Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Bryan N. Vu Date

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

By

Bryan N. Vu
Master of Science in Public Health

Environmental Health & Epidemiology

_____

Yang Liu, Ph.D.
Committee Chair

_____

Paige Tolbert, Ph.D.
Committee Member

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

By

Bryan N. Vu

Master of Public Health
University of California, Irvine
2016

Bachelor of Science
California State University, Long Beach
2014

Thesis Committee Chair: Yang Liu, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Environmental Health & Epidemiology
2018

Abstract

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

By Bryan N. Vu

Background: There is convincing evidence of adverse health effects induced by exposure to $PM_{2.5}$ in the growing body of literature. Lima's topography and aging vehicular fleet results in severe air pollution with limited amounts of monitors to effectively quantify measurements for epidemiologic studies.

Objectives: We propose to develop a high-performance satellite-driving exposure model to estimate daily $PM_{2.5}$ concentrations at a 1 km spatial resolution in Lima, Peru from 2010 to 2016 using a combination of ground measurements, aerosol optical depth (AOD), meteorological fields, parameters from atmospheric chemical transport models, and land use variables.

Methods: Parameters from the Weather Research and Forecasting model coupled with Chemistry (WRF-CHEM) and the European Centre for Medium-Range Weather Forecasts (ECMWF) were evaluated against ground monitoring stations from Weather Underground as well as ground $PM_{2.5}$ measurements from the DIGESA and SENAMHI sites in Lima, Peru. A random forest model was used to gap-fill non-random missing satellite AOD data due to cloud cover to enhance spatial coverage and quality. Both a linear mixed effects model and a random forest model was used to fit AOD, WRF-CHEM, ECMWF, and land use parameters against ground measurements from 16 monitoring stations with available data between 2014 to 2016. Both models were then used to predict daily $PM_{2.5}$ concentrations from 2010 to 2016.

Results: The model fitting R^2 for the LME model was 0.63 and random forest model was 0.73. The overall cross-validation (CV) R^2 value and (RMSE) for the linear mixed effects model and random forest model was 0.58 (7.08 $\mu g/m^3$) and 0.73 (5.66 $\mu g/m^3$), respectively. The intercept and slope of the LME model was 0 and 1, compared to -2 and 1 from the random forest model, suggesting that the random forest underestimates $PM_{2.5}$ compared to the LME model. Nonetheless, the random forest model performed better based on no change between model fitting R^2 and CV R^2.

Conclusions: Our prediction model allows for construction of long-term historical daily $PM_{2.5}$ levels to support fundamental and imperative epidemiological studies that will likely impact governmental policies on air pollution in Lima, Peru.

Developing Advanced PM2.5 Exposure Models in Lima, Peru

By

Bryan N. Vu

Master of Public Health
University of California, Irvine
2016

Bachelor of Science
California State University, Long Beach
2014

Thesis Committee Chair: Yang Liu, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Environmental Health & Epidemiology
2018

Acknowledgements

I would like to express my utmost appreciation to Dr. Yang Liu, Dr. Kyle Steenland, Dr, Odón Sánchez, and the members of The Remote Sensing Research Group at Emory for their help and expertise on the analyses of this thesis. Finally, I would also like to express my deepest appreciation to my family, friends, and colleagues for their continuing support in all my endeavors.

Table of Contents

1. Introduction

1.1 $PM_{2.5}$ and Health Impacts

The World Health Organization (WHO) establishes that joint exposure to household and ambient air pollutants, including particulate matter, contributes to 7 million global deaths in 2012 [2]. Ambient air pollutants, including $PM_{2.5}$ (fine particles with aerodynamic diameter of 2.5 μm or less), are emitted from a large variety of sources such as power plants, gasoline and diesel vehicles, wood burning, smelters, as well as natural sources including sea spray aerosols and wind-blown dust particles [5,17]. Epidemiologic studies have linked exposure to $PM_{2.5}$ with increased adverse health outcomes including asthma, cardiovascular diseases, type-2 diabetes, and obesity among adults and children [1, 3]. Results from a study conducted by Mirabelli et al. on the association between outdoor $PM_{2.5}$ in the United States and asthma symptoms in the previous 14 days among adults with active asthma, show a 3.4% increase in symptom prevalence among adults with active asthma for every 1 μg /m^3 increase in $PM_{2.5}$ [1]. Additionally, a meta-analysis of cohort studies conducted by Liu et al. showed significant increases in wheezing, coughing, and lower respiratory illness among children exposed to more than 25 μg/m^3 of $PM_{2.5}$ compared to children exposed to less than 25 μg/m^3 of $PM_{2.5}$. Increasing numbers of studies such as these, indicate that exposure to ambient $PM_{2.5}$ is significantly associated with the development of respiratory diseases among not only adults, but also children from North America, Europe, and Asia [8]. Conversely, there is a substantially limited number of air pollution studies conducted in South America where pollution levels far exceeds those of Europe and North America.

1.2 $PM_{2.5}$ in Lima, Peru

Lima, the capitol of Peru, is the third most populous city in the Americas according to a 2015 census survey, and is the second most polluted city in the Americas according to the WHO [18, 6]. Lima's air pollution problem stems from an aging fleet of public transportation in urban areas and the widespread use of indoor biomass stoves in rural areas [6, 10]. In 1991, the Lima government eliminated fare regulations and barriers to entry, creating an oversupply of aging minibuses. The citizen's group, Lima Como Vamos, reports that the average age of Lima's buses exceeds 20 years, far more than the average age of the bus fleet in Sao Paulo in Brazil at 4.2 years. Due to the densely populated urbanization of Lima, the resulting traffic congestion leads to particulate matter levels that exceed the WHO's standards, currently set at 10 $\mu g/m^3$ annual mean and 25 $\mu g/m^3$ 24-hour mean [20], by more than 200% [11]. 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 [10]. This creates high volumes of air pollution not only in urban areas but also in the mountainous rural areas as well. As a result, the rise in air pollution not only affect those living in Lima, but also the workers living in the rural communities on the outskirts of the city, whose average commute trip is between 90 to 180 minutes [11]. Yet, there is a limited number of studies on the effects of air pollution on health risks in Lima, especially on ambient air pollution outside the home. Accordingly, due to the known health risks associated with $PM_{2.5}$ exposure and the high volume of air pollution in Lima, more studies are needed to assess the effects of $PM_{2.5}$ in order to curtail Lima's air pollution problems and propose new policies to improve air quality standards.

