has a decentralized healthcare system and 60% of Peruvians have free medical coverage maintained by the Ministry of Health (MINSA), access to healthcare may be hindered by the large gap in health status between the poor and the rich [20]. Recent studies suggests that the asthma prevalence among children and adolescents in Lima hovers around 13% while other studies indicate asthma prevalence as high as 19.6% for the entire Lima population [21]. Yet, limited monitor measurements have made epidemiologic studies of air pollution in Lima difficult, and there exist few studies pertaining to asthma and air pollution in Lima. There is one prior cross-sectional study estimating the impact of traffic flow on the prevalence of asthma among schoolchildren, and one cohort study over an 8-month period in one neighborhood of Lima [22, 23]. Both studies found a significant association between asthma prevalence and increased exposure to $PM_{2.5}$; however, these studies are only representative of relatively short timeframes.

Recently, Vu et al. developed a machine learning satellite-driven model that estimated daily $PM_{2.5}$ at 1 km^2 spatial resolution, enabling the possibility of conducting time-series analyses which requires daily estimates of $PM_{2.5}$ [24]. The time-series analysis has several advantages including the ability to assess if short-term temporal variation in the exposure of interest is associated with changes in the outcome of interest [25]. The newly developed exposure model provides daily $PM_{2.5}$ estimates from 2010 through 2016, enabling studies with a longer study period, a finer spatial resolution, a larger population size, and stronger statistical power. Several studies have begun to utilize the satellite-derived $PM_{2.5}$ estimates to study health effects in Lima [7, 26, 27]. A time-series conducted by Davila et al. found a significant positive association between $PM_{2.5}$ and acute lower respiratory infections, pneumonia, and acute bronchiolitis/asthma in outpatient clinic visits, across different age groups in children up to age five between 2011-2015 [26]. Davila et al.'s study considered asthma together with acute bronchiolitis, was conducted using weekly visits as the outcome, and only examined children under age five. Studies by Tapia et al. found significant associations between $PM_{2.5}$ and cardiorespiratory outcomes; however, none focused on asthma morbidity [7, 27]. Here, we conduct daily time-series analyses to determine the association between counts of asthma ED visits from nine Lima hospitals and estimated ambient $PM_{2.5}$ levels in Lima, Peru from 2010 to 2016 across all ages.

METHODS

Study Domain

Lima is nestled 154 meters above sea level between the Pacific Ocean in the west and the Andes Mountains in the east. Home to about 10 million inhabitants (30% of Peru's entire population), Lima's air pollution stems from an aging vehicular fleet, biomass burning, and distinct topography. The study domain includes the 43 districts within the province of Lima as well as the Seaport of Callao, and is divided into five different zones (North, South, East, West, and Central) (Figure 1). Since Vu et al.'s model was calibrated to ground stations that are all located below 375 meters in altitude, extrapolation of $PM_{2.5}$ levels to locations above this height may contain large uncertainties. Thus, four districts (Carabayllo, Chaclacayo, Cienguilla, and Lurigancho) with an average altitude higher than 570 meters were excluded from this study; however, these districts only represented 4% of the total population of the study domain.

Satellite-derived $PM_{2.5}$ Estimates

Ground $PM_{2.5}$ measurements exist from ten monitors in the Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI) network, maintained by Peru's Ministry of Environment from April 2014 through December 2016. Ground measurements also exist from six Johns Hopkins University monitoring sites from November 2011 through March 2013. However, the total number of daily measurements only account for 19% of the days from March 2010 through December 2016 (our study period) where any monitor recorded a measurement on a given day. This results in an ineffective coverage of both spatial and temporal variability of $PM_{2.5}$ and therefore would not be sufficient for a time-series analysis. Consequently, daily ambient $PM_{2.5}$ concentrations from March 2010 to December 2016 were estimated by a random forest (RF) model developed by Vu et al. The RF model calibrated satellite aerosol optical depth (AOD), meteorological parameters from chemical transport models, and land use variables with available ground measurements from the two monitoring networks (SENAMHI and Johns Hopkins) [24].

Utilizing estimates derived from an advanced exposure model allows two advantages. First, there is extensive temporal coverage of daily estimates from March 2010 through December 2016, enabling a larger sample size and stronger statistical power to detect any association. Second, the modeled $PM_{2.5}$ estimates are also spatially resolved, with one estimate for every 1 km^2 of the study domain. Such specifications likely reduce many of the uncertainties and biases associated with traditional methods which assign single levels of $PM_{2.5}$ to all of Lima based on a limited number of monitors. Vu et al.'s RF model achieved a model training R^2 and cross-validation (CV) R^2 and root mean square prediction error (RMSE) of 0.70 (5.95 μg/m^3 model training RMSE and 5.97 μg/m^3 CV RMSE). These results indicate good fit and stable prediction capabilities. Since only the district of residence was available as the address for each participant in the health data, the $PM_{2.5}$ exposure estimates were aggregated to the district level. Each daily 1 km^2 $PM_{2.5}$ estimate was multiplied by the population density of the corresponding 1 km^2 LandScan^{TM} population grid. These grids were then summed within each district and divided by the total population of that district to derive a daily population-weighted district-average $PM_{2.5}$ estimate. In total, the exposure model provided $PM_{2.5}$ estimates for 2,236 (91%) of the 2,465 days during the study period.

Health Data

Electronic patient records for ED visits were obtained from nine large public hospitals belonging to the Ministry of Health of Peru (MINSA) in Lima during the period of March 2010 through December 2016. Information for each ED visit included the patient's primary International Classification of Disease 10th Revision (ICD10) diagnosis code, their district of residence, age, and gender. Asthma cases included ICD10 codes J45-J46. Validity of the electronic patient visit records were evaluated by comparing a random sample of 100 electronic medical records with hardcopies of medical history at each hospital. Since personal addresses within districts were not available in the electronic records, the number of visits were aggregated to the district level for each day.

Time-series Analysis

This study utilized a time-series approach through Poisson generalized linear models (GLMs) to estimate the associations between daily district-level $PM_{2.5}$ and counts of ED visits for asthma. An advantage of a time-series approach is that only time-varying variables can be assessed as confounders. We assessed the effects of same day (lag 0), previous day (lag 1), the day before previous (lag 2), and an average across all 3 prior days (3-day avg.) $PM_{2.5}$ exposure in separate models. Additionally, a categorical variable for district was added to control for spatial variability and allows the regression to be based only on temporal effects. The district indicator variables also control for spatial autocorrelation in the baseline asthma ED visits across all the districts as well as unmeasured factors that may vary between districts [18]. We also included indicator variables for each day of the week to account for daily fluctuations in $PM_{2.5}$, and added parametric cubic splines with monthly knots to control for long-term trends in ER visit rates. We controlled for meteorology using same day mean temperature (in linear, quadratic, and cubic forms), and same day mean relative humidity. We also included indicator variables for each hospital to indicate if that hospital contributed any cases for each day. Count data is usually over or under dispersed, meaning the mean and variance of the Poisson counts do not equal each other. Since the variance of the asthma ED visits is larger than the mean, we added a dispersion parameter (pscale in SAS) to account for overdispersion in all of our models.

We assessed effect modification by socioeconomic status (SES), gender, and age in our models. To assess interaction by SES, we obtained estimates of the percent of households above the poverty level for each district from the National Institute of Statistics and Informatics of Peru (INEI). A dichotomous variable was created to indicate whether districts were above or below the median poverty percentage (12.9%). To assess interaction by age, we stratified the asthma ED visit counts into three age strata: 1) 0 to 18 years, 2) 19 to 64 years, and 3) 65 years and above. We ran a separate model for each age-specific group, and effect modification would be indicated by differences in the observed association of $PM_{2.5}$ between the different age strata. Since 62% of the participants age 18 years and under were also age 6 years and under, we ran a separate model for the participants age 6 years and under as a sensitivity analysis. To assess gender as an effect modifier, we aggregated the number of asthma ED visits by district by gender and included gender as a dichotomous variable. We also aggregated the asthma ED visit counts by gender and district for all three age groups as a sensitivity analysis. Finally, we conducted sensitivity analyses using Zero-Inflated Poisson (ZIP) models to ensure robust results in the 65 years and above age group when many districts carried zero counts due to low ED visits (206 days out of 1,845 days contained zero counts in all districts). We conducted the ZIP model sensitivity analyses on all age groups to ensure results did not deviate from our original models. Data processing was conducted in R$^{©}$ (version 3.6.2) and model analyses were done using SAS v9.4 (SAS Institute Inc., Cary, NC, USA). Best model fit was determined via the lowest Akaike Information Criterion (AIC) value within each age strata.

RESULTS

In total, there were 103,974 cases of asthma ED visits in Lima from March 2010 through December 2016. Table S1 in the supplemental shows the number of asthma ED visits for each district in Lima along with the mean and standard deviation of the estimated population-weighted $PM_{2.5}$ during the entire study period. Table S1 also includes the median and interquartile range of the estimated population-weighted $PM_{2.5}$. Figure 2 shows the daily Lima-wide average $PM_{2.5}$ between 2010 and 2016 as well as the daily number of asthma ED visits during the same period. In general, the annual trends of $PM_{2.5}$ are usually high during the winter months of June-August and low during the summer months of December-March of each year. Asthma ED visits, on the other hand, follow a biannual pattern and peak during the fall and spring months.

Daily statistics stratified by districts show varying results. Figure 3 shows the time-series of mean daily $PM_{2.5}$ levels in Lima District and Ate District for the entire study period. Both districts show distinct seasonal patterns of high values in the summer months and lower values in the winter months. However, Lima District is closer to the coast and therefore has a smaller range compared to Ate District located further inland towards the Andes Mountains. Figure 4 shows the time-series of daily counts of asthma for all ages in Lima District and Ate District. Unlike $PM_{2.5}$, there does not seem to be a marked trend in seasonal or annual patterns of asthma ED visits in Lima during the study period, although in Lima District, asthma does appear somewhat higher in the summer.

Validation of patient electronic medical records against paper copies indicate that the date of the patient's emergency visit had the highest matching rate at 94%. Records with mismatched dates typically disagreed by one to two days, with the hardcopy record date occurring earlier than the date in the digital record. In contrast, ICD10 diagnosis code had the lowest matching rate at 86%. The discrepancy in ICD10 code matching rate was largely due to ambiguous diagnoses when patients presented a wide variety of symptoms during the visit and were subsequently hospitalized. These non-definitive diagnoses were recorded in the hard copy's emergency room records, but hospital discharge records matched the electronic patient records that we used. Thus, for most cases with discrepancy in diagnosis, we felt the electronic records are likely to be correct. Mismatch between electronic and hardcopy patient records for district, age, and sex, was a result of missing data in the hardcopy of the patient's records.

Table 1 shows the number of asthma ED visits stratified by sex, age groups, and region for the entire study period. The vast majority of the visits were from young people ages 18 years and below (73.0%) compared to 22.7% for adults ages 19 to 64 years and 4.3% for people ages 65 and older. For all cases of asthma ED visits across all age groups, same day (lag 0) $PM_{2.5}$ levels produced the best fitting model between daily air pollution and ED visits compared to models assessing lag 1, lag 2, and 3-day average effects. When stratified by age groups; however, we found that lag 0 produced the best fitting model in the 0 to 18 years age group. For the 19 to 64 years and the 65 years and older age groups, the 3-day average produced the best fitting model. All model results are shown in Table 2 in the form of rate ratios (RR) and their 95% confidence intervals for an interquartile range (IQR) increase of 6.02 $\mu g/m^3$ of $PM_{2.5}$ exposure. The overdispersion parameter (pscale) ranged from 1.05 to 1.13 indicating that asthma ED counts were not overly dispersed.

We found that an IQR increase in $PM_{2.5}$ exposure was associated with a 3.7% increase in asthma ED visits. For the 0 to 18 years age group, we found a corresponding 4.5% increase in asthma ED visits and for the 19 to 64 years age group, a 6.0% increase for every IQR increase in $PM_{2.5}$. In the 65 years and above age group, we found a 16% decrease in asthma ED visits for every IQR increase in $PM_{2.5}$ exposure. However, these results may be imprecise since the total number of asthma ED visits in this age group for the entire study period is relatively low (4,488 across 7 years). A sensitivity analysis for the 65 years and above age group using the ZIP model yielded similar results with a 18% decrease in asthma ED visits for every IQR increase in $PM_{2.5}$ exposure. We found no effect modification by SES for all ages combined or specific age groups with p-values ranging from 0.07 (18 years and under age group) to 0.29 (19 to 64 years age group).

Similarly, we found no effect modification by gender for all ages combined or specific age groups with p-values ranging from 0.22 (all ages) to 0.32 (18 years and under age group).

DISCUSSION

Although many South American countries have set up official air quality monitoring stations, only 104 cities record measurements for PM_{10} and 57 cities record measurements for $PM_{2.5}$ [28]. $PM_{2.5}$ is smaller and can be transported deeper into lung tissue leading to deadlier adverse health consequences. Many of the epidemiologic studies pertaining to air pollution and asthma have centered on ozone and children under 18 since children's immune systems and lungs are not fully developed at the start of exposure [29, 30]. Additionally, studies of particulate matter in South America largely focus on PM_{10} since measurements are more readily available, or are based on estimated particulate matter using chemical models such as CATT-BRAMS [31-33]. Before the development of the $PM_{2.5}$ exposure model, epidemiologic studies in Lima were hindered by sparse ground and mobile-based measurements and often utilized cross-sectional study designs that make causal inference difficult [22, 34]. One of the main strengths of the present study is the use of daily $PM_{2.5}$ estimates at 1 km^2 spatial resolution, which enables an investigation of the association between asthma and ambient air pollution in not only children but also adults using a time-series approach.

This study found moderately strong associations between ambient $PM_{2.5}$ exposure and asthma ED visits. The mean population-weighted $PM_{2.5}$ in Lima was 21.0 μg/m^3, while mean population-weighted $PM_{2.5}$ for individual districts ranged from 16.6 μg/m^3 in Magdalena del Mar (Central Lima) to 32.3 μg/m^3 in San Juan de Lurigancho (East Lima). This distribution is largely due to strong prevailing winds from the Pacific Ocean. Coastal winds drive the air pollution generated in the urban central region of Lima to the east and northeast regions against the Andes Mountains. Lima's topography and meteorological conditions lead to a thermal inversion layer that traps air pollution and reduces its dispersion resulting in lower $PM_{2.5}$ concentrations near the coast and rising concentrations towards the east. The heterogeneity in the spatial distribution of $PM_{2.5}$ among the districts in Lima indicates that district-specific $PM_{2.5}$ estimates should be used instead of Lima-wide averages. We conducted a sensitivity analysis using Lima-wide $PM_{2.5}$ estimates compared to district-specific estimates. Results from this sensitivity analysis indicate that the district-specific associations were stronger, suggesting lower exposure measurement error than the traditional assignment of central or nearest monitor measurement or city-wide averages. We also assessed different lags including same-day, previous day, day before previous, and an average of the three days since studies have shown that the timing of exacerbation of asthma by air pollution varies across ages and geographic locations [35-37].

We found a 3.7% (95% CI: 1.7%, 5.8%) increase in asthma ED visits per IQR increase in $PM_{2.5}$ exposure in all ages. These results are consistent with previous studies that also utilized a time-series approach with a similar IQR range [38]. A meta-analysis conducted by Zheng et al. looking at 37 studies found a 2.3% (95% CI: 1.5%, 3.1%) increase in asthma cases for a 10 $\mu g/m^3$ increase in $PM_{2.5}$ exposure [39]. All the studies in that meta-analysis were conducted in developed countries (20 in N. America, 7 in Europe, 7 in Asia, and 3 in Australia). These authors reported similar results for all age groups [39].

We also found a 4.5% (95% CI: 2.2%, 6.8%) increase in asthma ED visits in people ages 18 years and under for every IQR change in same day (lag 0) $PM_{2.5}$ exposure. Our findings are consistent with studies conducted in cities in developed countries. A prior multi-city study conducted in Dallas, St. Louis, and Atlanta in the United States reported a 2.0% (95% CI: 1.0%, 4.0%) increase in asthma ED visits per 8.0 $\mu g/m^3$ increase in $PM_{2.5}$ exposure in 5 to 18 year-olds [40]. Additionally, Gleason et al. found a 3.0% (95% CI: 2.0%, 4.0%) increase in asthma ED visits in children ages 3 to 17 years living in New Jersey, U.S. per 10 $\mu g/m^3$ increase in same day $PM_{2.5}$ exposure [41]. It is expected that same day exposure would result in the best model fit for the 0 to 18 years age group as well as the overall model since 73% of the total sample size consists of children. Studies have shown that the effects of $PM_{2.5}$ on asthma exacerbation may last up to six days, and that the most significant effects happen during same-day or day-before exposures [42]. The greater effect of $PM_{2.5}$ on asthma exacerbation in children may partly be explained by the biological mechanisms including inflammation in the alveolar region of the lung caused by deposition of finer particles [43].

We found a 6.0% (95% CI: 1.0%, 11.0%) increase in ED visits per IQR change in $PM_{2.5}$ for adults ages 19 to 64 years based on the previous 3-day average. Conversely, $PM_{2.5}$ appeared to be protective for people ages 65 years and above using exposure based on the previous 3-day average. Since the 65 years and above age group had a very low sample size (4,488 cases across 1,845 study period days and 39 districts), and a large number of days contained no cases, sensitivity analyses including ZIP models were conducted to ensure robust model performance. Results from the ZIP analyses were similar although less pronounced. The protective effect in the oldest age group contradicts previous studies including one by Park et al. that reported a cumulative risk greater than one for each lag strata from zero to 15 days per 10 μg/m^3 of $PM_{2.5}$ exposure in people above 65 years of age. It is possible that the 65 years and above population are less likely to present to the ER in general than other ages which may hinder the assessment of the effects of $PM_{2.5}$ on this age group. Another possibility is that people 65 years and above may have a less responsive immune system and more commonly have allergic asthma. Finally, another possibility is that misclassification of asthma is more common in the 65 years and above age group. Nonetheless, results for the 65 years and above population in this study should be interpreted with caution and that further studies are needed to investigate how to parse out asthma from other comorbidities often associated with the older adult population [44].

The associations found in this study pertain to $PM_{2.5}$ levels at or well below the Peruvian permitted 24h and annual standards, 50 μg/m^3 and 25 μg/m^3, respectively. The Peruvian standards are also higher than the 24h and annual standards set by the WHO, 25 μg/m^3 and 10 μg/m^3, respectively [45]. These two sets of standards highlight the differences in standard levels permissible in a LMIC compared to developed countries. Results from this study are consistent with others previously published that utilized the new $PM_{2.5}$ exposure estimates. Davila Cordova et al. found a 10% increase in weekly asthma outpatient visits in children under 5 years of age from 2001 to 2015 for every 7.1 μg/m^3 increase in $PM_{2.5}$ exposure [26]. Tapia et al. found a 4% increase in all respiratory ED visits for every 6.1 μg/m^3 increase in $PM_{2.5}$ exposure between 2010 through 2016 [7].

A strength of this study is the use of comprehensive health data from MINSA. We estimate that our population from nine large public hospitals represents about half the population of Lima. We think it is unlikely that the relationship between air pollution and ER visits for this part of the population differs substantially from the rest of Lima, although we have no data to confirm this. Another strength of this study is the ability to obtain large datasets for both the exposure and the outcome of interest which cover a long study period, something made possible by the ecologic design of our study. While we lack some individual level information, such that our study is potentially susceptible to confounding and ecologic bias, the time series design compares the same population to itself over time, lessening these concerns.

There are several limitations in our study. We did not have data to further stratified asthma by phenotypes (e.g. allergic, non-allergic, severe, etc.) which may differ in their association with $PM_{2.5}$. In addition, we did not have data on ozone and NOx (nitric oxide, NO, and nitrogen dioxide, NO_2). Ozone and NOx have been shown to be significantly associated with asthma, independent of the effects of $PM_{2.5}$. NOx is one of the major chemicals emitted from vehicle exhaust and is a precursor of ozone and the absence of both these compounds are a major limitation to this study. Body mass index (BMI) has also been shown to be associated with both outdoor air pollution exposure and asthma exacerbation. However, information on BMI for each ED visit case was not available and the availability of this information may modify the association between air pollution and asthma ED visits in Lima. Although SES is an effect modifier in the relationship between $PM_{2.5}$ and asthma in the literature, this study found no such relationship. One possibility might be due to having only limited ecological data on the percentage of the population living in poverty in each district. Furthermore, the effects of gender as an effect modifier on the association between air pollution and asthma has been unclear and results from our study suggests that gender is not an effect modifier in the Lima population. Lastly, another limitation of this study is that health records provided only district-level information on residence, not the exact address of residence, and we were unable to fully utilize the benefits of the 1 km^2 $PM_{2.5}$ exposure model. The aggregation of the $PM_{2.5}$ estimates to the district level may further pull the associations toward the null due to misclassification of exposure.

Past studies have indicated that asthma exacerbation may also be moderated by both indoor and outdoor $PM_{2.5}$ concentrations [46]. However, since Lima is a city with moderate climate, and windows are often open, we might expect indoor and outdoor air pollution to be similar [47]. Indoor air can have much higher $PM_{2.5}$ levels in Peru, but this occurs in the countryside where biomass fuel is used for cooking, which is very uncommon in Lima [48]. Although we do not have indoor air pollution measurements in Lima, differences in outdoor and indoor $PM_{2.5}$ concentrations may be assessed through a tracer element such as sulfur, which has few indoor generating sources [49].

CONCLUSIONS

This study is the first to utilize a time-series approach to investigate the association between satellite-derived $PM_{2.5}$ estimates and asthma ED visits in Lima, Peru. We found that short-term exposure to ambient $PM_{2.5}$ is associated with moderate increases in asthma ED visits in children under 19 years of age and among adults ages 19 to 64 years. Results from this study provide new support for such associations in the literature pertaining to LMIC, and also provides evidence for the Peruvian government to investigate the need to lower the current $PM_{2.5}$ standards in Lima. While we do not know whether lowering Peruvian standards for $PM_{2.5}$ would result in different findings for the exposure-response of asthma and $PM_{2.5}$, our findings indicate $PM_{2.5}$ increases asthma risk. Precautionary principle would argue for lowering $PM_{2.5}$ levels to lower asthma risk as well as risks for other health endpoints linked to $PM_{2.5}$ in Lima including ER visits for all cardiorespiratory diseases, overall mortality, and reproductive outcomes.

Future studies in LMICs should attempt to obtain more detailed address information for participants and improved data on potential effect modifiers like SES and BMI. They should also seek to incorporate exposure to ozone and NOx. In Lima specifically, information on the use of emergency rooms for the older adults with asthma and possible mid-diagnoses may help further investigate the seemingly protective effect of $PM_{2.5}$ in this study. Although the generalizability from our results to other LMICs are uncertain, the methods from this study hopefully provide guidance on how one can conduct epidemiologic studies in developing countries with high air pollution exposure but limited ground monitoring measurements.

ACKNOWLEDGEMENT

Research reported in this publication was supported by the NIH Fogarty International Center, National Institutes of Environmental Health Sciences, National Cancer Institute, Centers for Disease Control and the NIH under Award Number U01 TW0101 07 and 5T32ES12870. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The work of Vu and Liu were partially supported by NASA Applied Science Program (Grant # 80NSSC19K0191 and #NNX16AQ28G). This research was also supported by the HERCULES Center Pilot Project Program.

REFERENCE

1. Network, G.A., *The Global Asthma Report 2018*. 2018: Auckland, New Zealand.
2. Anenberg Susan, C., et al., *Estimates of the Global Burden of Ambient PM2.5, Ozone, and NO2 on Asthma Incidence and Emergency Room Visits.* Environmental Health Perspectives, 2018. **126**(10): p. 107004.
3. Masoli, M., et al., *The global burden of asthma: executive summary of the GINA Dissemination Committee report.* Allergy, 2004. **59**(5): p. 469-78.
4. Braman, S.S., *The global burden of asthma.* Chest, 2006. **130**(1 Suppl): p. 4s-12s.
5. Zhang, F.Y., et al., *Estimation of the Effects of Air Pollution on Hospitalization Expenditures for Asthma.* International Journal of Health Services, 2020. **50**(1): p. 100-109.
6. Nurmagambetov, T., R. Kuwahara, and P. Garbe, *The cost of asthma in the United States.* America Thoracic Society, 2017.
7. Tapia, V., et al., *Time-series analysis of ambient PM2.5 and cardiorespiratory emergency room visits in Lima, Peru during 2010–2016.* Journal of Exposure Science & Environmental Epidemiology, 2019.
8. Vardoulakis, S. and N. Osborne, *Air pollution and asthma.* Archives of Disease in Childhood, 2018. **103**(9): p. 813-+.
9. Bouazza, N., et al., *Fine particulate pollution and asthma exacerbations.* Archives of Disease in Childhood, 2018. **103**(9): p. 828-831.
10. Alhanti, B.A., et al., *Ambient air pollution and emergency department visits for asthma: a multi-city assessment of effect modification by age.* Journal of Exposure Science and Environmental Epidemiology, 2016. **26**(2): p. 180-188.
11. Bose, S., et al., *Association of traffic air pollution and rhinitis quality of life in Peruvian children with asthma.* Plos One, 2018. **13**(3): p. 13.
12. Chang, H.H., et al., *Time-series analysis of satellite-derived fine particulate matter pollution and asthma morbidity in Jackson, MS.* Environmental Monitoring and Assessment, 2019. **191**: p. 10.
13. Zuo, B.Q., et al., *Associations between short-term exposure to fine particulate matter and acute exacerbation of asthma in Yancheng, China.* Chemosphere, 2019. **237**: p. 6.
14. Strosnider, H.M., et al., *Age-Specific Associations of Ozone and Fine Particulate Matter with Respiratory Emergency Department Visits in the United States.* American Journal of Respiratory and Critical Care Medicine, 2019. **199**(7): p. 882-890.
15. Abrams, J.Y., et al., *Associations between Ambient Fine Particulate Oxidative Potential and Cardiorespiratory Emergency Department Visits.* Environmental Health Perspectives, 2017. **125**(10): p. 9.
16. WHO (World health Organization). *WHO Global Urban Ambient Air Pollution Database*. 2016 August 25, 2017]; Available from: http://www.who.int/phe/health_topics/outdoorair/databases/cities/en/.
17. Fischer, K., *Improving Sustainable Development in Lima Through Public Transportation.* Perspectives on Business and Economics, 2017. **35**(Leveraging Peru's Economic Potential): p. 6.
18. Ventura, L.M.B., et al., *Chemical composition of fine particles (PM2.5): water-soluble organic fraction and trace metals.* Air Quality, Atmosphere & Health, 2017. **10**(7): p. 845-852.
19. Silva, J., et al., *Particulate matter levels in a South American megacity: the metropolitan area of Lima-Callao, Peru.* Environmental Monitoring and Assessment, 2017. **189**(12): p. 635.
20. Global Health Workforce Alliance, W. *Peru*. 2020 06/15/2020]; Available from: https://www.who.int/workforcealliance/countries/per/en/#:~:text=Peru%20has%20a%20decentralized%20health,together%20provide%20services%20to%20the.
21. Lai, C.K.W., et al., *Global variation in the prevalence and severity of asthma symptoms: Phase Three of the International Study of Asthma and Allergies in Childhood (ISAAC).* Thorax, 2009. **64**(6): p. 476.

22. Rivero, K.M.R., et al., *Effects Of Long-Term Exposure To Pm2.5 On Asthma Control In Children: Longitudinal Study In A Peri-Urban Community In Lima, Peru.* American Journal of Respiratory and Critical Care Medicine, 2016. **193**: p. 1.
23. Carbajal-Arroyo, L., et al., *Impact of traffic flow on the asthma prevalence among school children in Lima, Peru.* Journal of Asthma, 2007. **44**(3): p. 197-202.
24. Vu, B.N., et al., *Developing an Advanced PM(2.5) Exposure Model in Lima, Peru.* Remote sensing, 2019. **11**(6): p. 641.
25. Bhaskaran, K., et al., *Time series regression studies in environmental epidemiology.* International journal of epidemiology, 2013. **42**(4): p. 1187-1195.
26. Davila Cordova, J.E., et al., *Association of PM(2.5) concentration with health center outpatient visits for respiratory diseases of children under 5 years old in Lima, Peru.* Environmental health : a global access science source, 2020. **19**(1): p. 7-7.
27. Tapia, V., et al., *PM(2.5) exposure on daily cardio-respiratory mortality in Lima, Peru, from 2010 to 2016.* Environ Health, 2020. **19**(1): p. 63.
28. Riojas-Rodríguez, H., et al., *Air pollution management and control in Latin America and the Caribbean: implications for climate change.* Rev Panam Salud Publica, 2016. **40**(3): p. 150-159.
29. Schwartz, J., *Air pollution and children's health.* Pediatrics, 2004. **113**(4): p. 1037-1043.
30. Peters, J.M., et al., *A study of twelve southern California communities with differing levels and types of air pollution - II. Effects on pulmonary function.* American Journal of Respiratory and Critical Care Medicine, 1999. **159**(3): p. 768-775.
31. Cesar, A.C.G., L.F.C. Nascimento, and J.A.d. Carvalho, Jr., *Association between exposure to particulate matter and hospital admissions for respiratory disease in children.* Revista de saude publica, 2013. **47**(6): p. 1209-1212.
32. Tuan, T.S., T.S. Venâncio, and L.F. Nascimento, *Air pollutants and hospitalization due to pneumonia among children. An ecological time series study.* Sao Paulo Med J, 2015. **133**(5): p. 408-13.
33. Sousa, S.I.V., et al., *Short-term effects of air pollution on respiratory morbidity at Rio de Janeiro — Part II: Health assessment.* Environment International, 2012. **43**: p. 1-5.
34. Underhill, J.L., et al., *Association of Roadway Proximity with Indoor Air Pollution in a Peri-Urban Community in Lima, Peru.* International Journal of Environmental Research and Public Health, 2015. **12**(10).
35. Chien, L.-C., Y.-A. Chen, and H.-L. Yu, *Lagged Influence of Fine Particulate Matter and Geographic Disparities on Clinic Visits for Children's Asthma in Taiwan.* International journal of environmental research and public health, 2018. **15**(4): p. 829.
36. Schildcrout, J.S., et al., *Ambient Air Pollution and Asthma Exacerbations in Children: An Eight-City Analysis.* American Journal of Epidemiology, 2006. **164**(6): p. 505-517.
37. Rosenquist, N.A., et al., *Acute associations between PM2.5 and ozone concentrations and asthma exacerbations among patients with and without allergic comorbidities.* Journal of Exposure Science & Environmental Epidemiology, 2020. **30**(5): p. 795-804.
38. Castner, J., L. Guo, and Y. Yin, *Ambient air pollution and emergency department visits for asthma in Erie County, New York 2007–2012.* International Archives of Occupational and Environmental Health, 2018. **91**(2): p. 205-214.
39. Zheng, X.-y., et al., *Association between Air Pollutants and Asthma Emergency Room Visits and Hospital Admissions in Time Series Studies: A Systematic Review and Meta-Analysis.* PloS one, 2015. **10**(9): p. e0138146-e0138146.
40. Alhanti, B.A., et al., *Ambient air pollution and emergency department visits for asthma: a multi-city assessment of effect modification by age.* Journal of Exposure Science & Environmental Epidemiology, 2016. **26**(2): p. 180-188.
41. Gleason, J.A., L. Bielory, and J.A. Fagliano, *Associations between ozone, PM2.5, and four pollen types on emergency department pediatric asthma events during the warm season in New Jersey: A case-crossover study.* Environmental Research, 2014. **132**: p. 421-429.

42. Slaughter, J.C., et al., *Effects of ambient air pollution on symptom severity and medication use in children with asthma.* Ann Allergy Asthma Immunol, 2003. **91**(4): p. 346-53.

43. Anderson, P.J., J.D. Wilson, and F.C. Hiller, *Respiratory tract deposition of ultrafine particles in subjects with obstructive or restrictive lung disease.* Chest, 1990. **97**(5): p. 1115-20.

44. Gillman, A. and J.A. Douglass, *Asthma in the elderly.* Asia Pacific allergy, 2012. **2**(2): p. 101-108.

45. Bicentenario, D.O.D., *Approve Environmental Quality Standards (EQS) for Air and established Complementary Provisions*. 2017.

46. Jie, Y., et al., *Do indoor environments influence asthma and asthma-related symptoms among adults in homes? A review of the literature.* Journal of the Formosan Medical Association, 2011. **110**(9): p. 555-563.

47. Organization), W.W.H. *Background information on urban outdoor air pollution*. 2021 [cited 2021 February 26, 2021]; Available from: https://www.who.int/phe/health_topics/outdoorair/databases/background_information/en/.

48. Kephart, J.L., et al., *Indoor air pollution concentrations and cardiometabolic health across four diverse settings in Peru: a cross-sectional study.* Environmental Health, 2020. **19**(1): p. 59.

49. Habre, R., et al., *The effects of PM2.5 and its components from indoor and outdoor sources on cough and wheeze symptoms in asthmatic children.* Journal of Exposure Science & Environmental Epidemiology, 2014. **24**(4): p. 380-387.

CHAPTER 1B TABLES AND FIGURES

Table 1. Number of asthma ED visits by sex, age groups, and region for the entire study period.

Total	**103,974**
Sex	**n (%)**
Males	53,944 (51.9)
Females	50,030 (48.1)
Age Group	**n (%)**
0-18	75,917 (73.0)
19-64	23,569 (22.7)
65+	4,488 (4.3)
Region	**n (%)**
Central	28,207 (27.1)
East	19,958 (19.2)
North	22,863 (22.0)
South	20,457 (19.7)
West	12,489 (12.0)

Table 2. Comprehensive model results in Rate Ratios (RR) for an interquartile range (IQR) increase of 6.02 µg/m^3 of PM$_{2.5}$ for all cases of asthma ED visits and stratified by age groups along with the model AIC values.