1.3 Limitations of Air Pollution Studies and Ground Measurements

To date, many studies have been conducted on the association between health outcomes and traffic-related air pollution. However, these studies have been cross-sectional in design with exposure measured as the distance from household to highway, or small cohort study designs that relied on a limited number of ground $PM_{2.5}$ measurements to infer correlation [12, 13]. One of the main limitations of the cross-sectional design with distance as a stand-in for exposure is the assumption that participants live at the address on the questionnaire. Additionally, in small cohort study designs, estimating exposures of $PM_{2.5}$ in a given population is traditionally done by assigning measurements of a central ground monitor to people living within a certain distance from it, from a few kilometers to tens of kilometers [4]. This method often leads to misclassification of exposure due to spatial misalignment and results in bias estimates of the health risks [4]. Furthermore, utilizing air-monitoring data presents many other limitations. Air quality monitoring networks are designed and implemented to focus on acquiring measurements of pollutants in highly populated areas. As a result, the monitor networks are usually densely concentrated in one area or region, and often omitted from rural and mountainous regions [21]. Additionally, air monitors usually collect samples once every three to four days due to time constraints and costs in collecting and analyzing the samples [21]. Due to these limitations, the biggest challenge in utilizing air quality monitors in health studies is obtaining an accurate and precise estimate of $PM_{2.5}$ concentrations through space and time. Since epidemiologic studies of $PM_{2.5}$ require long-term historical and accurate exposure data, relying on ground monitor measurements may not be the best suitable option.

1.4 Remote Sensing Techniques

Remote sensing techniques have proved useful in estimating ground $PM_{2.5}$ concentration due to its ability to provide comprehensive spatial and temporal coverage, making it a suitable supplement for $PM_{2.5}$ ground monitors [14]. Satellite sensors provide aerosol optical depth (AOD), a dimensionless measure between zero and one, of the aerosols such as smoke, particles, and dust distributed within a column of air from the Earth's surface to the top of the atmosphere [19]. Lower AOD values around 0.01 corresponds to a clean atmosphere while AOD values above 0.35 corresponds to a hazy environment [19]. 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 [15]. Most commonly used AOD products include those 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 [16]. AOD retrieved using the visible and near-IR bands is sensitive to aerosols with a size range of 0.1 to 2.0 μm, similar to the size of $PM_{2.5}$ [16]. Furthermore, a Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm has been applied on satellite retrieved AOD to achieve stronger correlations with $PM_{2.5}$. The MAIAC algorithm uses 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 [4].

Remote sensing techniques have proved successful in studies conducted by Liu et al. in China and the United States [14, 15]. Liu et al. compared model fit in a two-stage modeling technique with and without AOD [15]. Their results indicate that the AOD

model has higher predicting power compared to the non-AOD model, R^2 (0.79) for AOD compared to R^2 (0.48) for non-AOD model [15]. Furthermore, a study on estimating $PM_{2.5}$ concentrations in Southeastern United States using MAIAC AOD also proved successful [4]. In that study, Hu et al. also used a two-stage spatial statistical modeling approach to fit meteorological fields, land use parameters, and MAIAC AOD to ground observations [4]. Hu et al. achieved a model fitting R^2 of 0.83 with a mean prediction error of 1.89 $\mu g/m^3$, and a cross validation R^2 of 0.67 and mean prediction error of 2.54 $\mu g/m^3$, indication that MAIAC AOD can be used to estimate $PM_{2.5}$ concentrations. Finally, Ma et al. conducted a study to estimate $PM_{2.5}$ concentrations in China while also using satellite AOD as the primary predictor. Their national-scale geographically weighted regression model achieved a cross validation R^2 of 0.64, which also attests to the ability of AOD as an important predictor of $PM_{2.5}$ [14]. Furthermore, all the studies listed above found that the correlation between $PM_{2.5}$ and satellite AOD, derived from advanced statistical models including generalized linear regression and generalized additive modeling, can be greatly improved when land use and meteorological parameters are included [14,15].

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 Dirección General de Salud Ambiental e Inocuidad Alimentaria (DIGESA) stations from the Ministry of Health, and 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. Although the data quality is sparse, this presents an opportunity to implementing satellite remote sensing techniques in estimating ground-level $PM_{2.5}$ in a region with critically

high levels of air pollution couple with a limited number of epidemiological studies on its impact on health risks.

1.5 Study objectives

The goal of this project is to build a $PM_{2.5}$ exposure model to estimate daily $PM_{2.5}$ concentrations at 1km spatial resolution in Lima for year 2010 to 2016. This exposure model is derived from satellite AOD data, simulation data from chemical transport models (CTMs), meteorological fields from a forecast model, and land use parameters.

2. Data and Methods

2.1 Study Area

The study region is Lima City, the Capital of Peru spanning from -11.57º North to -12.52º South and -77.20º West to -76.62º East. The study region was divided into 2970 one kilometer-squared pixels. A 10km buffer around the study region was implemented to ensure accuracy of MAIAC AOD as well as any other parameters that need to be interpolated from coarser resolutions down to the desired 1km squared grid cells. The added buffer will also allow for better estimation of $PM_{2.5}$ concentrations near the outer boundaries of the study area. With the 10km buffer, the total number of pixels increased to 5959 during the model development and training period. Figure 1 shows the map of Lima's borders along with pixels within the study domain and the 10km buffer pixels.