All Cases	**RR (95% CI)**	**AIC**
Lag 0	1.04 (1.02, 1.06)	167,411.5
Lag 1	1.03 (1.01, 1.05)	167,419.5
Lag 2	1.02 (1.00, 1.04)	167,422.3
3-day avg.	1.04 (1.02, 1.06)	167,412.9
Aged 0-18 Years	**RR**	**AIC**
Lag 0	1.04 (1.02, 1.07)	143,285.5
Lag 1	1.03 (1.01, 1.06)	143,294.3
Lag 2	1.03 (1.01, 1.05)	143,294.6
3-day avg.	1.05 (1.02, 1.07)	143,287.1
Aged 19-64 Years	**RR**	**AIC**
Lag 0	1.05 (1.01, 1.10)	83,374.8
Lag 1	1.05 (1.00, 1.09)	83,376.0
Lag 2	1.02 (0.98, 1.07)	83,379.6
3-day avg.	1.06 (1.01, 1.11)	83,374.1
Aged 65+ Years	**RR**	**AIC**
Lag 0	0.86 (0.79, 0.95)	27,863.6
Lag 1	0.87 (0.80, 0.95)	27,864.6
Lag 2	0.87 (0.80, 0.95)	27,864.4
3-day avg.	0.84 (0.76, 0.93)	27,861.7

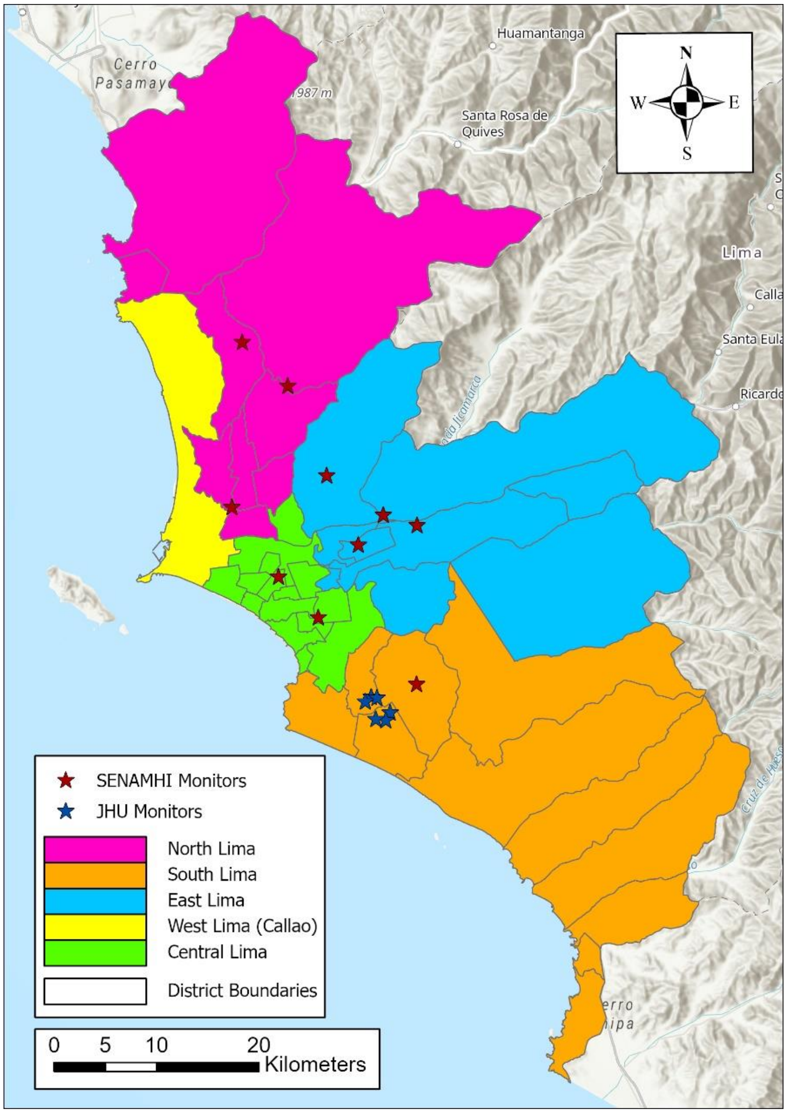


Figure 1. Study domain with the province of Lima and the Seaport of Callao divided into five different zones. The locations of the ten SENMAHI ground monitors and six Johns Hopkins University (JHU) monitor sites used to develop the exposure model that provided daily estimates for the present study are also included.

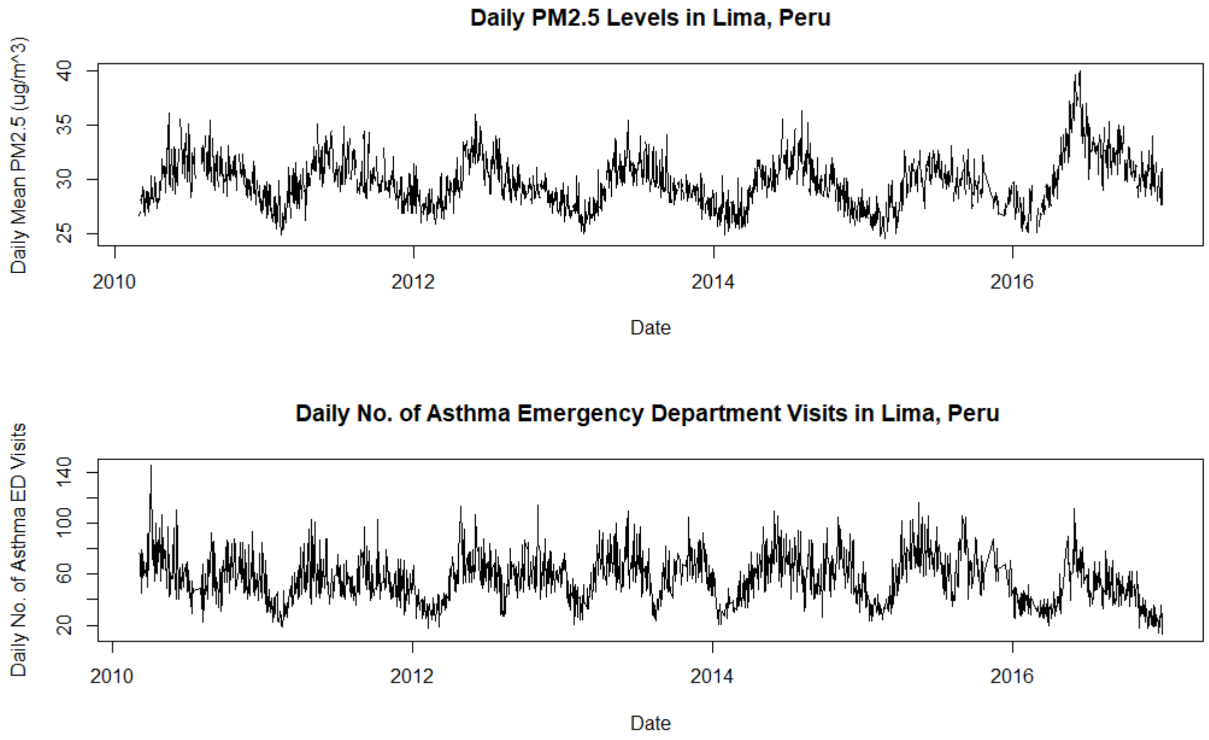


Figure 2. Daily Lima-wide population-weighted average $PM_{2.5}$ during the study period of March 2010 through December 2016. The bottom panel shows the total number of daily asthma ED visits in all districts during the same time period.

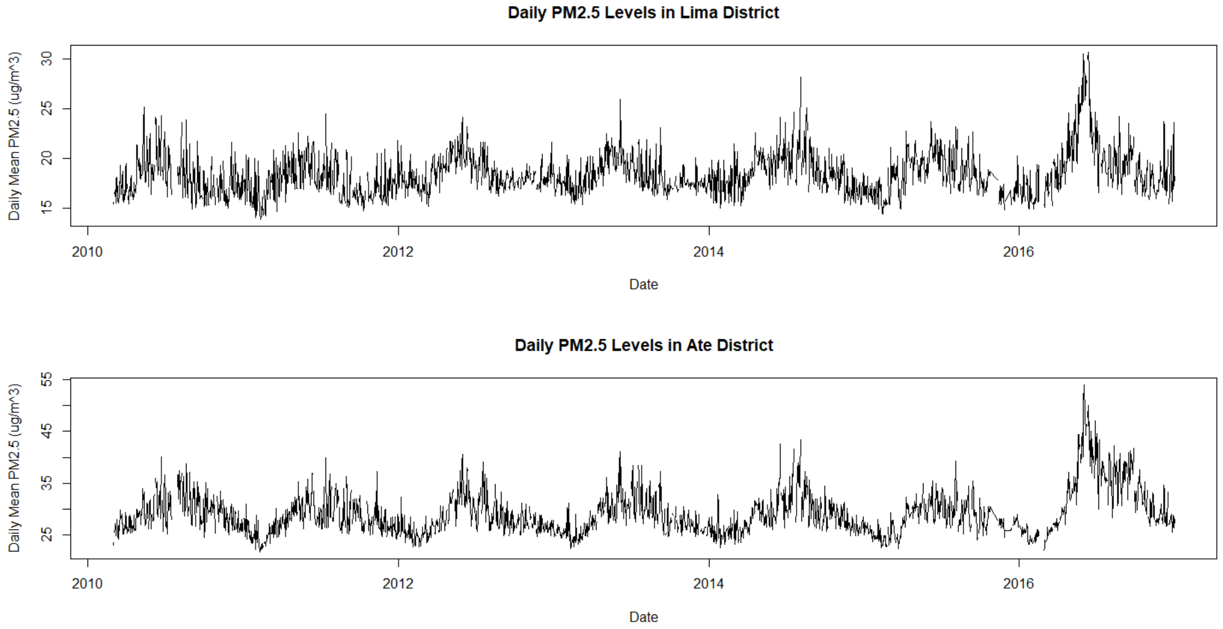


Figure 3. Daily population-weighted district-averaged PM$_{2.5}$ levels in Lima District (top) and Ate District (bottom) from March 2010 through December 2016. Lima District is closer to the coast and has a lower and smaller range compared to Ate District located further inland. Both districts show similar seasonal and annual trends with higher peaks during the winter and lower peaks in the summer.

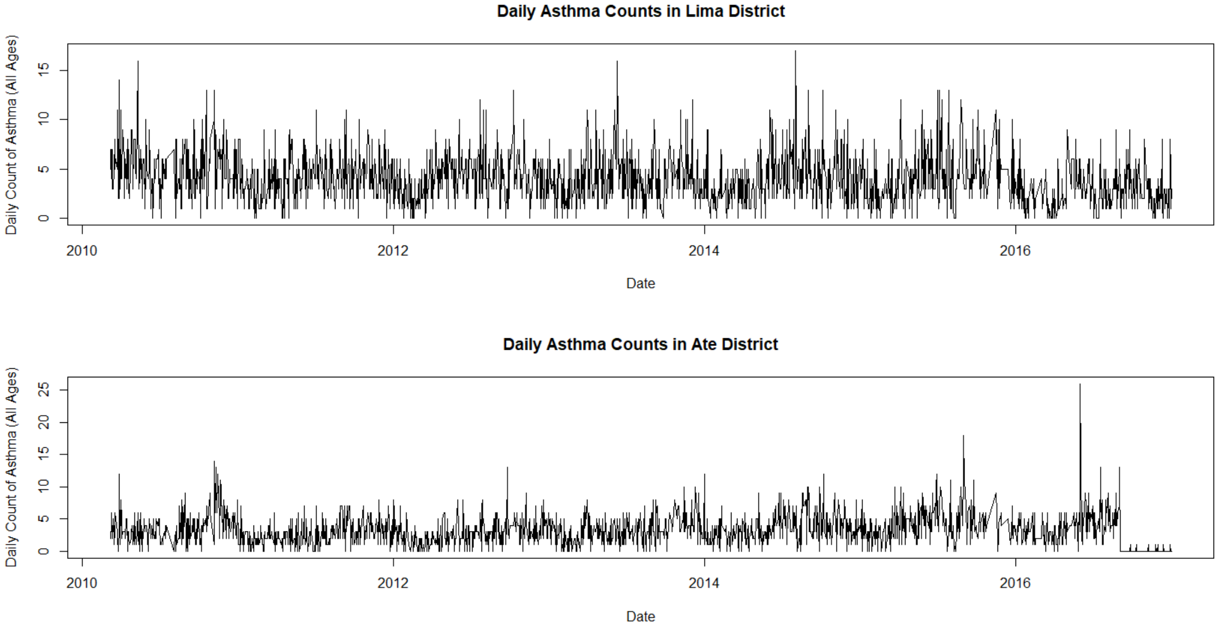


Figure 4. Total number of daily asthma ED visits in Lima District (top) and Ate District (bottom) from March 2010 through December 2016. There does not seem to be a significant difference in the number and trends of asthma ED visits between coastal and mountainous districts.

SUPPLEMENTAL CHAPTER 1B

The Province of Lima was divided into 43 districts plus the Seaport of Callao. The districts of Carabayllo, Chaclacayo, Cienguilla, and Lurigancho were removed since estimates of $PM_{2.5}$ from the exposure model may contain large uncertainties in high elevation and these districts have an average altitude higher than 570 meters. Table S1 summarizes the number of emergency department (ED) visits for each district along with the mean and standard deviation of the estimated population-weighted district-averaged $PM_{2.5}$ during the entire study period.

Table S1. Number of asthma ED visits and population weighted estimated district mean $PM_{2.5}$ and standard deviation in μg/m^3 during the study period for each district including the district name, and region.

NAME	REGION	AREA (km^2)	COUNT	MEAN $PM_{2.5}$ ± SD	MEDIAN(25%-75%)
Barranco	Central	3.33	195	17.0 ± 1.5	16.5 (15.5-17.5)
Breña	Central	3.32	3,493	17.7 ± 2.1	17.0 (15.7-18.5)
Jesús María	Central	4.57	1,871	16.8 ± 2.1	16.0 (14.5-17.6)
La Victoria	Central	8.74	1,356	19.4 ± 2.4	18.5 (17.2-20.2)
Lima	Central	21.88	7,450	18.5 ± 2.1	17.8 (16.6-19.4)
Lince	Central	3.03	1,227	17.6 ± 2.2	16.8 (15.3-18.2)
Magdalena del Mar	Central	3.61	1,941	16.6 ± 1.5	16.1 (15.0-17.1)
Miraflores	Central	9.62	252	17.0 ± 1.6	16.6 (15.5-17.6)
Pueblo Libre	Central	4.38	3,730	17.0 ± 1.6	16.4 (15.4-17.5)
Rímac	Central	11.87	2,244	20.6 ± 2.4	19.8 (18.3-21.6)
San Borja	Central	9.96	284	20.0 ± 2.2	18.9 (17.8-20.7)
San Isidro	Central	11.10	173	17.0 ± 2.0	18.9 (17.8-20.7)
San Miguel	Central	10.72	2,870	17.2 ± 1.2	16.8 (16.0-17.7)

Santiago de Surco	Central	34.75	768	20.6 ± 1.8	19.9 (18.9-21.1)
Surquillo	Central	3.46	353	17.6 ± 1.7	17.0 (15.8-18.2)
Ate	East	77.72	6,066	29.5 ± 4.4	27.7 (25.6-30.7)
El Agustino	East	12.54	524	27.6 ± 3.7	26.6 (24.1-28.9)
La Molina	East	65.75	342	29.5 ± 3.7	28.1 (26.6-30.3)
San Juan de Lurigancho	East	131.25	11,793	32.4 ± 4.6	30.8 (27.8-34.4)
San Luis	East	3.49	269	20.4 ± 2.1	19.6 (18.3-21.4)
Santa Anita	East	10.69	964	29.1 ± 5.0	27.5 (24.4-30.8)
Ancón	North	299.22	601	22.1 ± 2.4	21.7 (20.1-23.5)
Comas	North	48.75	1,357	27.8 ± 3.3	26.6 (24.5-28.8)
Independencia	North	14.56	2,240	23.5 ± 2.5	22.7 (21.2-24.5)
Los Olivos	North	18.25	2,471	19.7 ± 2.3	18.8 (17.4-20.5)
Puente Piedra	North	71.18	8,518	27.3 ± 2.8	26.4 (24.8-28.4)
San Martín de Porres	North	36.91	7,555	18.6 ± 2.2	17.8 (16.6-19.3)
Santa Rosa	North	21.50	121	21.2 ± 2.0	20.4 (19.3-21.9)
Chorrillos	South	38.94	1,097	18.0 ± 1.2	17.6 (16.9-18.6)
Lurín	South	181.12	240	18.9 ± 1.4	18.4 (17.3-19.5)
Pachacamac	South	160.23	343	27.8 ± 1.3	27.4 (26.3-28.5)
Pucusana	South	37.83	29	17.9 ± 1.0	17.5 (16.9-18.5)
Punta Hermosa	South	119.50	19	18.8 ± 1.4	18.9 (18.0-20.7)
Punta Negra	South	130.50	17	19.0 ± 1.3	19.0 (18.0-21.0)
San Bartolo	South	45.01	22	22.5 ± 3.5	20.1 (19.0-22.6)
San Juan de Miraflores	South	23.98	7,742	20.5 ± 2.0	19.9 (18.6-21.3)

Santa María del Mar	South	9.81	3	18.0 ± 1.3	17.9 (17.0-18.7)
Villa El Salvador	South	35.46	3,369	19.7 ± 2.0	19.0 (17.5-20.4)
Villa María del Triunfo	South	70.57	7,576	25.0 ± 2.2	24.2 (22.8-25.8)
Callao	West	147.85	12,489	18.7 ± 1.6	18.2 (17.3-19.4)

CHAPTER 2

Application of geostationary satellite and high-resolution meteorology data in estimating hourly $PM_{2.5}$ levels during the Camp Fire episode in California

Bryan N. Vu, Jianzhao Bi, Wenhao Wang, Amy Huff, Shobha Kondragunta, Yang Liu

* This manuscript is currently under review at Remote Sensing of Environment.

ABSTRACT

Wildland fire smoke contains large amounts of $PM_{2.5}$ that can traverse tens to hundreds of kilometers, resulting in significant deterioration of air quality and excess mortality and morbidity in downwind regions. Estimating $PM_{2.5}$ levels while considering the impact of wildfire smoke has been challenging due to the lack of ground monitoring coverage near the smoke plumes. We aim to estimate total $PM_{2.5}$ concentration during the Camp Fire episode, the deadliest wildland fire in California history. Our random forest (RF) model combines calibrated low-cost sensor data (PurpleAir) with regulatory monitor measurements (Air Quality System, AQS) to bolster ground observations, Geostationary Operational Environmental Satellite-16 (GOES-16)'s high temporal resolution to achieve hourly predictions, and oversampling techniques (Synthetic Minority Oversampling Technique, SMOTE) to reduce model underestimation at high $PM_{2.5}$ levels. In addition, meteorological fields at 3 km resolution from the High-Resolution Rapid Refresh model and land use variables were also included in the model. Our AQS-only model achieved an out of bag (OOB) R^2 (RMSE) of 0.70 (13.93 μg/m^3) and cross-validation (CV) R^2 (RMSE) of 0.71 (17.40 μg/m^3), our combined AQS and PurpleAir weighted model achieved OOB R^2 (RMSE) of 0.87 (9.20 μg/m^3) and CV R^2 (RMSE) of 0.76 (12.89 μg/m^3), and our RF+SMOTE model achieved OOB R^2 (RMSE) of 0.92 (10.37 μg/m^3) and CV R^2 (RMSE) of 0.81 (17.90 μg/m^3). Hourly predictions from our model may aid in epidemiological investigations of intense and acute exposure to wildland fire $PM_{2.5}$.

KEYWORDS

$PM_{2.5}$; GOES16; wildland fire; remote sensing; AOD; SMOTE, Weighted Random Forest

INTRODUCTION

Smoke from wildland fires can traverse tens to hundreds of kilometers away, carrying harmful pollutants that can affect adjacent and downwind communities resulting in excess morbidity and mortality. Although composition of wildland fire smoke is dependent on fuel type, temperature, and wind conditions, smoke from combustion of biomass is generally a mixture of particulate matter, carbon dioxide, water vapor, carbon monoxide, other organic chemicals, and trace minerals [1]. The size of the particulate matter from smoke emitted directly from wildland fires ranges considerably; however, larger particles are likely to deposit in the near field while smaller particles may remain in the atmosphere for days before depositing downwind. Particulate matter, especially $PM_{2.5}$ (particulate mass of particles 2.5 μm or smaller in diameter), composes 90% of total particle mass emitted from wildland fires and has been linked to multiple adverse health outcomes including respiratory and cardiovascular diseases [2-4].

The health effects of $PM_{2.5}$ have been widely documented in studies involving asthma, heart diseases, and premature death, and there is growing evidence that toxicity from particles generated by wildland fires differ from those emitted from other sources [5-8]. For example, Leibel et al. found that the mean daily age-adjusted rate of respiratory emergency departments per 10,000 children in communities located downwind of a wildland fire increased from 55 in the week before the fire to 75 during the week of the fire [9]. Stowell et al. found that a 1 $\mu g/m^3$ increase in wildland fire smoke $PM_{2.5}$ was associated with an odds ratio of 1.08 in asthma emergency department visits yet found null associations with non-smoke $PM_{2.5}$ [10]. Furthermore, studies have also indicated that not only prolonged but also acute exposure to $PM_{2.5}$ results in long lasting effects including persistent coughs, wheezing, and exacerbation of previous conditions such as asthma [2].

Human activities including energy production, industrial activities, and land-use change have led to increased greenhouse gas emissions and consequently climate change. The Center for Research on the Epidemiology of Disasters reported 315 natural disasters globally in 2018 relating to climate change, with

10 of those cases being wildland fires [11]. Climate change has also rendered California's once temperate climate drought stricken while the dried forests act as the perfect fuel once a fire ignites. The Camp Fire, which originated in Butte County in Northern California on November 8 and lasted for 17 days, for example, is considered the deadliest wildland fire in California history. It resulted in 85 casualties, 153,336 acres burnt, 18,804 structures destroyed, and completely destroyed two towns [12]. Economic losses from the Camp Fire stands at $16.65 billion, including $16.5 billion in insured losses and $150 million in firefighting costs [13]. The Camp Fire not only resulted in massive economic losses, but may also result in increased long-term adverse health effects. Due to the intensity and duration of the fire, prolonged and cumulative exposure to fire smoke may result in progressive decline in lung function and increases the overall lifetime risk of heart disease and cancer. Additionally, California is marked by annual Santa Ana winds (SAWs) in the early fall, which is associated with the state's most damaging wildland fires. Studies have suggested that under normal conditions, SAWs improve visibility inland by sweeping polluted air masses out to sea, resulting in lower $PM_{2.5}$ levels near the coast [14]. However, in the presence of fires upwind, SAWs have been shown to increase $PM_{2.5}$ in areas downwind of the fires [9, 14]. Due to the ability of fine particles in smoke to traverse great distances and the harmful effects previously documented in the literature, further research is needed to model the spatiotemporal trends of wildland fire $PM_{2.5}$.

$PM_{2.5}$ exposure assessments based solely on measurements from ground monitors such as the Environmental Protection Agency (EPA) Air Quality System (AQS) are often hindered by inadequate and uneven spatial coverage of the ground measurements. Recent studies also turned to real-time and near real-time $PM_{2.5}$ sensors to improve the coverage of ground measurements [15]. Studies estimating the $PM_{2.5}$ concentrations during wildland fire episodes have relied on ground monitors, chemical transport models (CTMs), and more recently, remotely sensed data including satellite aerosol optical depth (AOD). AOD is a unitless measure of light extinction within the atmospheric column, and previous studies have shown its efficacy in predicting surface level $PM_{2.5}$ [16-20]. Various data fusion approaches have been proposed to model $PM_{2.5}$ from wildland fire smoke [21-24]. For example, fusing ground measurements with satellite remote sensing data can significantly expand the scope of $PM_{2.5}$ exposure modeling without substantially compromising the accuracy of the fused data, but is limited by the coverage of satellite data. Merging ground observations with CTM simulations tends to improve spatiotemporal coverage, and CTMs can also provide speciation information useful in targeting wildland fire smoke PM from other sources. However, the quality of the merged data in areas with limited numbers of monitors depends heavily on the accuracy of CTM simulations, which often contain large errors during fire events [25].

In addition to estimating daily $PM_{2.5}$ concentrations, the rapidly changing characteristics of fire smoke motivates the assessment of $PM_{2.5}$ concentrations at the hourly level. For example, Marsha and Larkin relied on previous days' satellite AOD and fire radiative power (FRP) to make hourly predictions at 10 km resolution. Their model achieved an R^2 of 0.78 and a normalized RMSE of 4.9% [26]. Sanchez-Balseca and Perez-Foguet used dynamic linear models that employ Gaussian Field principles to model hourly $PM_{2.5}$ concentrations in wildland fire events [27]. However, this approach has several limitations including the need for a sizable amount of ground monitors to both calibrate and validate the model, and the requirements of the presence of monitoring stations that acquire both $PM_{2.5}$ and PM_{10} measurements to produce $PM_{2.5}/PM_{10}$ ratios needed in the modeling process. Li et al. used the Hybrid Single Particle Lagrangian Integrated Trajectory (HYSPLIT) model, coupled with emissions inventory and meteorological parameters to forecast hourly $PM_{2.5}$ concentrations during the Camp Fire episode [28]. The researchers configured the model with various combinations of biomass burning emissions data sets, plume rise schemes, meteorological inputs, mixing layer depth options, and vertical motion options to produce an ensemble that predicted $PM_{2.5}$ at 0.1° resolution [28]. However, the spatial resolution of this study is relatively coarse, and the ensemble model sometimes estimated $PM_{2.5}$ levels 10 times higher than the EPA AQS measurements during the first 6 days of the fire and underestimated $PM_{2.5}$ levels for the rest of the fire period, indicating large uncertainties. Such studies do not sufficiently account for the fine spatial and temporal resolution necessary to assess adverse health effects associated with wildland fire smoke.

With the advent of the Geostationary Operational Environmental Satellite-16 (GOES-16), satellite remotely sensed data including AOD, aerosol detection parameters, and fire spot characterization variables are now available at the sub-hour level to aid in modeling processes of events that will benefit from the aid of fine-scale temporal variables [29]. Data from GOES-16 have also been successfully utilized in estimating daily and hourly ambient $PM_{2.5}$ [30]. Recent studies have also shown the effectiveness of low-cost air quality sensors as promising supplements to regulatory ground monitors, by bolstering the number and coverage of ground observations during model training [31, 32]. For example, PurpleAir is a network of low-cost sensors providing continuous measurements of ambient $PM_{2.5}$ and has been shown to accurately report air pollution measurements after calibration with gold-standard collocated monitors [33]. In this study, we reported a machine learning modeling method in conjunction with the Synthetic Minority Over-Sampling Technique (SMOTE) to fuse satellite remote sensing data, assimilated meteorological parameters, land use variables, and EPA AQS and PurpleAir measurements. The resulting model was used to estimate hourly $PM_{2.5}$ levels at 3x5 km^2 resolution during the Camp Fire period in California.

METHODS

Study Domain

California is the third largest and the most populous state with 39 million residents spanning 423,970 km^2 of the United States' western region bordering the Pacific Ocean. The two most populous urban centers, the Greater Los Angeles Area in the south and the San Francisco Bay Area in the north, are the second and fifth largest metropolitan areas in the U.S., respectively. We created a modeling grid at 3x5 km^2 spatial resolution for spatial alignment of all model parameters, and our study region includes 40,578 grid cells. There are 108 AQS air monitoring stations providing hourly measurements and 2,090 available outdoor low-cost $PM_{2.5}$ sensors from the PurpleAir network. Figure 1 shows the study domain and location of ground monitors from the AQS and PurpleAir sensors.

Ground $PM_{2.5}$ Data

Hourly ground $PM_{2.5}$ measurements were taken from the AQS network and the PurpleAir sensors from October 1 through November 30, 2018. AQS is a database of ground monitoring measurements maintained by the EPA (https://www.epa.gov/aqs) with 157 possible stations providing daily measurements in California, 108 of which provided 41,947 hourly measurements during the study period. PurpleAir is a citizen-based, real-time low-cost PM sensor network started in 2015 (https://www.purpleair.com/) with over 8,000 sensors worldwide, measuring $PM_{1.0}$, $PM_{2.5}$ and PM_{10}. PurpleAir measurements in this study were calibrated through a method previously published, and subsequently 848 sensors in California contributed 207,103 hourly measurements during the study period [32]. Because AQS measurements were considered gold-standard in this study compared to PurpleAir, PurpleAir measurements were deleted in grid cells that contain both AQS and PurpleAir and only AQS measurements were kept. Furthermore, grid cells with more than one AQS or PurpleAir measurements were averaged to maintain one measurement per grid cell. On average, AQS monitors only provided six hourly measurements per station per day; therefore, the addition of PurpleAir measurements supplemented more ground measurements to ensure better model fitting results.

Satellite Data

GOES-16 is a geostationary weather satellite operating in the east position at 75.2°W and provides high spatial and temporal resolution imagery through 16 spectral bands at visible and infrared wavelengths using Advanced Baseline Imager (ABI) [34]. Launched in November of 2016, GOES-16 was fully operational in December 2017 providing many different products at 2 km resolution near GOES-16's final longitude (75.2º W), and up to 5 km in Western US. We collected AOD, aerosol detection binary variables, and fire spot detection variables during October and November of 2018. In our study period, fire spot detection and AOD products are available every 5 minutes and aerosol detection product is available every 15 minutes for the continental U.S. Aerosol detection product includes binary aerosol, dust, and smoke mask values. Fire (hot spot characterization) product provides four values: fire mask, a quantitative flag characterizing the quality of a particular pixel; temperature in kelvin; area in square kilometers; and radiative power in megawatts. All products were aggregated to the hourly level.

Meteorological variables

The High-Resolution Rapid Refresh (HRRR, https://rapidrefresh.noaa.gov/hrrr/) is a real-time atmospheric model run by the NOAA National Centers for Environmental Prediction that assimilates radar data every 15 min over a 1-hour period to add further detail to the data provided by the hourly output from the Rapid Refresh model. HRRR has been shown to accurately simulate the observations of near-surface air and dew-point temperature [35]. Furthermore, HRRR has been used to evaluate near-surface wind, temperature, and humidity conditions during wildland fire episodes, and when used in conjunction with GOES estimates, it proved beneficial in improving model performance in calculating geophysical processes such as actual evapotranspiration [36]. In this study, we obtained hourly meteorological parameters including 2-meter temperature, surface pressure, u- and v- wind, planetary boundary layer height, and relative humidity at 3-km spatial resolution from HRRR. HRRR data were joined to the GOES-16 grid through a nearest neighbor match while ensuring that no two HRRR data points were joined to the same GOES-16 grid cell.

Ancillary variables

Land-use parameters including percentage cultivated, barren, shrub, etc. were obtained from the 2011 National Land Cover Database at 30-meter resolution (https://www.mrlc.gov/), elevation information was obtained from the Advanced Spaceborne Thermal Emission and Reflection radiometer Global Digital Elevation (https://asterweb.jpl.nasa.gov/), and distances to nearest primary and secondary roads were computed from the U.S. Census TIGER/Line Shapefiles (https://www.census/). A convolutional layer was calculated for each ground $PM_{2.5}$ measurement by taking an inverse-distance weighted average of the nearest five measurements from the same day and hour to enhance spatial and temporal correlation between ground measurements.

Modeling Approach

A random forest (RF) model is a supervised machine learning ensemble method that aggregates sets of decision trees, or predictions, calculated from the best subset of predictors [37]. The RF model works by selecting a bootstrap sample from all observations with replacement, and subsequently selects the best set of predictors that provides the best split at each node. Advantages of the RF model include its accuracy in learning and classifying features, ability to include large numbers of predictors, and ability to provide variable importance measures that explain the relative contribution of each predictor. Furthermore, individual weights may be assigned to each observation in instances when certain observations are favored over others (e.g., higher accuracy). The RF model has two major hyper-parameters to tune, number of decision trees to grow (n_{tree}), and the number of predictors randomly tried at each split (m_{try}).

We trained the RF model through three different approaches. In the first approach, a RF model is trained with AQS-only measurements. The second approach incorporated PurpleAir measurements to bolster the number of ground observations and enhance measurements of high $PM_{2.5}$ concentrations near the Camp Fire site. In this RF model, full weight is given to AQS measurements and 15% weight is given to PurpleAir measurements. The lower weight of the PurpleAir measurements reflects the higher measurement errors of PurpleAir sensors as well as the lack of consideration in spatial representativeness of this citizen-based network. A more detailed discussion was provided elsewhere [32]. Because high $PM_{2.5}$ concentrations account for only a small fraction of the AQS and PurpleAir data, the third approach applies a Synthetic Minority Over-sampling Technique (SMOTE) to the model training dataset to enhance model performance at high $PM_{2.5}$ levels. SMOTE is a statistical technique that generates synthetic samples using information about the minority class available in the training data [38]. In this study, we set the minority class as any ground measurement at or above 100 $\mu g/m^3$ (2% of the total number of ground observations), nearly three times the daily U.S. national ambient air quality standard (NAAQS) of 35 $\mu g/m^3$. For each measurement in the minority class, the SMOTE function synthetically produces an observation along with its predictors from the five nearest neighbors [38]. Due to the small number of minority observations in our model training dataset, the application of SMOTE enhances the distribution of ground measurements and does not skew the distribution in any way. All three model approaches included the same predictors shown in Table S1 and the same m_{try} (set as default, the square root of the number of predictors rounded up, 6) and n_{tree} (500). Finally, a 10-fold cross-validation (CV) technique is implemented in all three approaches to evaluate model performance. The 10-fold CV works by randomly dividing the total number of observations into 10 segments. Measurements from nine segments are used to train the model and the remaining segment is used to test predictions. This process is repeated 10 times to achieve predictions for all measurements [39, 40]. All data analyses were conducted in R Studio version 3.6.2 and mapping was conducted in ArcGIS version 10.7.1.

RESULTS

Panels A, B, and C in Figure 2 show the scatter plots of model predicted vs. measured $PM_{2.5}$ concentration for the 10-fold cross-validation from the three models, A) AQS-only Model, B) Combined AQS and Weighted PurpleAir Model, and C) SMOTE applied to Combined AQS and Weighted PurpleAir Model. Panels D, E, and F in Figure 2 show the same CV scatter plots of the three models; however, restricted to measurements below 50 μg/m^3 since roughly 94% of all observations are under this level in all models.

AQS-only Model

In total, there were 40,399 grid-averaged hourly AQS observations spanning the modeling period, between October 1 and November 30, 2018 with $PM_{2.5}$ ranging from 0.1 to 657 μg/m^3. The model out of bag (OOB) R^2 is 0.84 (RMSE = 12.00 μg/m^3), and the 10-fold CV R^2 (RMSE) is 0.85 (12.16 μg/m^3). Variable importance ranking from this RF model indicates that aside from the convolutional layer, elevation, pressure, and percent herbaceous land cover were the top three predictors. GOES-16 AOD is the 12th most important predictor while detection of aerosol, detection of smoke, and smoke mask ranked 21st, 22nd, and 27th, respectively. HRRR variables including pressure and planetary boundary layer height

(PBLH) rank 1st and 8th, respectively. Other HRRR variables such as friction velocity, wind components, and radiation flux vary after rank 10.

AQS + Weighted PurpleAir Model

In total, there were 246,181 grid-averaged hourly combined AQS and PurpleAir observations during the study period with ground level $PM_{2.5}$ measurements ranging from 0 to 707 μg/m^3. The model OOB R^2 is 0.86 (RMSE = 9.52 μg/m^3) showing some improvement when PurpleAir measurements are included both in terms of model fit and residual error, likely since PurpleAir measurements captured more higher values. The 10-fold CV resulted in an R^2 (RMSE) of 0.86 (9.60 μg/m^3). Variable importance ranking from this model is similar to the AQS-only Model, with the convolutional layer, pressure, nearest distance to roads and elevation being the top three predictors. GOES-16 AOD ranked 14th highest in importance while detection of aerosol, smoke, and smoke mask ranked 22nd, 26th, and 29th, respectively. Similar to the AQS-only Model, pressure and PBLH rank 1st and 8th, respectively.

AQS + Weighted PurpleAir + SMOTE Model

Of the 246,181 hourly combined AQS and PurpleAir $PM_{2.5}$ observations, 4,819 are at or above the 100 μg/m^3 minority cutoff. The SMOTE application produced an additional two synthetic observation for each minority observation, resulting in 255,819 total grid-averaged hourly combined AQS and PurpleAir observations. The OOB R^2 is 0.92 (RMSE = 10.44 μg/m^3), and 10-fold CV R^2 (RMSE) was 0.91 (9.23 μg/m^3), indicating a substantial increase in model performance compared to the first two models due to the application of SMOTE. However, since AQS measurements were given full weight, the RMSE increased slightly. Variable importance shows that aside from the convolutional layer, nearest distance to roads, pressure, and 10-meter u-wind component were the most important predictors in this model. GOES-16 AOD ranked 9th highest important while smoke mask, detection of smoke and detection of aerosol ranked 14th, 25th, and 26th, respectively. In all three models, the binary dust detection, area, temperature, and radiative power of fire spot variables consistently ranked lowest in the models and were consequently excluded from all models. In this model, HRRR pressure ranked 2nd, 10-meter U-wind component and PBHL ranked 4rd and 5th, respectively, while 10-meter V-wind component ranked 10th. Table S2 of the supplemental shows the variable importance ranking for all three models.

As a sensitivity analysis, predictions from all three models were made for the same hour on the day the Camp Fire reached its peak, November 16th, to assess model performance and prediction capabilities. Figure 3 shows the estimated $PM_{2.5}$ at noon on November 16th. Although the shape of the smoke plumb does not change, $PM_{2.5}$ estimates increased with the addition of PurpleAir and SMOTE.