2.2 Datasets and Processing

2.2.1 Ground Data

There are six DIGESA stations and ten SENAMHI stations that recorded $PM_{2.5}$ measurements in Lima, Peru. Data availability for DIGESA stations include monthly mean measurements for $PM_{2.5}$ from 2001 to 2005 and a daily average measurement every 4 days starting from 2007 to 2016. However, there were many missing data from the DIGESA network, often months at a time. In total, the six DIGESA sites contributed 1,120 daily observations from 2010 to 2016. Table 1 shows the number of observations for each monitoring station in the DIGESA network from 2010 to 2016. SENAMHI stations recorded daily mean measurements of $PM_{2.5}$ from 2014 to 2016. Table 2 shows the number of observations for each monitoring station in the SENAMHI network from 2010 to 2016. The ten SENAMHI sites contributed 7,363 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 William Checkley from Johns Hopkins University. These monitors provided weekly estimates from November 2011 to March 2013, and were extrapolated to the daily level by giving the six preceding days the same concentration as the measured value on the seventh day. Due to the dense location of these 15 monitors, spanning six grid cells in the southern region of the study domain, an average of each pixel was calculated if more than one monitor fell within a certain pixel. In total, Checkley sites provided an additional 2,265 daily observations to the model fitting dataset. Table 3 shows the numbering of each Checkley site and the monitors used to compose the average for that site along with the total number of observations each site contributed to the model fitting dataset. Figure 2 shows the location of each monitor station in relation to the study domain. Finally, results of a time series analysis of each station to assess monthly and yearly trends of $PM_{2.5}$ in $\mu g/m^3$ for both monitoring networks can be seen in Figure 3.

2.2.2 Satellite Remote Sensing Data

Satellite aerosol optical depth (AOD) at 1km spatial resolution retrieved from MODIS (Moderate Resolution Imaging Spectroradiometer) aboard NASA's Terra and Aqua satellites operating since 1999 and 2002, respectively, is calculated through a MAIAC (Multi-Angle Implementation of Atmospheric Correction) algorithm [16]. The MAIAC algorithm accomplishes atmospheric correction by first gridding the data to a fixed 1 km grid and accumulating of up to 16 days of measurements [16]. 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 [16]. Furthermore, the MAIAC algorithm has been shown to

be the most accurate algorithm when compared to aerosol optical thickness (AOT) from the Aerosol Robotic Network (AERONET) stations [16]. AERONET is a system of sun-photometers established by NASA in conjunction with other partnerships to measure atmospheric aerosol properties [7]. MAIAC AOD in Lima was compared to AOD measurements from ARICA, the nearest AERONET site to Lima to assess validity and accuracy. Subsequently, MAIAC AOD was gap-filled through a random forest method discussed in Bi et al. [23]. Finally, MAIAC AOD was linked to each grid cell through a one-to-one spatial link as both were in 1km spatial resolution. Figure 4 shows the annual average gap-filled AOD in the study domain from 2010 to 2016. Additionally, Figure 5 shows monthly mean gap-filled AOD for 2010, 2012, 2014, 2015, and 2016.

2.2.3 Chemical Transport Model Data

SENAMHI produces WRF-CHEM (Weather Research and Forecast with Chemistry) simulations for air quality forecasts in Lima at 5 km spatial resolution. WRF-CHEM is a next generation CTM (atmospheric chemical transport model) developed by NOAA (National Oceanic and Atmospheric Administration) and NCAR (National Center for Atmospheric Research) [24]. CTMs simultaneously simulates the emission, 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 [24]. WRF-CHEM data outputs were packaged in monthly files with 26 vertical levels in the atmosphere every 6 hours (00:00, 06:00, 12:00 and 18:00 UTC). Parameters including in each WRF-CHEM files include: surface pressure, temperature, u- and v- wind components, simulated $PM_{2.5}$, and planetary boundary layer height (PBL, a measure of earth's lower atmosphere where

surface radiative forces causes turbulent mixing of chemicals) in HDF format. WRF-CHEM HDF files were processed through Interactive Data Language (IDL) to extract daily averaged measurements of parameters of interests at layer 0 (the lowest layer in the 26 vertical levels of the atmosphere). Parameters of interest including pressure, PBL, $PM_{2.5}$ aerosol dry mass, precipitation, temperature, and both wind components (u and v) were each extracted and analyzed for monthly, seasonal and yearly trends from 2010 to 2016 and for vertical level 0, 4, 9, 14, 19 and 24. For the purpose of modeling $PM_{2.5}$, only parameters in vertical level 0 were used as it was the level closest to earth's surface. Furthermore, due to the coarseness of the data at 5km spatial resolution, interpolation to 1 km spatial resolution using an inverse distance weighting method in statistical software R was used to create a smoother surface of the WRF-CHEM parameters. Figure 6 shows the contrast of temperature before and after interpolation on April 1 2015. Table 4 shows the correlations between ground measurements of $PM_{2.5}$ and WRF-CHEM simulated $PM_{2.5}$ for all SENAMHI stations, with N representing the number of days with available data between 2014 and 2016. Figure 7 shows the yearly average simulated $PM_{2.5}$ in $\mu g/m^3$ from the WRF-CHEM output. Figure 8 shows a comparison between daily, monthly mean, and yearly mean PBL height in meters from WRF-CHEM in vertical layer 0. a sample of the monthly and yearly average of the parameters produced by WRF-CHEM. Figures 9, 10, and 11 shows the monthly average concentrations of $PM_{2.5}$ in $\mu g/m^3$ in 2014 for vertical layers 0, 4, and 9, respectively. Figure 12 shows the time-series comparison of three ground monitors from the SENAMHI sites with the WRF-CHEM simulated $PM_{2.5}$ in $\mu g/m^3$ from 2014 to 2016.

2.2.4 Forecast Model

Data from the ECMWF (European Centre for Medium-Range Weather Forecasts) was used to compare and supplement the output from WRF-CHEM simulations. ECMWF is an independent intergovernmental organization with membership from 34 countries [25]. ECMWF was established in 1975 to produce numerical weather forecasts and currently archives the data freely for public use [25]. Data for 28 parameters including dew point, temperature, wind and pressure was downloaded from the ECMWF archive (http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/) in HDF format at the highest resolution available, 12.5km. Extraction of parameters to daily values was done using IDL, and interpolated to 1km spatial resolution using inverse distance weighting in statistical software R. Figure 13 shows the contrast of temperature before and after interpolation on April 1 2015. Additionally, a daily average was calculated for each variable and a time series analysis to assess monthly, seasonal and yearly trends was performed. 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 through an equation from University of Miami's Rosenstiel School of Marine & Atmospheric Science (http://andrew.rsmas.miami.edu/bmcnoldy/Humidity.html). ECMWF's parameters including temperature and calculated relative humidity was used in the process of gap-filling the MAIAC AOD data as mentioned in Bi et al. [23].

2.2.5 Miscellaneous Data

2.2.5.1 Elevation

Elevation information for the study region was downloaded from EARTHDATA (https://search.earthdata.nasa.gov/search). Advanced Spaceborne Thermal Emission and

Reflection Radiometer Global Digital Elevation Map (ASTER GDEM) is a satellite image product released through the joint collaboration between the Ministry of Economy, Trade, and Industry (METI) of Japan and the United States National Aeronautics and Space Administration (NASA) on October 17, 2011 [22]. ASTER GDEM data for Lima was downloaded as four Geo TIFF segments and was compiled in ArcGIS. After compilation, ASTER GDEM data was converted from a raster to points, where each point represents a 1km pixel with the elevation height in meters as an attribute. Finally, the elevation shapefile was clipped and spatially joined to the grid cells in the study domain. Figure 14 shows the map of the elevation of Lima, Peru by 1km pixels.

2.2.5.2 Population Density

The Oak Ridge National Laboratory produces yearly global ambient population distribution data using the LandScan algorithm comprising of spatial data and imagery analysis technologies along with a multi-variable dasymetric modeling approach to disaggregate census counts within an administrative boundary. Yearly LandScan data was downloaded for years 2010 to 2016 through the LandScan website (http://wms.cartographic.com.proxy.library.emory.edu/landscan/portal.aspx) as raster files and processed using ArcGIS. Each annual LandScan raster file was converted from raster to point in GIS and clipped to the size of the study region including the 10 km buffer. Each clipped LandScan shapefile was then spatially joined (point to point) to the study region. Subsequently, population density was calculated by dividing the total number of people within each pixel by the area of each pixel. Figure 15 shows Lima's population density in 2010.

2.2.5.3 Meteorology

Ground meteorological data was downloaded from the Weather Underground website for four personal weather stations along with one airport station. Table 5 contains the information and data availability of each weather station. Historical data for each station was retrieved through a dropdown menu on the Weather Underground website (https://www.wunderground.com/personal-weather-station/dashboard?ID=STATIONID#history). Data processing to convert HTML format to comma delimited was done in statistical software R. Meteorological data was used to correlate ground $PM_{2.5}$ observations from SENAMHI stations, MAIAC AOD and parameters from both WRF-CHEM and ECMWF. For $PM_{2.5}$ ground observations, each ground $PM_{2.5}$ monitor was linked to the closest weather station. For satellite remote sensing data, the nine pixels in a 3x3 km square area that was closest to a weather station was averaged for comparison. Both WRF-CHEM and ECMWF data were processed in the same manner for correlation analyses.

2.2.5.4 Land Use Information

Land use parameters were taken from GlobeLand30, a 30-meter Global Land Cover Dataset. GlobeLand30 is a product from the National High Technology Research and Development Program of China from the Ministry of Science and Technology of China. Furthermore, it is a 30-meter spatial resolution mapping-product for 2010 derived from remote sensing images through Landsat images downloaded from the U.S. Geological Survey and through the HJ-1 satellite images retrieved from the China Centre For Resources Satellite Data and Application. The land use parameters contained within the product consisted of 10 categories: cultivated land, forest, grassland, shrubland,

wetland, water bodies, tundra, artificial surfaces (urban areas), bareland, and permanent snow and ice. The major land use parameter to be used for the modeling $PM_{2.5}$ is percent urbanization, which was calculated by first conducting an unsupervised reclassification of the land use parameters in ArcGIS from ten categories down to four categories (open shrubland, bare/sparse vegetation, water bodies, and artificial/urban areas). Afterwards, the land use product was overlaid and spatially joined to the 1x1 km study domain to calculate the percent of urbanization by dividing the area of artificial areas by the entire area of each 1 km pixel. Figure 16 shows the map of the land use categories after reclassification, and Figure 17 shows the maps of percent urbanization for each pixel.

2.2.5.5 NDVI Data

Normalized difference vegetation index (NDVI) is a MODIS vegetation index product produced at 16-day intervals. It provides consistent spatial and temporal comparisons of vegetation canopy greenness, and effectively characterizes the global range of vegetation states and processes. The vegetation indices are retrieved from daily, atmosphere-corrected, bidirectional blue, red, and near-infrared surface reflectance based on a MODIS-specific composition methods [28]. Low quality pixels from the surface reflectance resulting from water, clouds, heavy aerosols, and cloud shadows are first removed and the remaining good-quality pixels are used to calculate an NDVI value that best represent the composition period [28]. NDVI data at 500 meter 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/search/) for years 2010 to 2016 in HDF format. IDL was used to extract 16-day interval data within the latitude and longitude range of

the study domain and further processing was conducted in R statistical software. Since NDVI data has 16-day intervals, each 15 days preceding the day with measured NDVI was given the same NDVI values. Finally, daily NDVI values were merged with the study domain through the use of R software. Figure 18 shows an example map of NDVI values on the 361st day of 2016.