Hourly predictions were made for the extent of California in grid cells where and when all predictors are present. Figure 4 shows an example of hourly $PM_{2.5}$ predictions by the weighted RF and SMOTE model on November 16, the day ground measurements recorded the highest levels of $PM_{2.5}$ from 6am to 4pm PST. Furthermore, comparison of hourly prediction maps with the true color composite images from MODIS, suggests predictions from the weighted RF and SMOTE model largely aligns with the true-color images of the smoke plumes. Figure 5 shows this comparison with images at noontime on November 8 and November 16. Minor differences in the true-color images and our prediction maps could be caused by our model capturing the $PM_{2.5}$ levels on the surface rather than the column-integrated smoke plumes captured in the satellite image.

DISCUSSION

In this study, we developed a model to predict surface wildland fire $PM_{2.5}$ concentrations during the Camp Fire using three approaches. To the best of our knowledge, this is the first study that integrated low-cost sensor data to bolster ground observations and GOES-16 satellite remote sensing data to achieve high temporal resolution concurrently. To date, many studies documenting the effects of wildland fires on human health have focused on exposures ranging from days to months, often limited by the lack of fine-temporal exposure estimates [41-43]. Especially in California, where wildland fires seem to ravage both the northern and southern regions repeatedly each year with growing intensity as climate change progresses, evidence-based guidelines regarding vulnerable populations are needed to mitigate risks and parse out the adverse health effects of wildland fire smoke from those caused by other environmental hazards [44].

Results from the AQS-only Model show that using EPA's AQS ground measurements alone to model wildland fire $PM_{2.5}$ may underestimate $PM_{2.5}$ levels. Model performance is adversely affected by lack of extensive hourly measurements as well as the lack of AQS monitors near the Camp Fire site to pick up high concentrations. Although AQS monitors are relatively evenly distributed across California, there are only a few located near the site of the Camp Fire. Integrating PurpleAir measurements increased ground observations by over 500%, bolstered the number of measurements at and around the Camp Fire, and improved both model fitting OOB R^2 and RMSE. Even though the model fitting OOB R^2 only improved by 2%, the intercept reduced by 23% from 1.62 to 1.25. Furthermore, the average $PM_{2.5}$ prediction based on the AQS-only Model on November 16th at noon was 30.6 $\mu g/m^3$while the average $PM_{2.5}$ prediction from the AQS + PurpleAir Model on the same day and time was 34.3 $\mu g/m^3$ suggesting that the addition of PurpleAir measurements reduced the amount of underestimation from the AQS-only model. However, these improvements are only minimal since there are still uncertainties in the low-cost sensor measurements due to the light-scattering principle associated with laser particle counters and manufacturing calibration and maintenance. For example, uncertainty may be present in the monitors recording incorrect particle counts and in the conversion between particles counts and mass concentrations. Furthermore, sensors may degrade over time. As a result, data quality may differ based on sensor location and condition. Nonetheless, calibration by Bi et al. indicates that the density of the PurpleAir network partially offsets the impact of measurement errors. As a result, a weight of 15% is given to the PurpleAir measurements, allowing AQS measurements to still have a major role in model training. Predictions in Figure 3 show that the addition of PurpleAir intensified the $PM_{2.5}$ estimates in the north where the fire originated and also in the west due to Santa Ana winds blowing the smoke from east to west. Although the addition of PurpleAir measurements augmented the number of high values, the model still underestimates at high $PM_{2.5}$ levels. Nonetheless, the addition of PurpleAir measurements enables us to calibrate the model for more accurate predictions throughout the entire study domain, especially where the fire originated and downwind from it.

Due to the nature of modeling a wildland fire event in a large domain, the distribution of the ground measurements is right-skewed since the majority of monitors and subsequently their measurements are outside the vicinity of the fire smoke. To compensate, we applied the SMOTE technique to artificially inflate the high values and improve model performance. The duplication of measurements at or above 100 $\mu g/m^3$ ensures that these measurements are more wildland fire-related. Additionally, SMOTE synthetically duplicates the high observations by inverse weighting the nearest five neighbors which also ensure that these duplicates have similar predictors without being exactly identical. The number of high measurements after duplication remains well below 6% of the total number of observations, which we deem as not significantly altering the original distribution. There were 3.7 times more PurpleAir measurements at or above the minority cut off compared to AQS and the implementation of SMOTE improves model performance even though PurpleAir measurements were still given only 15% weight compared to AQS. We utilized a 10-fold cross-validation method instead of a spatial cluster method reported in a few recent $PM_{2.5}$ exposure modeling studies [39, 40]. Monitors neighboring the wildland fire will inevitably measure higher $PM_{2.5}$ levels compared to monitors across the study domain, and these monitors will also measure high values consistently through the extent of the wildland fire period. Consequently, the traditional 10-fold CV was chosen over a spatial cluster CV to reduce the inability of monitors outside of the fire smoke to predict measurements in and around the wildland fire. Although CV results suggest that all three models underestimate the high levels of the wildland fire $PM_{2.5}$, predictions from each model indicate the need to incorporate both PurpleAir measurements and the SMOTE technique. Our full model captures both the spatial extent and the intensity of the smoke plume produced by the wildfire that the AQS-only Model inadequately achieved.

In the end, the model that utilized both AQS and PurpleAir measurements with the incorporation of a weighted sampling scheme and SMOTE is our best performing model. By utilizing low-cost sensor measurements from PurpleAir, we are able to expand both the spatial and temporal coverage of the ground measurements to improve model calibration and performance in predicting wildland fire $PM_{2.5}$. Using a machine learning model with a weighted sampling scheme ensures that the gold-standard bearer AQS are still prominently featured in the models and the implementation of SMOTE allowed us to slightly inflate high levels to reduce model underestimation. To date, few studies exist that predict wildland fire $PM_{2.5}$ due to lack of ground measurements and good predictor variables. A recent model by Li et al. utilized deep learning techniques to predict weekly $PM_{2.5}$ during a 10-year timespan between 2008-2017 in California using MAIAC AOD, variables from MERRA-2, and meteorological and land cover parameters [45]. Although their model did not focus specifically on wildland fire $PM_{2.5}$, it achieved a similar training R^2 of 0.94 and validation R^2 of 0.82 [45], indicating that our model is able to perform sufficiently compared to those with similar predictors in the same region. Furthermore, our model is capable of estimating hourly $PM_{2.5}$ levels during the Camp Fire episode. Therefore, results from our model will enable researchers to investigate the spatial extent at which wildland fire smoke $PM_{2.5}$ traverses as well as the acute temporal fluctuations in concentrations and link this exposure to adverse health outcomes targeted at communities downwind of the event.

Comparison of our hourly predictions to true color composite images from MODIS Aqua show very similar smoke plumes in spatial extent, indicating good model performance. Although predictions are limited to sunlight hours due to the availability of AOD and other GOES-16 predictors used in this study, our model is able to pick up the heterogeneity in $PM_{2.5}$ distribution even inside the smoke plume. However, there are a few limitations. First, GOES-16 is the east position geostationary satellite with a skewed view of the Pacific West. As a result, many of the quality flags associated some of the variables such as AOD suggests low quality. A possible reason for the low quality AOD is the geometrics of the geostationary satellite [46]. Unfortunately, GOES-17, the west position geostationary satellite was not suitable for scientific analyses until January of 2019 after the Camp Fire event. Furthermore, there is also uncertainty if GOES-16 is able to accurately pick up AOD directly inside the smoke plume due to the heavy aerosol loading. Therefore, we are unsure if the missingness is directly due to failure to retrieve an AOD value or if the true AOD value is higher than GOES16's capable range. Future research may consider using GOES16's visible band reflectance as input instead of AOD. Second, although the integration of SMOTE within the model improves the OOB R^2; the implementation of such an approach is arbitrary. Other methods to deal with imbalanced data include under-sampling; however, removing instances in the majority class when two or more observations are similar may result in loss of information.

CONCLUSIONS

The present study is the first to incorporate high temporal resolution geostationary satellite data with low-cost sensors to model wildland fire $PM_{2.5}$ during a major wildland fire, the Camp Fire episode in California. We found that only using ground observations from EPA's AQS network alone was not sufficient for modeling hourly $PM_{2.5}$, and that the addition of PurpleAir low-cost sensors not only bolstered our number of observations but also improved the R^2 and RMSE. Furthermore, the implementation of SMOTE to synthetically enhance high values in our model training dataset further enhanced the model's ability to estimate high $PM_{2.5}$ values. Predictions from our model may be used for epidemiological studies investigating both long-term cumulative exposure to wildland fire $PM_{2.5}$ but also acute intense short-term exposure as well.

ACKNOWLEDGEMENTS

This work was supported by the NASA Applied Sciences Program [Grant # 80NSSC19K0191] and by the National Institutes of Environmental Health Sciences [Grant # 5T32ES12870]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCE

1. Sokolik, I.N., et al., *Progress and Challenges in Quantifying Wildfire Smoke Emissions, Their Properties, Transport, and Atmospheric Impacts.* Journal of Geophysical Research-Atmospheres, 2019: p. 21.
2. Stone, S.L., et al., *Wildfire Smoke: A Guide for Public Health Officials*. 2019.
3. Reid, C.E., et al., *Associations between respiratory health and ozone and fine particulate matter during a wildfire event.* Environment International, 2019. **129**: p. 291-298.
4. Stowell, J.D., et al., *Associations of wildfire smoke PM2.5 exposure with cardiorespiratory events in Colorado 2011–2014.* Environment International, 2019. **133**: p. 105151.
5. Davila Cordova, J.E., et al., *Association of PM(2.5) concentration with health center outpatient visits for respiratory diseases of children under 5 years old in Lima, Peru.* Environmental health : a global access science source, 2020. **19**(1): p. 7-7.
6. Tapia, V., et al., *Time-series analysis of ambient PM2.5 and cardiorespiratory emergency room visits in Lima, Peru during 2010–2016.* Journal of Exposure Science & Environmental Epidemiology, 2019.
7. Tapia, V., et al., *PM2.5 exposure on daily cardio-respiratory mortality in Lima, Peru, from 2010 to 2016.* Environmental Health, 2020. **19**(1): p. 63.
8. Gan, R.W., et al., *The association between wildfire smoke exposure and asthma-specific medical care utilization in Oregon during the 2013 wildfire season.* Journal of Exposure Science and Environmental Epidemiology, 2020. **30**(4): p. 618-628.
9. Leibel, S., et al., *Increase in Pediatric Respiratory Visits Associated with Santa Ana Wind-Driven Wildfire Smoke and PM2.5 Levels in San Diego County.* Annals of the American Thoracic Society, 2020. **17**(3): p. 313-320.
10. Stowell, J.D., et al., *Associations of wildfire smoke PM2.5 exposure with cardiorespiratory events in Colorado 2011-2014.* Environment International, 2019. **133**: p. 11.
11. Fawzy, S., et al., *Strategies for mitigation of climate change: a review.* Environmental Chemistry Letters, 2020: p. 26.
12. U.S. Census Bureau. *Camp Fire - 2018 California Wildfires*. 2018 [cited 2020 05/01/2020]; Available from: https://www.census.gov/topics/preparedness/events/wildfires/camp.html.
13. Reyes-Velarde, A. *California's Camp fire was the costliest global distaster last year, insurance report shows*. 2019 [cited 2020 05/30/2020]; Available from: https://www.latimes.com/local/lanow/la-me-ln-camp-fire-insured-losses-20190111-story.html.
14. Aguilera, R., et al., *Santa Ana Winds of Southern California Impact PM2.5 With and Without Smoke From Wildfires.* Geohealth, 2020. **4**(1): p. 9.
15. Mehadi, A., et al., *Laboratory and field evaluation of real-time and near real-time PM2.5 smoke monitors.* Journal of the Air & Waste Management Association, 2020. **70**(2): p. 158-179.
16. Vu, B.N., et al., *Developing an Advanced PM(2.5) Exposure Model in Lima, Peru.* Remote sensing, 2019. **11**(6): p. 641.
17. Bi, J., et al., *Impacts of snow and cloud covers on satellite-derived PM2.5 levels*. Vol. 221. 2018. 665-674.
18. Hu, X., et al., *Estimating ground-level PM2.5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model.* Remote Sensing of Environment, 2014. **140**: p. 220-232.
19. Liang, F., et al., *MAIAC-based long-term spatiotemporal trends of PM2.5 in Beijing, China.* Science of The Total Environment, 2018. **616-617**: p. 1589-1598.
20. Meng, X., et al., *Estimating PM2.5 speciation concentrations using prototype 4.4 km-resolution MISR aerosol properties over Southern California.* Atmospheric Environment, 2018. **181**: p. 70-81.

21. Zou, Y.F., et al., *Machine Learning-Based Integration of High-Resolution Wildfire Smoke Simulations and Observations for Regional Health Impact Assessment.* International Journal of Environmental Research and Public Health, 2019. **16**(12): p. 20.
22. Mirzaei, M., et al., *Estimation of local daily PM2.5 concentration during wildfire episodes: integrating MODIS AOD with multivariate linear mixed effect (LME) models.* Air Quality Atmosphere and Health, 2020. **13**(2): p. 173-185.
23. Baker, K.R., et al., *Meteorological and air quality modeling for Hawaii, Puerto Rico, and Virgin Islands.* Atmospheric Environment, 2020. **234**: p. 14.
24. Guan, S.H., et al., *Impact of wildfire on particulate matter in the southeastern United States in November 2016.* Science of the Total Environment, 2020. **724**: p. 11.
25. Diao, M.H., et al., *Methods, availability, and applications of PM2.5 exposure estimates derived from ground measurements, satellite, and atmospheric models.* Journal of the Air & Waste Management Association, 2019. **69**(12): p. 1391-1414.
26. Marsha, A. and N.K. Larkin, *A statistical model for predicting PM2.5 for the western United States.* Journal of the Air & Waste Management Association, 2019. **69**(10): p. 1215-1229.
27. Sanchez-Balseca, J. and A. Perez-Foguet, *Modelling hourly spatio-temporal PM2.5 concentration in wildfire scenarios using dynamic linear models.* Atmospheric Research, 2020. **242**: p. 9.
28. Li, Y., et al., *Ensemble PM2.5 Forecasting During the 2018 Camp Fire Event Using the HYSPLIT Transport and Dispersion Model.* Journal of Geophysical Research-Atmospheres, 2020. **125**(15): p. 19.
29. Schmit, T.J., et al., *A CLOSER LOOK AT THE ABI ON THE GOES-R SERIES.* Bulletin of the American Meteorological Society, 2017. **98**(4): p. 681-698.
30. Zhang, H. and S. Kondragunta, *Daily and Hourly Surface PM2.5 Estimation from Satellite AOD.* Earth and Space Science, 2021. **n/a**(n/a): p. e2020EA001599.
31. Lin, C.Q., et al., *Observation of PM2.5 using a combination of satellite remote sensing and low-cost sensor network in Siberian urban areas with limited reference monitoring.* Atmospheric Environment, 2020. **227**: p. 11.
32. Bi, J.Z., et al., *Incorporating Low-Cost Sensor Measurements into High-Resolution PM2.5 Modeling at a Large Spatial Scale.* Environmental Science & Technology, 2020. **54**(4): p. 2152-2162.
33. Johnson, K., et al., *PurpleAir PM2.5 performance across the U.S. #2*. 2020, Resarch Triangle Park, NC: Meeting between ORD, OAR/AirNow, and USFS.
34. Nachamkin, J.E., J. Schmidt, and C. Mitrescu, *Verification of Cloud Forecasts over the Eastern Pacific Using Passive Satellite Retrievals.* Monthly Weather Review, 2009. **137**(10): p. 3485-3500.
35. Lee, T.R., et al., *Evaluation of the High-Resolution Rapid Refresh (HRRR) Model Using Near-Surface Meteorological and Flux Observations from Northern Alabama.* Weather and Forecasting, 2019. **34**(3): p. 635-663.
36. Ha, W.S., G.R. Diak, and W.F. Krajewski, *Estimating Near Real-Time Hourly Evapotranspiration Using Numerical Weather Prediction Model Output and GOES Remote Sensing Data in Iowa.* Remote Sensing, 2020. **12**(14): p. 25.
37. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.
38. Blagus, R. and L. Lusa, *SMOTE for high-dimensional class-imbalanced data.* BMC Bioinformatics, 2013. **14**(1): p. 106.
39. Young, M.T., et al., *Satellite-Based NO2 and Model Validation in a National Prediction Model Based on Universal Kriging and Land-Use Regression.* Environmental science & technology, 2016. **50**(7): p. 3686-3694.
40. Murray, N.L., et al., *A Bayesian ensemble approach to combine PM2.5 estimates from statistical models using satellite imagery and numerical model simulation.* Environmental Research, 2019. **178**: p. 8.

41. Woo, S.H.L., et al., *Air pollution from wildfires and human health vulnerability in Alaskan communities under climate change.* Environmental Research Letters, 2020. **15**(9): p. 13.

42. Agyapong, V.I.O., et al., *Long-Term Mental Health Effects of a Devastating Wildfire Are Amplified by Socio-Demographic and Clinical Antecedents in Elementary and High School Staff.* Frontiers in Psychiatry, 2020. **11**: p. 11.

43. Ontawong, A., et al., *Impact of long-term exposure wildfire smog on respiratory health outcomes.* Expert Review of Respiratory Medicine, 2020. **14**(5): p. 527-531.

44. Maxmen, A., *CALIFORNIA SCIENTISTS RACE TO ASSESS HEALTH RISKS OF WILDFIRE SMOKE.* Nature, 2019. **575**(7781): p. 15-16.

45. Li, L., et al., *Ensemble-based deep learning for estimating PM2.5 over California with multisource big data including wildfire smoke.* Environment International, 2020. **145**: p. 106143.

46. Zhang, H., et al., *Improving GOES Advanced Baseline Imager (ABI) aerosol optical depth (AOD) retrievals using an empirical bias correction algorithm.* Atmos. Meas. Tech., 2020. **13**(11): p. 5955-5975.