2.2.5.6 Road Network Data

Road Network Data was downloaded as an ArcGIS-ready shapefile from the OpenStreetMap project through Geofabrik (http://download.geofabrik.de/south-america/peru.html). Geofrabrik is a consulting and software development firm based in Karlsruhe, Germany that specializes in OpenStreetMap services. The OpenStreetMap project is a free mapping and data service built by volunteers. The road network map was clipped to the area of the study region in ArcGIS and reclassified into three classes: motorways, primary and trunk roads, and secondary and tertiary roads. For each road network class, a distance was calculated between each study domain pixel to the nearest segment of road based on class. Figure 19 shows the road network for primary and trunk roads with the correspondence nearest distance in meters.

2.2.5.7 Cloud Fraction Data

Cloud fraction data, fraction of clouds covering each pixel, was used in the MAIAC AOD gap-filling processes. Daily data for cloud fraction at 5km 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/search/order/2/MOD06_L2--6) 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. [23].

2.3 Modeling Approach

2.3.1 LME Model

A linear mixed effects (LME) model was used to fit predictors to 8,491 $PM_{2.5}$ ground observations (6,410 observations from SENAMHI stations and 2,081 observations from the Checkley sites). The LME model includes a month-specific random intercepts and slopes for relative humidity and PBL (both of which are time-varying variables) to account for the temporally varying relationship between $PM_{2.5}$ and humidity and between $PM_{2.5}$ and vertical mixing height. Furthermore, the LME model also includes a day-specific random intercept and slope for AOD to account for the temporally varying relationship between $PM_{2.5}$ and AOD. The LME model allows incorporation of both fixed-effects terms and random-effects terms to account for different parameters. The fixed-effects affect parameters such as population density, elevation, road distance and NDVI, which are mostly static over time. In contrast, the random-effects affects parameters that are associated with certain sampling procedures and contribute to the covariance structure of the data. Due to small sample size compared to the number of parameters, all parameters used in the model were centered and scaled to allow model convergence. This performance-driven LME model can account for the day-to-day variability in the $PM_{2.5}$-AOD relationship by generating a daily AOD slope for each monitoring site on each day. Furthermore, the model can also account for the month-to-month variability in the $PM_{2.5}$-relative humidity and $PM_{2.5}$-PBL relationship by generating a monthly relative humidity and monthly PBL slope for each monitoring site for each month. Equation 1 shows the model structure of the LME model.

(1)

$$PM_{2.5,st} = [\mu + (\mu'_{Month}, \mu'_{Day})] + (\beta_1 + \beta_1'_{Day})AOD_{st} + (\beta_2 + \beta_2'_{Month})RH_{st}$$
$$+ (\beta_3 + \beta_3'_{Month})PBL_{st} + \beta_4 PM_s + \beta_5 Wind\_U_s + \beta_6 Wind\_V_s + \beta_7 Temp_s + \beta_8 NDVI_s$$
$$+ \beta_9 SP_s + \beta_{10} AL_s + \beta_{11} LCC_s + \beta_{12} SSRD_s + \beta_{13} Dist\_3_s + \beta_{14} Elev_s + \beta_{15} Pop_s$$
$$+ \beta_{16} Perurb_s + \varepsilon$$

where $PM_{2.5,st}$ is the measured $PM_{2.5}$ concentration in $\mu g/m^3$ at site s on day t; $\mu + (\mu'_{Month}, \mu'_{Day})$ are the fixed and random intercepts; $(\beta_1 + \beta_1'_{Day})AOD$ is the day-specific fixed and random effects of MAIAC gap-filled AOD; $(\beta_2 + \beta_2'_{Month})RH$ is the month-specific fixed and random effects of interpolated ECMWF relative humidity (in percent); $(\beta_3 + \beta_3'_{Month})PBL$ is the month-specific fixed and random effects of interpolated ECMWF planetary boundary layer height (vertical mixing depth) in meters; *PM* is the interpolated WRF-CHEM simulated $PM_{2.5}$ concentrations in $\mu g/m^3$; *Wind_U* is the interpolated ECMWF wind-u component; *Wind_V* is the interpolated ECMWF wind-v component; *Temp* is the interpolated WRF-CHEM temperature in Kelvins; *NDVI* is the normalized difference vegetation index; *SP* is the interpolated ECMWF surface pressure in pascal; *AL* is interpolated ECMWF albedo (unitless); *LCC* is interpolated ECMWF low cloud cover in percent; *SSRD* is interpolated ECMWF surface solar radiation downwards (in J/m^2); *Dist_3* is the distance (in meters) to the nearest secondary/tertiary road; *Elev* is the elevation (in meters); *Pop* is the population density (number of people per kilometer square); and *Perurb* is the percent urbanization.

2.3.2 Random Forest Model

A random forest model was also used to fit predictors to the same dataset used for the linear mixed effects model. A random forest model is comprised of a set of decision trees constructed from the best split for each node among a subset of predictors randomly chosen at that node [26]. The two main parameters in a random forest model is m_{try} and n_{tree}, which stands for the number of predictors sampled for splitting at each node and the number of trees grown, respectively. Algorithm for the random forest model works by first drawing n_{tree} bootstrap-samples from the model fitting dataset [26]. Subsequently, the algorithm grows an unpruned classification or regression tree with m_{try} of predictors randomly sampled for each bootstrap-sample and the best split is consequently chosen at each node [26,27]. Predictions are then made by aggregating the predictions of n_{tree} trees (e.g., a simple majority vote for classification and average for regression) [27]. In a random forest model, the error rate is calculated using predictions of "out-of-bag" samples, which are the data samples not in the bootstrap sample [27]. Comparison of results with different settings of m_{try} and n_{tree}, was conducted to achieve the best prediction accuracy. Variables used in the random forest model is the same as those used in the LME model, with m_{try} set at 6 and n_{tree} at 1000.