CHAPTER 2 TABLES AND FIGURES

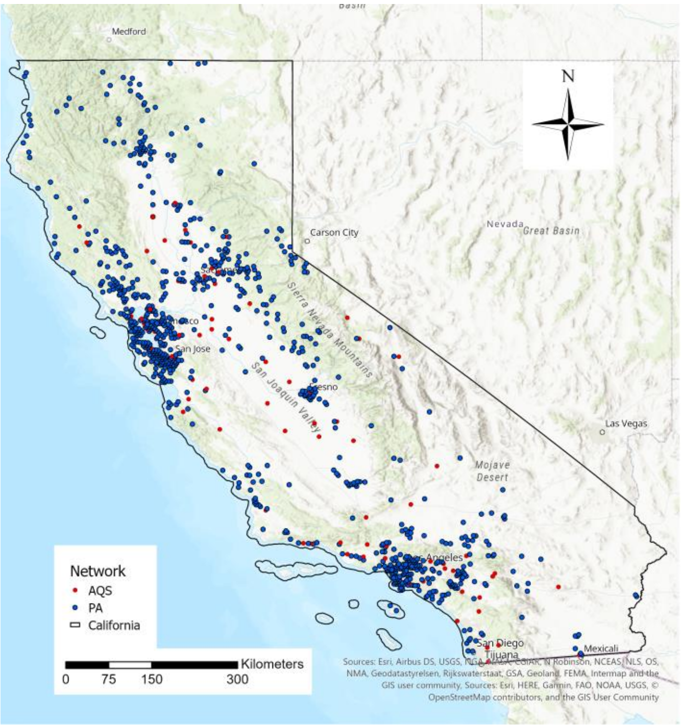

Basin
Medford
N
Carson City
Nevada
Great Basin
Sierra Nevada Mountains
San Francisco
San Jose
San Joaquin Valley
Fresno
Las Vegas
Mojave Desert
Los Angeles
San Diego
Tijuana
Mexicali
Network
AQS
PA
California
Kilometers
0
75
150
300
Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community, Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community

Figure 1. Study domain of California. EPA AQS monitors are pictured in red, PurpleAir sensors are in blue.

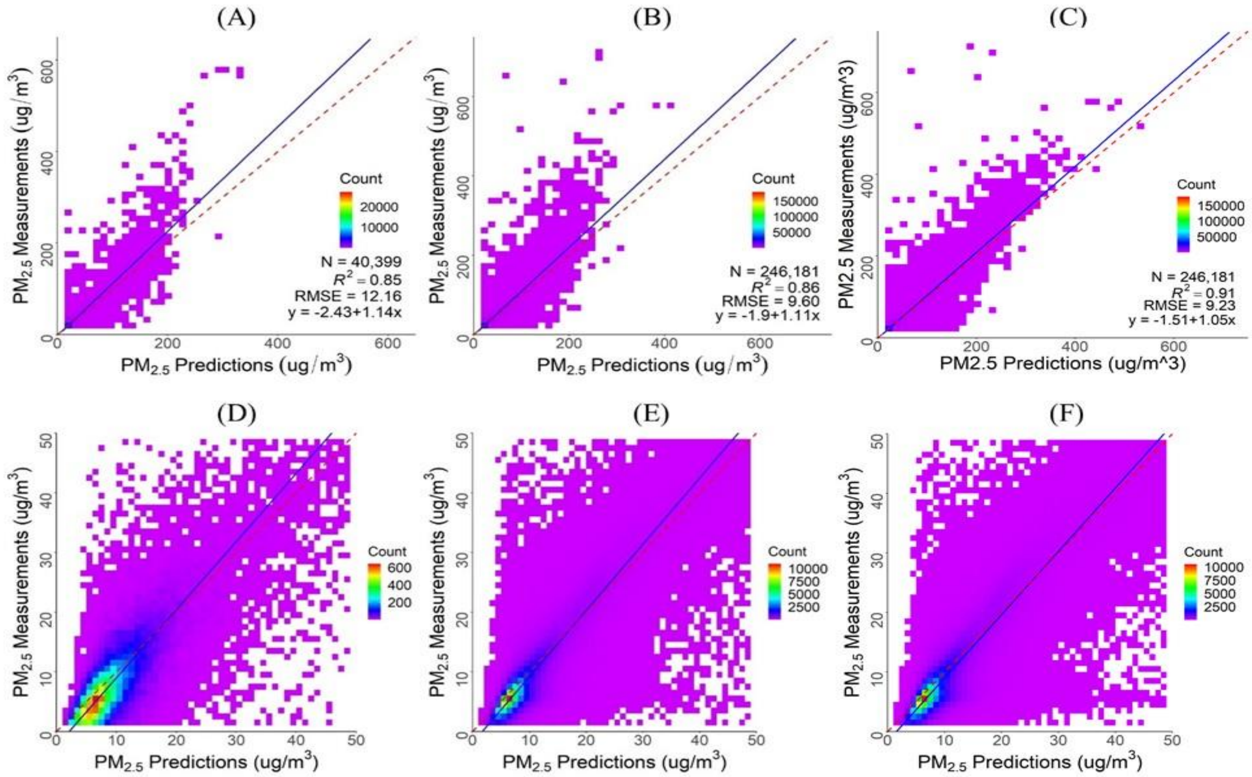


Figure 2. Panel of density scatter plots of full model predicted vs. measured $PM_{2.5}$ concentration from the three models, (A) AQS-only Model, (B) AQS + Weighted PurpleAir Model, and (C) AQS + Weighted PurpleAir + SMOTE Model, and 10-fold cross-validation predicted vs. measured $PM_{2.5}$ concentration from the three models at lower $PM_{2.5}$ levels (<50 μg/m^3), (D) AQS-only Model, (E) AQS + Weighted PurpleAir Model, and (F) AQS + Weighted PurpleAir + SMOTE Model.

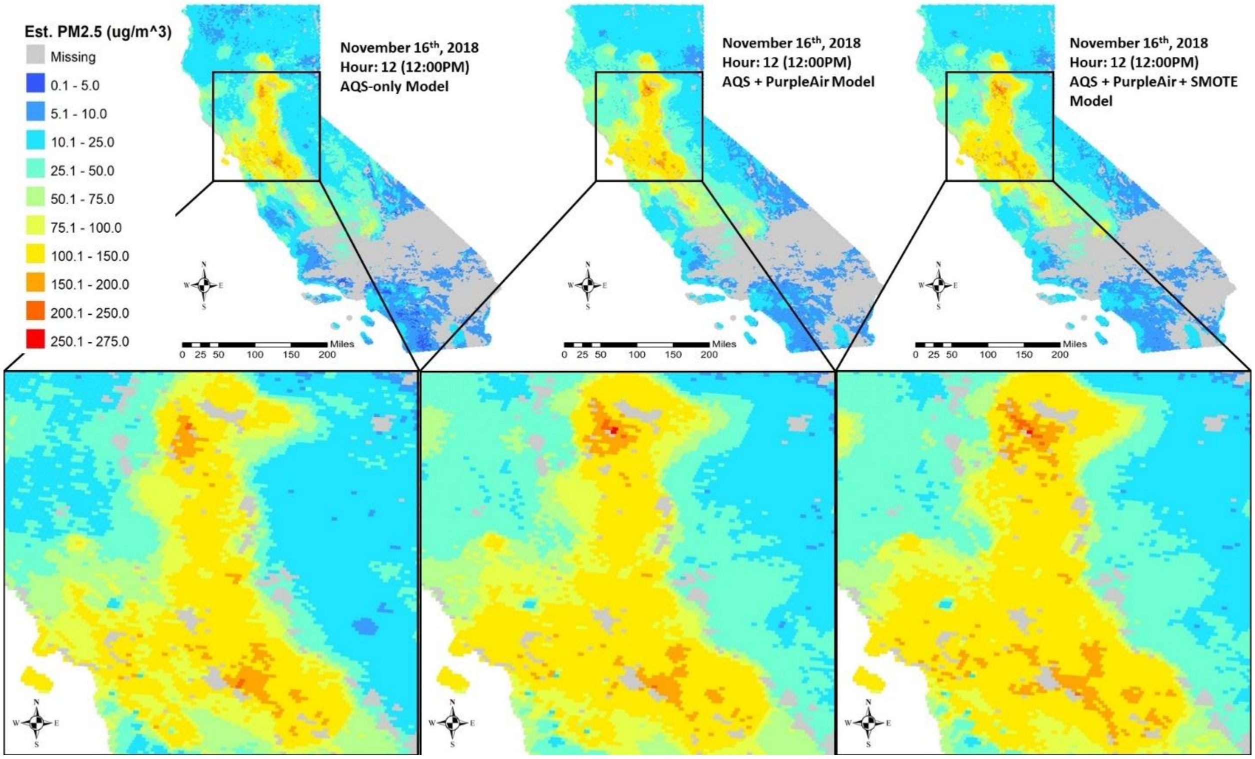


Figure 3. Predictions from all three models (AQS-only, AQS + PurpleAir, AQS + PurpleAir + SMOTE) on 12:00pm on November 16th, 2018. Area of the smoke plume remains the same in all models; however, $PM_{2.5}$ levels increase as PurpleAir and SMOTE is added.

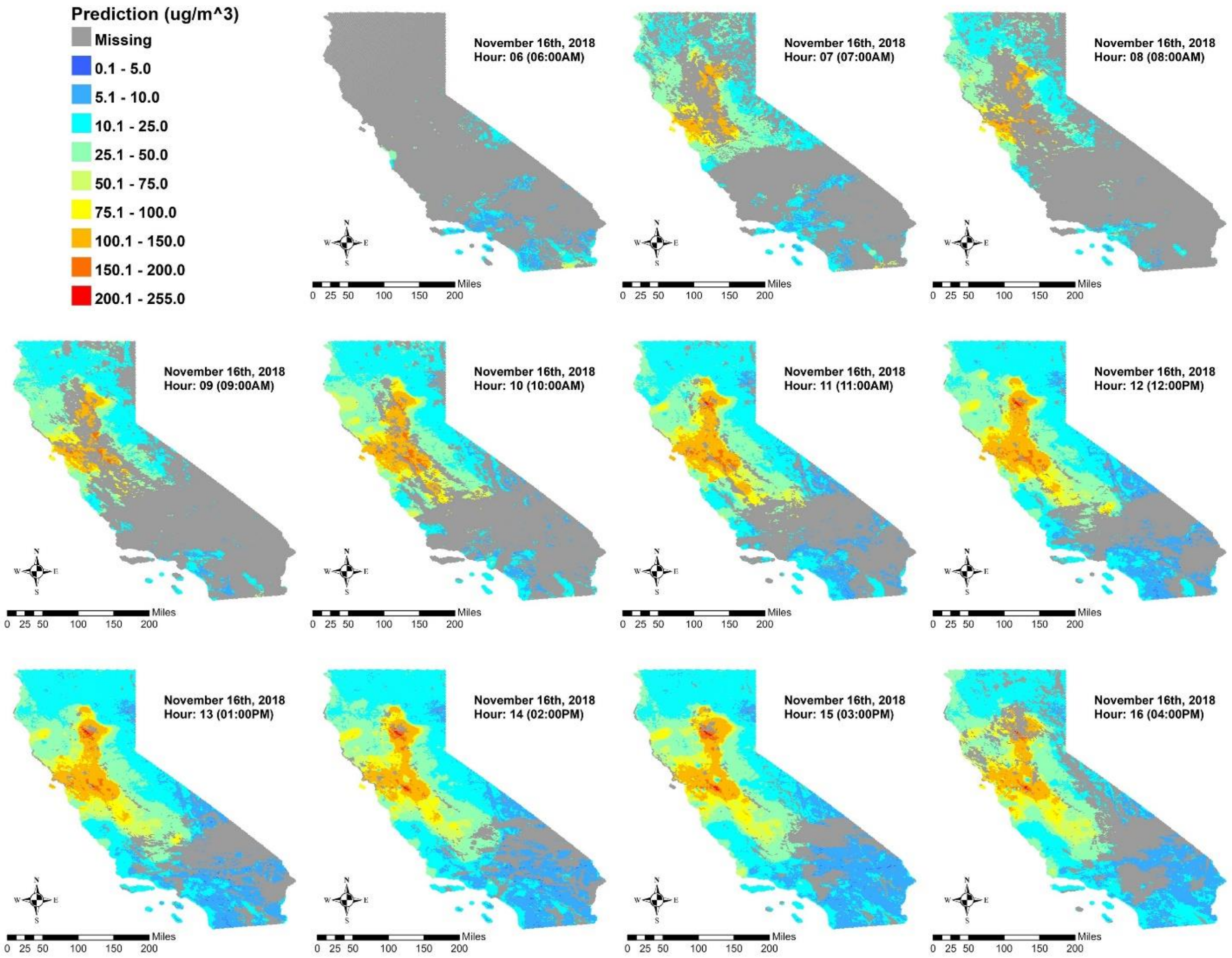


Figure 4. Hourly prediction maps of $PM_{2.5}$ in µg/m^3 from the weight RF and SMOTE model in California on November 16, 2018. Recorded ground measurements were highest on this day.

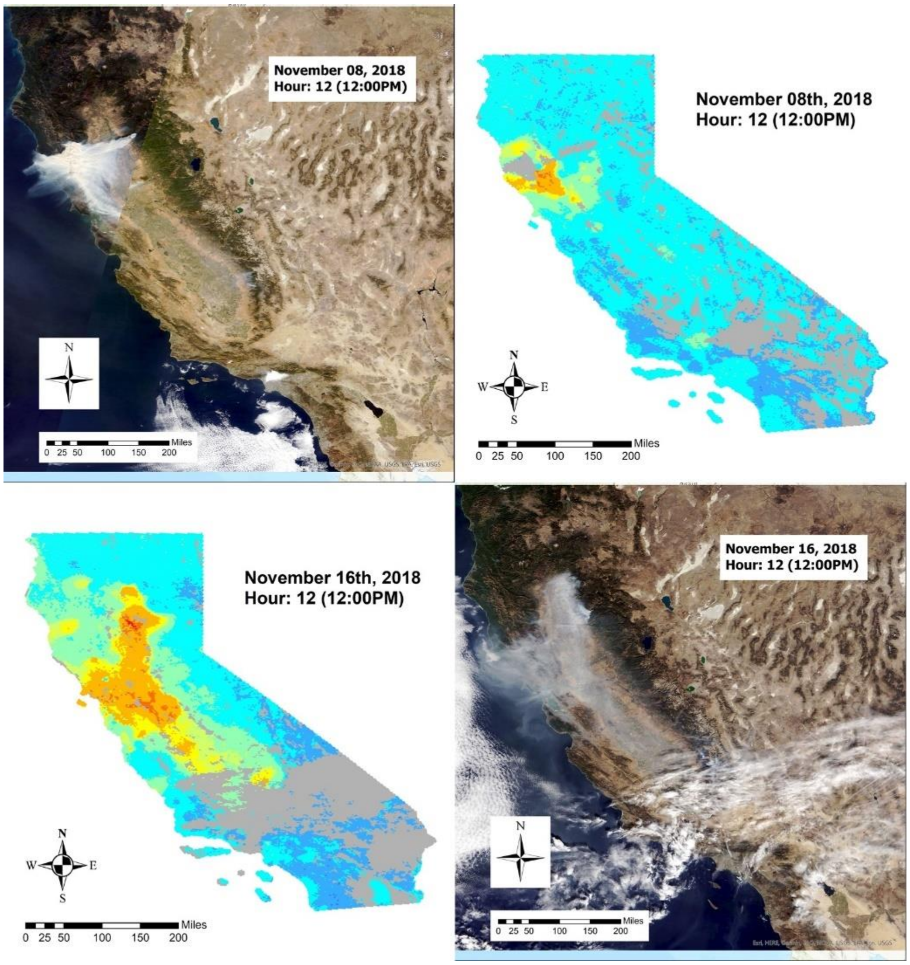


Figure 5. Comparison of hourly prediction maps of $PM_{2.5}$ in μg/m^3 with the true color composite images from MODIS at noontime on November 8 and November 16, the day the Camp Fire started and the day with the highest recorded ground measurements, respectively.

SUPPLEMENTAL CHAPTER 2

All three model approaches utilize the same variables listed below in Table S1. As well as the same parameter specifications in the random forest model, m_{try} set to default (square root of the total number of parameters rounded up, 6) and n_{tree} of 500.

Table S1. Predictor variables used in random forest models of all three approaches.

GOES-16	**HRRR**	**HRRR**
AOD	Planetary boundary layer height (PBLH)	Total cloud cover (TCC)
Aerosol detection	Pressure	Low cloud cover (LCC)
Smoke detection	Upward longwave radiation flux (ULWRF)	Specific humidity
Smoke mask flag	Downward longwave radiation flux (DLWRF)	Relative humidity
Land-use	Sensible heat net flux (SHNF)	Friction velocity
Elevation	Upward shortwave radiation flux (USWRF)	Dew point
Nearest distance to roads	Downward shortwave radiation flux (DSWRF)	2-meter temperature
Population	10-meter U-wind component	Wind speed
% of shrub lands	10-meter V-wind component	Wind direction
% of herbaceous areas	Visibility	Wind gust
% of developed areas	**Land-use**	**Ancillary**
% of cultivated areas	% of forest	Convolutional layer
% of barren lands	% of water bodies	

One of the major advantages of the random forest (RF) model is the output of variable importance from the model, which indicates how much each variable in the model affects the root mean square error (RMSE). Table S2 lists the variable importance output for each of the three models.

Table S2. Variable importance output from the three Random Forest Models.

Rank	**AQS-only Model**	**AQS+PurpleAir and Weighted Model**	**AQS+PurpleAir and Weighted + SMOTE Model**
1	Pressure	Pressure	Road Distance
2	Elevation	Road Distance	Pressure

3	% of herbaceous	Elevation	% herbaceous
4	Population	% herbaceous	U-wind
5	Road Distance	% shrub	PBLH
6	% of shrub	% developed	Elevation
7	% of developed	% forest	% shrub
8	PBHL	PBLH	SHNF
9	% barren	Population	AOD
10	% water	U-wind	V-wind
11	V-wind	% water	% barren
12	AOD	% barren	% water
13	% cultivated	% cultivated	ULWRF
14	Temperature	AOD	Det. Mask
15	% forest	DLWRF	Rel. Humidity
16	DSWRF	Specific Humidity	% cultivated
17	ULWRF	Temperature	Friction Velocity
18	SWRF	ULWRF	DSWRF
19	Specific Humidity	V-wind	TCC
20	Dew Point	TCC	DLWRF
21	Det. AOD	Wind Direction	Specific Humidity
22	Det. Smoke	Det. AOD	Visibility
23	Friction Velocity	USWRF	Dew Point
24	SHNF	Visibility	Temperature
25	DLWRF	SHNF	Det. Smoke
26	Rel. Humidity	Det. Smoke	Det. AOD
27	Det. Mask	Dew Point	LCC
28	Wind Direction	Friction Velocity	USWRF
29	U-wind	Det. Mask	Wind Direction
30	Visibility	DSWRF	% forest
31	LCC	LCC	Population

32	Wind Speed	Wind Speed	% developed
33	Gust	Rel. Humidity	Wind Speed
34	TCC	Gust	Gust

CHAPTER 3

Association between $PM_{2.5}$ and emergency department outpatient visits for cardiovascular outcomes in California

Bryan N. Vu, Rohan D'Souza, Danlu Zhang, Ana Rappold, Matthew Strickland, Kyle Steenland, Yang Liu, Howard Chang

ABSTRACT

Background: Hundreds of millions of people are affected by cardiovascular disease (CVD). Air pollution, particularly $PM_{2.5}$ (particulate matter that are 2.5 micrometer of less in aerodynamic diameter), has been shown to adversely affected respiratory disease. However, the relationship between $PM_{2.5}$ and CVD are still uncertain in many parts of the world. California, one of the largest and most populated state in United States, has been prone to wildland fires in recent decades due to climate change with limited amount of studies that investigate the association between $PM_{2.5}$, both from smoke and non-smoke sources, and CVD.

Objective: We conducted a case cross-over study to determine the association between several CVD outcomes and $PM_{2.5}$ in California between 2016 to 2018.

Methods: We used a conditional logistic regression model to regress daily cases of CVD outcomes with total $PM_{2.5}$, smoke $PM_{2.5}$, and non-smoke $PM_{2.5}$. CVD outcomes include acute myocardial infarction (AMI), arrythmia, heart failure (HF), ischemic heart disease (IHD), stroke, and total CVD. Total $PM_{2.5}$ was obtained from a random forest exposure model and smoke $PM_{2.5}$ was obtained from an interpolated dataset from the environmental Protection Agency Air Quality System with a Hazard Mapping System indicator for smoke. Non-smoke $PM_{2.5}$ was calculated as total $PM_{2.5}$ minus smoke $PM_{2.5}$. We included temperature and dewpoint to account for meteorology and indicator variables for active fire spots and prescribe burns to control for wildland fires. We also assigned control using 2 methods: 1) assigning the same day of week within the month, 2) assigning 2 bi-weekly same day of week before the case date and 2 bi-weekly same day of week after the case date.

Results: In general, we found positive but nonsignificant associations between the 3 $PM_{2.5}$ exposure types and the CVD outcomes when controls were assigned as the same day of week within the month. When controls are assigned on a bi-weekly basis, we found positive and significant association between total $PM_{2.5}$ and non-smoke $PM_{2.5}$ and the CVD outcomes but not for smoke $PM_{2.5}$.

Discussion: Results from this study align well with previous studies regarding the relationship between CVD and $PM_{2.5}$ and provides additional literature on the association between smoke $PM_{2.5}$ and CVD.

KEYWORDS

$PM_{2.5}$; smoke $PM_{2.5}$; ED visits, case-crossover, CVD

INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of disease burden in the world with over 500 million cases worldwide [1]. When stratified by specific outcomes, ischemic heart diseases (IHDs) including acute myocardial infarction (AMI) and heart failure (HF) pertained to 197 million of the 500 million cases, and there were 101 million stroke survivors [1]. The burden of CVDs results in high health care costs and may impact not only developed countries but also low- and middle-income countries (LMIC). There are many risk factors that drive CVD, including social, metabolic, behavioral, and environmental influences. Previous studies have demonstrated the effects of risk factors such as smoking and body mass index (BMI) on CVD, often with significant adverse associations [2-5].

There is also an increasing number of studies investigating the effects of air pollution, particularly $PM_{2.5}$ (particulate matter with an aerodynamic diameter of 2.5 microns or less), on CVD since $PM_{2.5}$ has been shown to be related to adverse cardiopulmonary function [6]. For example, one study has shown that reduced long-term exposure to $PM_{2.5}$ also reduced the rate of cardiovascular mortality across 619 counties in the United States [7]. However, the effects of smoke $PM_{2.5}$ on CVD have not been well documented in current literature. The composition of smoke $PM_{2.5}$ generated from wildland fires differ from $PM_{2.5}$ generated from other sources including vehicular combustion and industrial sources such as energy generation, and there have been increasing evidence suggesting that particulate matter generated from wildland fires may be more toxic [8]. Thus, more investigation is needed in understanding the association between not only anthropogenic sources of $PM_{2.5}$ but also smoke $PM_{2.5}$ and cardiovascular diseases.

The state of California in the United States has been prone to wildland fire activities in recent decades. California's drought-stricken climate and dried forests act as fuel during each fire season, which are burning more intensely and lasting longer each year [9]. Due to climate change, wildland fire conditions in California are not expected to improve in the near future. Additionally, heart disease is one of the leading causes of death in California, according to the American Heart Association [10]. Although mortality from

CVD has declined in recent decades, the number of deaths in California resulting from CVD is more than the two leading causes, cancer and respiratory diseases, combined in 2014 [11]. Furthermore, 1 in 3 Californian are living with at least one form of CVD, resulting in an estimated $37 billion in annual health care costs [11]. As one of the major mortality and morbidity burden in California, more studies are needed to understand the effects of air pollution on CVD.

Although there have been several studies investigating the association between $PM_{2.5}$, particularly wildland fire $PM_{2.5}$, and respiratory outcomes in California [12, 13], few studies have investigated the association between $PM_{2.5}$ and various CVD morbidity outcomes at once for the entire state of California in recent years, especially in 2017 and 2018 when the state recorded record numbers of wildland fires [14-17]. Nonetheless, these recent studies either focused on mortality and not morbidity [18], investigated total CVD without parson out specific forms [14], or did not investigate smoke $PM_{2.5}$ in conjunction with CVD. Understanding the effects of $PM_{2.5}$ on CVD morbidity will drive prevention and intervention strategies to mitigate and reduce the burden in California. Furthermore, parsing out the different forms of CVD will enable future researchers and physicians to pinpoint and target specific outcomes for intervention and prevention. Moreover, with the increase in climate change and wildland fire activities, more studies are needed to assess the effects of smoke $PM_{2.5}$ on CVD outcomes in a region prone to wildland fires.

This present study aims to investigate the association between 5 common forms of CVD, acute myocardial infarction (AMI), arrhythmia, heart failure (HF), ischemic heart disease (IHD), stroke, and total CVD, with total $PM_{2.5}$, non-smoke $PM_{2.5}$, and smoke $PM_{2.5}$ in California between 2016 to 2018. We propose to use a case cross-over approach to reduce the effects of unmeasured confounders, and include prescribe burn and active fire spot indicators to control for fire location. Results from this study will help to bolster the current state of limited literature on the topic of air pollution and cardiovascular diseases in California, especially during years with strong wildland fires.

METHODS

Study Domain

California is the most populous and third-largest U.S. state with over 39.5 million residents across over 423 thousand squared-kilometers. There are 1,719 zip codes in California; however, analyses were restricted to 1,625 zip codes with at least 1 case per day. In this study, we will focus on the time period between 2016 and 2018. Figure 1 shows the study domain of California divided into zip codes.

Health Data

Health information was obtained from the California Office of Statewide Health Planning and Development (OSHPD), which provides nonpublic datasets on hospital emergency department (ED) visits and licensed freestanding ambulatory surgery clinics. For this study, only the outpatient encounter, also known as service visit, is used and data for each encounter includes the diagnosis and the patient's zip code. CVD outcomes in this study include arrhythmia, AMI, HF, IHD, stroke and total CVD.

Satellite-derived $PM_{2.5}$ estimates

$PM_{2.5}$ estimates were obtained from a Random Forest (RF) model that calibrated remote sensing data, and meteorological and land use variables to ground monitoring measurements from the Environmental Protection Agency's (EPA) Air Quality System (AQS) and low-cost particular matter sensor network called PurpleAir. In short, aerosol optical depth (AOD) retrieved from NASA satellites along with variables including elevation, population, vegetation index, and meteorological variables such as temperature, humidity, and wind speed were calibrated to daily ground $PM_{2.5}$ measurements from the AQS regulatory monitoring network and the PurpleAir low-cost sensors through a decision tree-based model [19]. The RF model achieved a random cross-validation (CV) R^2 of 0.86 with relatively low prediction error and produced daily $PM_{2.5}$ predictions at 3 km^2 spatial resolution [19].

Smoke $PM_{2.5}$

Smoke $PM_{2.5}$ was calculated from the Spatially Interpolated $PM_{2.5}$ Concentrations for the US from 2006-2018, version 2 dataset downloaded from the EPA's AQS (https://aqs.epa.gov/aqsweb/documents/data_mart_welcome.html). This is a 15 km^2 gridded dataset kriged from raw $PM_{2.5}$ in-site measurements from the AQS. Each pixel also contains a smoke plume flag from the National Oceanic and Atmospheric Administration (NOAA) Hazard Mapping System (HMS) smoke product, indicating if a pixel on a particular day is part of a smoke plume. Only grids with an HMS indicator are considered smoke $PM_{2.5}$, all other grids without an HMS indicator are set to zero. Each zip code was matched to the nearest 15 km grid, and in cases where more than 1 grid fall within a zip code, an average of the grids was taken.

Fire Indicators and Meteorological Variables

To control for fires, active fire spot information at 1 km resolution was downloaded from the Moderate Resolution Imagine Spectroradiometer (MODIS) Collection 6 (https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms/active-fire-data). To consider the impact of the extent of each active fire, a 50 km buffer was added to each fire spot and matched to the zip codes. Similarly, information on prescribed burns was obtained from the Prescribed Fire Information Reporting System (PFIRS) (https://ssl.arb.ca.gov/pfirs/) with information on the latitude and longitude of each prescribe burn. Since prescribed burns occur in a maintained setting, a 10 km buffer was added to each burn and matched to the zip codes. Meteorological variables including daily maximum temperature and dewpoint temperature is obtained from the Daily Surface Weather and Climatological Summaries (DAYMET V4) at 1 km resolution. Each 1 km DAYMET grid was matched to the zip codes and zip codes with more than 1 DAYMET grid is averaged.

Case Cross-Over Approach

A case cross-over approach was used to investigate the association between $PM_{2.5}$ and the CVD outcomes. First, we assessed the association between total $PM_{2.5}$ and the 6 CVD outcomes. We ran separate models for each outcome with the RF $PM_{2.5}$ estimates as the main exposure. Second, we investigated the effects of smoke $PM_{2.5}$ by parsing out smoke and non-smoke $PM_{2.5}$. Non-smoke $PM_{2.5}$ is calculated as total $PM_{2.5}$ (RF $PM_{2.5}$) minus smoke $PM_{2.5}$. We also assessed controls in 2 different scenarios. First, we utilized the traditional case cross-over approach of assigning controls by assigning the exposures of the same day of week within the case-month as controls. Second, we assigned the exposures of the same day of week 4 weeks before, 2 weeks before, 2 weeks after, and 4 weeks after the case date. The additional length in time between the controls aim to reduce misclassification of exposure with the assumption that fires may last longer than a week and that controls may have smoke $PM_{2.5}$ assigned as exposure. For all models, we assessed same-day exposure, lag 1, lag 2, lag 3, and a moving average of same-day, lag 1, and lag 2. To control for fluctuation in meteorology, splines (quadratic and cubic) were added for daily maximum temperature and dewpoint temperature. As sensitivity analyses, we restricted the models to only include zip codes with a smoke $PM_{2.5}$ value to ensure that exposure to smoke $PM_{2.5}$ is captured in the models. We also categorized smoke $PM_{2.5}$ into quartiles to determine if a dose response exists in the smoke $PM_{2.5}$ and CVD relationships.

RESULTS

In total there were 884,290 CVD cases in California between 2016 to 2018. Table 1 shows the total number of cases for each outcome and the total number of observations used in the model (cases + controls) for both of the two types of controls that were assessed.

For every 10 μg/m^3 of total $PM_{2.5}$ same-day exposure, the rate of AMI outpatient ED visits increased by 1.7% (95%CI: -0.7%,4.0%), the rate of arrhythmia outpatient ED visits increased by 0.3% (95%CI: -0.6%,1.2%), the rate of HF outpatient ED visits increased by 1.7% (95%CI: -0.2%,3.6%), the rate of IHD outpatient ED visits increased by 1.4% (95%CI: -0.2%,3.1%), the rate of stroke outpatient ED visits decreased by 0.6% (95%CI: -2.4,1.1%), and total CVD outpatient ED visits increased by 0.3% (95%CI: -0.2%, 0.7%) with weekly controls in the case-month. Conversely, for models with biweekly controls set at

4-weeks before, 2-weeks before, 2-weeks after, and 4-weeks after case date, for every 10 µg/m^3 of total PM$_{2.5}$ same-day exposure, the rate of AMI outpatient ED visits increased by 0.9% (95%CI: -1.2%,3.0%), the rate of arrhythmia outpatient ED visits increased by 1.6% (95%CI: 0.8%,2.5%), the rate of HF outpatient ED visits increased by 2.8% (95%CI: 1.1%,4.5%), the rate of IHD outpatient ED visits increased by 1.4% (95%CI: -0.1%,2.9%), the rate of stroke outpatient ED visits decreased by 0.1% (95%CI: -1.7,1.0%), and total CVD outpatient ED visits increased by 1.4% (95%CI: 1.0%, 1.8%).

For every 10 µg/m^3 of non-smoke PM$_{2.5}$ same-day exposure, the rate of AMI outpatient ED visits increased by 1.8% (95%CI: -1.0%,5.0%), the rate of arrhythmia outpatient ED visits increased by 0.6% (95%CI: -0.5%,1.6%), the rate of HF outpatient ED visits increased by 1.8% (95%CI: -0.2%,4.0%), the rate of IHD outpatient ED visits increased by 1.7% (95%CI: -0.3%,3.6%), the rate of stroke outpatient ED visits decreased by 0.3% (95%CI: -3.3,1.7%), and total CVD outpatient ED visits increased by 0.3% (95%CI: -0.2%, 0.8%) with weekly controls in the case-month. Conversely, for models with biweekly controls set at 4-weeks before, 2-weeks before, 2-weeks after, and 4-weeks after case date, for every 10 µg/m^3 of non-smoke PM$_{2.5}$ same-day exposure, the rate of AMI outpatient ED visits increased by 1.1% (95%CI: -1.4%,4.0%), the rate of arrhythmia outpatient ED visits increased by 2.1% (95%CI: 1.2%,3.0%), the rate of HF outpatient ED visits increased by 3.6% (95%CI: 1.8%,5.5%), the rate of IHD outpatient ED visits increased by 2.2% (95%CI: 0.5%,3.9%), the rate of stroke outpatient ED visits increased by 1.2% (95%CI: -0.5,3.0%), and total CVD outpatient ED visits increased by 1.9% (95%CI: 1.5%, 2.3%).

For every 10 µg/m^3 of smoke PM$_{2.5}$ same-day exposure, the rate of AMI outpatient ED visits increased by 1.2% (95%CI: -2.1%,4.0%), the rate of arrhythmia outpatient ED visits decreased by 0.8% (95%CI: -2.2%,0.7%), the rate of HF outpatient ED visits increased by 1.3% (95%CI: -1.9%,4.0%), the rate of IHD outpatient ED visits increased by 0.4% (95%CI: -2.0%,2.8%), the rate of stroke outpatient ED visits decreased by 0.18% (95%CI: -0.46,1.0%), and total CVD outpatient ED visits decreased by 0.1% (95%CI: -0.8%, 0.6%) with weekly controls in the case-month. Conversely, for models with biweekly controls set at 4-weeks before, 2-weeks before, 2-weeks after, and 4-weeks after case date, for every 10 µg/m^3 of smoke

$PM_{2.5}$ same-day exposure, the rate of AMI outpatient ED visits increased by 0.6% (95%CI: -2.4%,4.0%), the rate of arrhythmia outpatient ED visits increased by 0.4% (95%CI: -0.9%,1.8%), the rate of HF outpatient ED visits increased by 0.6% (95%CI: -2.3%,3.5%), the rate of IHD outpatient ED visits decreased by 0.1% (95%CI: -2.3%,2.1%), the rate of stroke outpatient ED visits decreased by 2.9% (95%CI: -5.5,-0.4%), and total CVD outpatient ED visits increased by 0.3% (95%CI: -0.4%, 0.9%).

Figure 2 shows the plotted odds ratio of each CVD outcome for a 10 µg/m^3 increase in exposure of total $PM_{2.5}$, non-smoke $PM_{2.5}$, and smoke $PM_{2.5}$. Results for the sensitivity analyses including restricting the models to include only zip codes that have a smoke estimate found that effect estimates remain largely unchanged. Furthermore, categorizing smoke $PM_{2.5}$ estimates resulted in model non-convergence since the distribution of smoke $PM_{2.5}$ is highly skewed with only a handful of cases and controls that have exposures higher than the median smoke $PM_{2.5}$ in the respective outcome. Results for fire indicators were not significant in any of the models.

DISCUSSION

This study is the first to evaluate the effects of total $PM_{2.5}$, non-smoke $PM_{2.5}$, and smoke $PM_{2.5}$ on several CVD outcomes in California between 2016 to 2018. The utilization of spatially and temporally resolved $PM_{2.5}$ exposure estimates from a machine learning model help to minimize measurement error and exposure misclassification. Few studies in California have focused on evaluating the effects of PM2.5 on CVD outcomes in recent years as wildland fires are more frequent and burn more intensely. Smoke plumes generated from wildland fires may traverse hundreds of kilometers and affect millions of people downwind.