2.3.3 Cross Validation and Predictions

A 10-fold cross-validation (CV) process was carried out on both the LME and random forest model in the same manner to validate the prediction results from both models. The model fitting dataset consisting of 8,491 ground observations were randomly divided into 10 segments or subsets 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 10^{th} 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. A correlation between the predictions and the original ground observations is conducted to produce a CV R^2. After cross-validation, daily datasets consisting of the set of variables used in the LME and random forest models except for ground measurements were created to make predictions. Predictions were made using the predict function in statistical software R using both models on the same daily datasets. Once predictions were made, daily files were aggregated to the monthly and yearly level for mapping using ArcGIS.

3. Results

3.1 Descriptive Statistics

$PM_{2.5}$ concentrations from the DIGESA network were found to be unreliable, reducing the model fitting R^2 when included in the model, and were consequently omitted from all analyses. Histograms of all the predictors used in the modeling approach can be seen in Figure 20. Most of the parameters including albedo, planetary boundary layer height, AOD, NDVI, simulated $PM_{2.5}$, surface solar radiation, temperature, and both wind components are unimodal and log linearly distributed. Relative humidity and surface pressure is bimodal, suggesting that the distribution of the monitors may play a role in the distribution of these parameters. Parameters including elevation, road distance, percent urbanization, and population density are not normally distributed due to the likely nonrandom placement of the ground monitors, especially the lack of spatial distribution as a direct result of the clustered Checkley sites. The overall mean $PM_{2.5}$ concentration in $\mu g/m^3$ of each monitor site within their respective network is shown in Table 6. The mean $PM_{2.5}$ concentration for all combined sites used in the modeling dataset was 23.6 $\mu g/m^3$. There is no distinctive secular trends in $PM_{2.5}$, both on the yearly and seasonal level. For most of the SENAMHI stations, $PM_{2.5}$ levels tend to start low in months January while rising during the months of May to September and then dip slightly down again during November and December. Nonetheless, this is only a suggestive indication and not all monitors follow this trend. The Checkley monitors show a slightly different trend, with $PM_{2.5}$ concentrations peaking around April and decreasing during the months between Jun and October before slightly increasing during November and December. These trends may be due to the fact that Checkley monitors are only 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 which are located in the central region of Lima. Furthermore, trends for Checkley monitors are only available from late 2011 to early 2013 while trends for SENAMHI are observed from 2014 to 2016, so a fair and continuous comparison of seasonal and yearly trends may not be conducted between the two monitor networks.

3.2 Model Fitting and Validation

The model described in Equation 1 is the linear mixed effects model used to fit predictors of $PM_{2.5}$ to ground observations from both the SENAMHI and Checkley sites. AOD was allowed to have daily random effects while relative humidity and planetary boundary layer height is expected to vary but not significantly through the month and as a result was set on the random effects at the monthly level. Overall, the regression R^2 for the LME model was 0.63 and the cross-validation (CV) R^2 and RMSE is 0.58 and 7.08 $\mu g/m^3$, respectively. Figure 21 shows a density plot of the correlations between predicted and measured $PM_{2.5}$ values from the cross-validation of the LME model. Table 7 shows the beta coefficients, standard error, degrees of freedom, t-value, and p-value for each parameter. All predictors except wind U-component, temperature, NDVI, and relative humidity, were highly significant. Wind U-component and temperature were parameters from the WRF-CHEM simulation, and were not highly correlated with the ground observations from Weather Underground. Therefore, insignificance of these parameters were expected. NDVI is the normalized vegetative index, categorizing the vegetative canopy of the particular area from negative one with no vegetative canopy to one with full vegetative canopy. Since air monitors are centrally placed in urban environments,

with NDVI staying constant over time, it is therefore expected that NDVI would not be a significant predictor of $PM_{2.5}$ in the LME model.

A random forest model was used to fit predictors of $PM_{2.5}$ to ground observations from both the SENAMHI and Checkley sites. The random forest model was specified with a *nodesize* of 6, *maxnode* of 2048, m_{try} of 6 and the n_{tree} at 1000. The "out of bag" R^2 from the random forest model using the entire dataset is 0.73 with an RMSE of 5.61 $\mu g/m^3$, with a cross-validation (CV) R2 and RMSE of 0.73 and 5.66 $\mu g/m^3$, respectively. Figure 22 shows a density plot of the correlations between predicted and measured $PM_{2.5}$ values from the cross-validation of the random forest model. Table 8 shows the name of each predictor along with the importance, or percent increase in MSE. Although random forest is a "black-box" machine learning method, the importance output is a measure of parameter predictive power based on a permutation test [26]. 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 degradation in prediction accuracy [26]. The higher the importance or percent increase in MSE, the higher the predictive accuracy for that variable [27]. The random forest model indicates that temperature, albedo, and surface solar radiation are the most important predictors of $PM_{2.5}$. This is in direct contrast to the LME model, which indicates that temperature is not a good predictor of $PM_{2.5}$. This is a result of the random forest method compartmentalizing and categorizing temperature and not analyzing this variable as a continuous variable. The random forest model also indicates that percent urbanization, elevation, and residential road distance has the least predictive accuracy of $PM_{2.5}$. This is in direct conjunction with the histograms shown in Figure 20, which indicates that these parameters are not normally

distributed and likely led to inconsistencies in compartmentalizing and categorizing these variables. Figure 23 and 24 shows the time series between estimated $PM_{2.5}$ concentrations using the random forest model and ground $PM_{2.5}$ observations from each monitor aggregated to the monthly mean in 2012 for the Checkley sites and 2015 for the SENAMHI sites, respectively. Figure 25 shows a map of the mean concentration of each monitor station in the study domain next to the mean estimated concentration of each monitor from the cross-validation results.