Results from this study indicate that, in general, there is a positive association between $PM_{2.5}$ and outpatient ED visits of CVD between 2016 to 2018 in California. However, in models where controls were assigned weekly based on the case-month, all associations are not significant. Furthermore, in all three $PM_{2.5}$ exposures, stroke continues to produce a protective effect between 0.1%-2.9% decrease in outpatient ED visits for every 10 µg/m^3 increase in $PM_{2.5}$. This result is different from previous literature. For example, a study conducted in China investigating the association between $PM_{2.5}$ and stroke, ischemic stroke, and hemorrhagic stroke found a 0.37% (95%CI: 0.15%,0.60%) increase, 0.46% (95%CI: 0.21%, 0.72%) increase, and a -0.13% (95%CI: -0.73%,0.48%) decrease in acute incidence for every 10 µg/m^3 increase in $PM_{2.5}$, respectively [20]. The difference in estimates may be a result of a lower case-count in our data, ~52,000 cases compared to the ~132,000 cases in the Chinese study.

This study also found that when controls are assigned bi-weekly, twice before the case date and twice after the case date, many of the estimates for the CVD outcomes are significant compared to the traditional method of assigning controls to the same day of week within the case-month. These significant associations are observed only in the total $PM_{2.5}$ and non-smoke $PM_{2.5}$ exposures likely because the correlation between total $PM_{2.5}$ and non-smoke $PM_{2.5}$ is quite high, with a correlation coefficient around 0.8 for each CVD outcome. Moreover, AMI and Stroke are both nonsignificant in the bi-weekly control models for total $PM_{2.5}$ and non-smoke $PM_{2.5}$ exposure, also likely due to wide confidence interval as a result of a smaller number of observations.

Results from this study shows an increase of ~0.6%-3.6% increase in outpatient ED visits for heart failure for all $PM_{2.5}$ exposure types. This result is consistent with other studies in the literature including one conducted in Chile that found a 1.6% (95%CI: 0.9%-3.0%) increase in heart failure emergency hospitalizations for every 10 µg/m^3 increase in $PM_{2.5}$ [21]. Additionally, a study conducted by Li et al. in China showed a 0.35% (95%Ci: 0.06%-0.64%) increase in the number of hospital admissions for every 10 µg/m^3 increase in same day ambient $PM_{2.5}$ exposure [22]. This study also found an increase of ~0.4%-2.2% increase in outpatient ED visits for ischemic heart disease in total $PM_{2.5}$ and non-smoke $PM_{2.5}$ exposures. However, we found a slight protective effect (0.1% decrease) in the bi-weekly controlled smoke $PM_{2.5}$ exposure model. Again, our results are consistent with previous literature that indicates a 0.27% (95%CI: 0.21%-0.33%) increase in IHD morbidity for every 10 µg/m^3 increase in ambient $PM_{2.5}$ in a study conducted in Beijing between 2010 and 2012 [23]. The association between $PM_{2.5}$ and acute myocardial infarction is ~0.6%-1.8% increase in outpatient ED visits for every 10 µg/m^3 increase in $PM_{2.5}$ exposure. In contrast, a study conducted in Massachusetts, U.S. found a 4% increase in the odds of AMI for every 1.05 μg/m^3 increase in total PM$_{2.5}$ exposure [24]. Moreover, for every 10 μg/m^3 increase in PM2.5 exposure, the association for arrythmia ranged between -0.08%-2.1%. Similarly, a study conducted by Zheng et al. found a 2.09% (95%CI: 1.58%-2.60%) increase in hospital admission for arrythmia for every interquartile range (47.5 μg/m^3) increase in PM$_{2.5}$ exposure [25].

Based on the two methods for assigning controls, assigning bi-weekly controls to each case produced large effect estimates that are significant compared to assigning controls within the same day of week for each case-month. One of the main concerns regarding this assignment of controls is that wildland fires may last longer than a week. If controls are assigned based on the same day of week within a month, the effect of the association may be diluted if the controls are also assigned a high PM$_{2.5}$ exposure. Another concern is that the effect estimates for smoke PM2.5 are often protective and are always nonsignificant. This is likely due to only a small percentage of the total number of observations having a measurement for smoke. For each CVD outcome, only about 0.3% of the total observations have a smoke PM$_{2.5}$ measurement greater than 0. This may partially be due to the course resolution of the smoke PM$_{2.5}$ dataset at 15 km^2, and the crude method of parsing smoke PM$_{2.5}$ from total PM$_{2.5}$. Therefore, estimates for smoke PM$_{2.5}$ may not be robust. Similarly, indicators for active fire spots and prescribe burns were found to be nonsignificant in all models, regardless of PM$_{2.5}$ type and control assignment. Assignment of the 50 km buffer for active fire spots and 10 km buffer for prescribe burns were made arbitrarily with scientific inference for wildland fire and prescribe burn behaviors.

In general, results from this study aligns well with previous studies reported from China and other parts of the U.S. However, results in this study tend to be nonsignificant due to a few limitations. One of the biggest limitations of this study is that only outpatient data is used. Generally, for CVD outcomes, especially for severe forms including stroke and AMI, patients will likely not be released the same day [26]. As such, the current health data utilized in this study may not be representative of the general population living with these CVD outcomes. Due to an error in diagnostic date and codes in the health data, inpatient cases were not included. Future research will entail correctly identifying inpatient records for inclusion. Second, as mentioned earlier, the smoke $PM_{2.5}$ was obtained at 15 km^2 resolution and may lead to high uncertainties and bias in exposure misclassification. Future research should focus on identifying a better data source for smoke $PM_{2.5}$. Finally, future research should also focus on identifying major wildland fire events to bolster the active fire spot indicator and ensure that the smoke $PM_{2.5}$ estimates are accurate both temporally and spatially.

CONCLUSIONS

This is the first study to look at various CVD outcomes including acute myocardial infarction, heart failure, arrythmia, ischemic heart disease, stroke, and total CVD in association with total $PM_{2.5}$, non-smoke $PM_{2.5}$, and smoke $PM_{2.5}$. Results from this study indicate that the effect estimates for each of the CVD outcomes are consistent with previous studies conducted elsewhere including China and other part of the United States. However, most of the estimates are nonsignificant, suggesting that the relationship between $PM_{2.5}$ and CVD remains unclear. Future research should focus on correcting the inpatient records for inclusion in these models and to obtain a better smoke $PM_{2.5}$ dataset with finer spatial resolution to reduce exposure misclassification. Results from this study adds to the current literature on the relationship between cardiovascular diseases and air pollution in the context of with and without wildland fire influence.

ACKNOWLEDGEMENTS

This work was supported by the NASA Applied Sciences Program [Grant # 80NSSC19K0191] and by the National Institutes of Environmental Health Sciences [Grant # 5T32ES12870]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCE

1. Roth, G.A., et al., *Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update From the GBD 2019 Study*. Journal of the American College of Cardiology, 2020. **76**(25): p. 2982-3021.
2. Pope, C.A., et al., *Lung Cancer and Cardiovascular Disease Mortality Associated with Ambient Air Pollution and Cigarette Smoke: Shape of the Exposure&#x2013;Response Relationships.* Environmental Health Perspectives, 2011. **119**(11): p. 1616-1621.
3. Mons, U., et al., *Impact of smoking and smoking cessation on cardiovascular events and mortality among older adults: meta-analysis of individual participant data from prospective cohort studies of the CHANCES consortium.* BMJ : British Medical Journal, 2015. **350**: p. h1551.
4. Khan, S.S., et al., *Association of Body Mass Index With Lifetime Risk of Cardiovascular Disease and Compression of Morbidity*. JAMA Cardiology, 2018. **3**(4): p. 280-287.
5. Bastien, M., et al., *Overview of Epidemiology and Contribution of Obesity to Cardiovascular Disease.* Progress in Cardiovascular Diseases, 2014. **56**(4): p. 369-381.
6. Du, Y., et al., *Air particulate matter and cardiovascular disease: the epidemiological, biomedical and clinical evidence.* Journal of thoracic disease, 2016. **8**(1): p. E8-E19.
7. Corrigan, A.E., et al., *Fine particulate matters: The impact of air quality standards on cardiovascular mortality*. Environmental Research, 2018. **161**: p. 364-369.
8. Reid, C.E. and M.M. Maestas, *Wildfire smoke exposure under climate change: impact on respiratory health of affected communities.* Curr Opin Pulm Med, 2019. **25**(2): p. 179-187.
9. Li, S. and T. Banerjee, *Spatial and temporal pattern of wildfires in California from 2000 to 2019.* Scientific Reports, 2021. **11**(1): p. 8779.
10. Association, A.H., *California State Fact Sheet*. 2017.
11. Conroy, S.M. *Burden of Cardiovascular Disease in California, 2016*. 2016 07/10/2021]; Available from: http://healthpolicy.ucla.edu/publications/search/pages/detail.aspx?PubID=1587.
12. Aguilera, R., et al., *Fine Particles in Wildfire Smoke and Pediatric Respiratory Health in California.* Pediatrics, 2021. **147**(4): p. e2020027128.
13. Aguilera, R., et al., *Wildfire smoke impacts respiratory health more than fine particles from other sources: observational evidence from Southern California.* Nature Communications, 2021. **12**(1): p. 1493.
14. Bi, J., et al., *Temporal changes in short-term associations between cardiorespiratory emergency department visits and PM2.5 in Los Angeles, 2005 to 2016.* Environmental Research, 2020. **190**: p. 109967.
15. Ebisu, K., et al., *Age-specific seasonal associations between acute exposure to PM2.5 sources and cardiorespiratory hospital admissions in California.* Atmospheric Environment, 2019. **218**: p. 117029.
16. Ostro, B., et al., *Associations of Source-Specific Fine Particulate Matter With Emergency Department Visits in California.* American Journal of Epidemiology, 2016. **184**(6): p. 450-459.
17. Jones, C.G., et al., *Out-of-Hospital Cardiac Arrests and Wildfire-Related Particulate Matter During 2015–2017 California Wildfires.* Journal of the American Heart Association, 2020. **9**(8): p. e014125.
18. Liao, N.S., et al., *Particulate Air Pollution and Risk of Cardiovascular Events Among Adults With a History of Stroke or Acute Myocardial Infarction.* Journal of the American Heart Association, 2021. **10**(10): p. e019758.
19. Bi, J.Z., et al., *Incorporating Low-Cost Sensor Measurements into High-Resolution PM2.5 Modeling at a Large Spatial Scale.* Environmental Science & Technology, 2020. **54**(4): p. 2152-2162.

20. Ban, J., et al., *Associations between short-term exposure to PM2.5 and stroke incidence and mortality in China: A case-crossover study and estimation of the burden.* Environmental Pollution, 2021. **268**: p. 115743.

21. Vera, J., et al., *Fine Particulate Air Pollution (PM2.5) Increases Emergency Hospital Admissions due to Decompensate Heart Failure.* Epidemiology, 2009. **20**(6).

22. Li, M., et al., *Association Between PM2.5 and Daily Hospital Admissions for Heart Failure: A Time-Series Analysis in Beijing.* International Journal of Environmental Research and Public Health, 2018. **15**(10): p. 2217.

23. Xie, W., et al., *Relationship between fine particulate air pollution and ischaemic heart disease morbidity and mortality.* Heart, 2015. **101**(4): p. 257.

24. Madrigano, J., et al., *Long-term Exposure to PM2.5 and Incidence of Acute Myocardial Infarction.* Environmental Health Perspectives, 2013. **121**(2): p. 192-196.

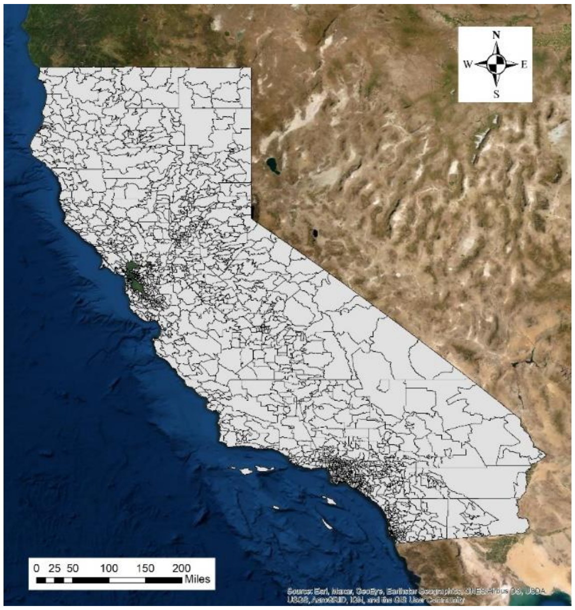
25. Zheng, Q., et al., *The effect of ambient particle matters on hospital admissions for cardiac arrhythmia: a multi-city case-crossover study in China.* Environmental Health, 2018. **17**(1): p. 60.

26. Eichelberger, C., et al., *Emergency Department Visits and Subsequent Hospital Admission Trends for Patients with Chest Pain and a History of Coronary Artery Disease.* Cardiology and therapy, 2020. **9**(1): p. 153-165.

CHAPTER 3 TABLES AND FIGURES

Table 1. Total number of cases and total number of observations (cases + controls) by CVD outcomes with weekly controls compared to bi-weekly controls.

	Weekly Controls					
	AMI	Arrhythmia	HF	IHD	Stroke	Total CVD
# of Cases	28,921	191,791	55,650	62,701	51,929	884,290
Total Obs.	127,104	843,305	244,852	275,438	227,985	3,886,751
	Bi-weekly Controls					
	AMI	Arrhythmia	HF	IHD	Stroke	Total CVD
# of Cases	28,921	191,791	55,650	62,701	51,929	884,290
Total Obs.	137,653	912,641	264,941	298,910	247,440	4,213,915



Source: Esri, Maxar, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

Figure 1. Study domain of California divided into zip codes.

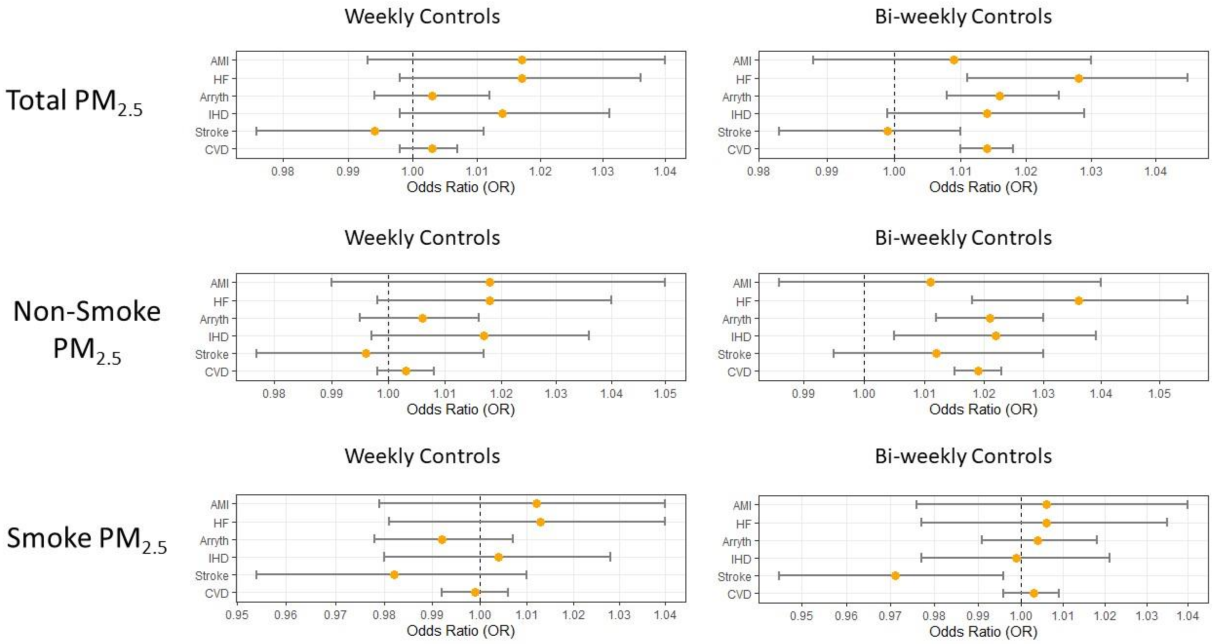


Figure 2. Plots of odds ratios by CVD outcome for every 10 µg/m^3 increase in exposure of total $PM_{2.5}$, non-smoke $PM_{2.5}$, and smoke $PM_{2.5}$. The left column of plots denotes models where controls are assigned the same day of week within the case-month. The right column of plots denotes models where controls were assigned 4-weeks before, 2-weeks before, 2-weeks after, and 4-weeks after the case-date.

CONCLUSIONS

To our knowledge these studies are the first of its kind in their respective topic. Results from these three aims ultimately add more evidence to the body of literature pertaining to the utilization of satellite remote sensing data in modeling air pollution in not only low- and middle-income countries but also during extreme weather events. The publication of the $PM_{2.5}$ exposure model in Lima, Peru enables researchers to investigate the association between air pollution and a variety of adverse health outcomes in Lima. Results from those epidemiological studies will provide evidence for air pollution mitigation in Lima since the permissible levels of $PM_{2.5}$ in Peru is about double the standards set forth by the World Health Organization. Furthermore, results from aim two indicate that even with an adequate network of ground monitors, the addition of low-cost sensor observations to bolster the number of ground measurements tremendously improve model performance. The addition of SMOTE also reduces the root mean error, suggesting that implementation of such techniques is promising in modeling extreme events such as wildland fires. Finally, results from aim three indicate the need to effectively parse out non-smoke $PM_{2.5}$ and smoke $PM_{2.5}$ from total $PM_{2.5}$. Although the effect estimates are very similar to those previously published, the non-significance of the estimates when utilizing traditional assignment of weekly controls compared to the significance of the estimates when utilizing a bi-weekly assignment of controls suggests that further investigation is needed to determine the true relationship between $PM_{2.5}$ and cardiovascular diseases in California.