3.3 Prediction of $PM_{2.5}$ Concentrations

Due to time constraints, prediction maps of $PM_{2.5}$ based on the LME model have not been finished and are not included at this time. When completed, these figures will be inserted in the appendix. The predicted annual mean $PM_{2.5}$ concentrations in $\mu g/m^3$ using the random forest model are shown in Figure 26. Figure 27 shows the monthly mean $PM_{2.5}$ concentrations in $\mu g/m^3$ of 2015 using the random forest model. Due to data availability from WRF-CHEM along with available days with MAIAC gap-filled AOD (some days are missing AOT values due to cloud cover), daily predictions of $PM_{2.5}$ started on the 61st day of 2010 or March 2 2010. The last day of predictions is on the 366th day of 2016 of December 31 2016. Mean $PM_{2.5}$ concentrations range from 14.62 to 44.32 $\mu g/m^3$. Predictions from the random forest 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 in the mountains. The annual prediction maps suggest that $PM_{2.5}$ concentrations at lowest in the valleys and urban areas while highest in the mountains and remote areas of Lima, Peru.

4. Discussion and Conclusion

The LME model achieved decent fit of the ground monitor data with an R^2 of 0.63 while the random forest model achieved better fit with an R^2 of 0.73. The CV R^2 from both models (0.58 for LME model and 0.73 for random forest model) suggest that overfitting is not likely a serious issue. Due to time constraints, predictions of the LME model have not been mapped and comparison of $PM_{2.5}$ spatial distribution from both models cannot be completed at this time. Nonetheless, important or significant variables are not consistent between the two models, which warrants further investigation into the accuracy of each model's algorithm. Additionally, the slope and intercept of the LME model does show a better-fitted line with an intercept at nearly zero and a slope of one compared to the intercept of negative two and a slope of one from the random forest model. Although the LME model achieved a better intercept and slope in the fit between estimated $PM_{2.5}$ and measured $PM_{2.5}$ concentrations in the cross-validation dataset, the drop between model fitting and cross-validation R^2 (from 0.63 to 0.58) indicates that the model may not be accurately estimating $PM_{2.5}$. Conversely, while the intercept and slope for the fit between estimated and measured $PM_{2.5}$ concentrations of the random forest's cross-validation dataset is not as perfect as the LME model, the model fitting R^2 value does not differ from the cross-validation R^2. Furthermore, results of the slopes and intercepts between the two models indicate that the random forest underestimates $PM_{2.5}$ levels more compared to the LME model.

One of the limitations for these models is the uneven distribution of the ground monitors across Lima, Peru. All monitors were located around the urbanized city with no monitors near the rural areas and up in the Andes Mountains. The 15 Checkley monitors

were clustered all within a few kilometers of each other in the southern region of the study domain, affecting their predictive capabilities on the rest of the study domain. Furthermore, many of the ground monitors for both DIGESA and SENAMHI lack daily measurements, influencing the temporal distribution of $PM_{2.5}$ ground measurements in the model fitting dataset. Additionally, the Checkley ground measurements was collected only from late 2011 to early 2013 while the SENAMHI data was collected from mid-2014 through 2016, which impacts model predictive abilities. Although many of the ground monitors also recorded PM_{10} (particulate matter with aerodynamic diameter of 10 μm), due to time constraints, PM_{10} measurements cannot be converted to $PM_{2.5}$ measurements to maximize ground observations in the model fitting process.

Future research should focus on converting PM_{10} to $PM_{2.5}$ from both the SENAMHI and DIGESA monitors to maximize ground observations and bridge the gap between SENAMHI and Checkley datasets both spatially and temporally. Furthermore, intuition dictates that $PM_{2.5}$ concentrations should be highest in urbanized areas due to vehicular and factory emissions. However, our prediction maps indicate that $PM_{2.5}$ is lowest in urbanized areas and highest in the remote mountainous areas, warranting further investigation into this issue. A possibility might be population density or elevation driving the increase in $PM_{2.5}$ from low areas to high areas since mean measurements of the ground monitors also contains this pattern (as seen in Figure 25). Another possible solution may be to determine the mean planetary boundary height near the base of the Andes Mountains and restrict the study domain to this area.

5. References

[1]Mirabelli, M. C., Vaidyanathan, A., Flanders, W. D., Qin, X., & Garbe, P. (2016). Outdoor PM2.5, Ambient Air Temperature, and Asthma Symptoms in the Past 14 Days among Adults with Active Asthma. *Environmental Health Perspectives*, *124*(12), 1882-1890. doi:10.1289/EHP92

[2]WHO, 2014 World Health Organization Burden of Disease from the Joint Effects of Household and Ambient Air Pollution for 2012 (2014) http://www.who.int.proxy.library.emory.edu/phe/health_topics/outdoorair/databases/AP_jointeffect_BoD_results_March2014.pdf?ua=1

[3]Kim, K.-H., Kabir, E., & Kabir, S. (2015). A review on the human health impact of airborne particulate matter. *Environmental International, 74*, 136-143. Doi: https://doi.org/10.1016/j.envint.2014.10.005

[4]Xuefei, H., Waller, L. A., Lyapustin, A., Wang, Y., Al-Hamdan, M. Z., Crosson, W. L., Estes Jr., M. G., Estes, S. M., Quattrochi, D. A., Puttaswamy, S. J., Liu, Y. (2014). Estimating ground-level PM2.5 concentration in the Southeastern United States using MAIAC AOD retrievavls and a two-stage model. *Remote Sensing of Environment, 140*, 220-232. Doi: https://doi.org/10.1016/j.rse.2013.08.032

[5]Prieto-Parra, L., Yohannessen, K., Brea, C., Vidal, D., Ubilla, C. A., Ruiz-Rudolph, R. (2017). Air pollution, PM2.5 composition, source factors, and respiratory symptoms in asthmatic and nonasthmatic children in Santiago, Chile. *Environmental International, 101*, 190-200. Doi: https://doi-org.proxy.library.emory.edu/10.1016/j.envint.2017.01.021

[6]World Health Organization. WHO Global Urban Ambient Air Pollution Database (update 2016). Accessed 25 August 2017. Available from: www.who.int/phe/health_topics/outdoorair/databases/cities/en/.

[7]Goddard Space Flight Center. (2018). AERONET: Aerosol Robotic Network. Accessed April 11 2018. < https://aeronet.gsfc.nasa.gov/>

[8]Liu, Q., Xu, C., Ji, G., Liu, H., Shao, W., Zhang, C. … Zhao, P. (2017). Effect of exposure to ambient PM2.5 pollution on the risk of respiratory tract diseases: a meta-analysis of cohort studies. *Journal of Biomedical Research, 31*(2), 130-142. Doi: 10.7555/JBR.31.20160071

[10]WHO, 2015. Climate and Health Country Profile – 2015, Peru. Accessed: 07 July 2017. http://www.who.int/globalchange/resources/PHE-country-profile-Peru.pdf?ua=1

[11]Scholl, L., Guerrero, A., Quintanilla, O., & L'Hoste, M. C. (2015). Comparative Case Studies of Three IDB-Supported Urban Transport Projects. Inter-American Development Bank. Accessed: 07 July 2017. http://docs.trb.org/prp/16-6544.pdf

[12]Baumann, L. M., Robinson, C. L., Combe, J. M., Gomez, A., Romero, K., Gilman, R. H., … Hansel, N. N. (2011). 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, 127*(4), 875-882. Doi: https://doi.org/10.1016/j.jaci.2010.11.031

[13] Carbajal-Arroyo, L., Barraza-Villarreal, A., Durand-Pardo, R., Moreno-Macías, H., Espinoza-Laín, R., Chiarella-Ortigosa, P., & Romieu, I.(2007) Impact of Traffic Flow on the Asthma Prevalence Among School Children in Lima, Peru. *Journal of Asthma, 44*(3), 197-202. DOI: 10.1080/02770900701209756

[14]Ma, Z., Hu, X., Huang, L., Bi, J., & Liu, Y. (2014). Estimating Ground-Level PM2.5 in China Using Satellite Remote Sensing. *Environmental Science & Technology, 48*, 7436-7444. Doi: dx.doi.org/10.1021/es5009399

[15]Liu, Y., Paciorek, C. J., Koutrakis, P. (2009). Estimating Regional Spatial and Temporal Variability of PM2.5 Concentrations Using Satellite Data, Meteorology, and Land Use Information. *Environmental Health Perspectives, 117*(6), 886-892.

[16]Remer, L. A., Kaufman, Y., Tanré, D., Mattoo, S., Chu, D., Martins, J., … Kleidman, R. (2005). The MODIS aerosol algorithm, products, and validation. *Journal of Atmospheric Science, 62*(4), 947−973.

[17]European Environmental Agency (EEA). (2012). Particulate matter from natural sources and related reporting under the EU Air Quality Directive in 2008 and 2009. EEA Technical report, 10. Doi: 10.2800/55574

[18]Instituto Nacional de Estradistica e Informatica (INEI). 2010. Perú: Estimaciones y Proyecciones de Población total y edades quinquenales, según Departamento, Provincia y Distrito. Accessed 25 August, 2017. < http://proyectos.inei.gob.pe/web/biblioineipub/bancopub/Est/Lib1010/cuadros/d01011.xls >

[19]Earth System Research Laboratory, Global Monitoring Division. (N/A). Surfrad Aerosol Optical Depth. Accessed 25 August 2017. < https://www.esrl.noaa.gov/gmd/grad/surfrad/aod/>

[20]World Health Organization. Ambient (outdoor) air quality and health. (Updated 2016). Accessed March 19 2018. Available from: http://www.who.int/mediacentre/factsheets/fs313/en/

[21]California Environmental Health Tracking Program. (N/A). Air Quality: Measures and Limitations. Accessed April 11 2018. < http://www.cehtp.org/faq/air/air_quality_measures_and_limitations>

[22]Jet Propulsion Laboratory: California Institute of Technology. 2004. ASTER: Advanced Spaceborne Thermal Emission and Reflection Radiometer. Accessed April 11 2018. < https://asterweb.jpl.nasa.gov/gdem.asp>

[23]Bi, J., Wildani, A., Wang, Y., Lyapustin, A., Liu, Y. (2018). Incorporating Snow and Cloud Fractions in Random Forest to Estimate High Resolution PM2.5 Exposures in New York State. Under Reivew.

[24]Grell, G. A., Peckham, S. E., Schmitz, R.., McKeen, S. A., Frost, G., Skamarock, W. C., Eder, B. (2005). Fully coupled "online" chemistry within the WRF model. *Atmospheric Environment*, *39*, (37), 6957-6975.

[25]ECMWF. (N/A). About. Accessed April 12 2018. < https://www.ecmwf.int/en/about>

[26]Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5−32.

[27]Liaw, A., Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*(3), 18−22.

[28]Earth Observatory. (N/A). Measuring Vegetation (NDVI & EVI). Accessed April 16 2018. <https://earthobservatory.nasa.gov/Features/MeasuringVegetation/measuring_vegetation_2.php>

6. Tables and Figures

Figure 1. Maps of study domain and pixel categorization.

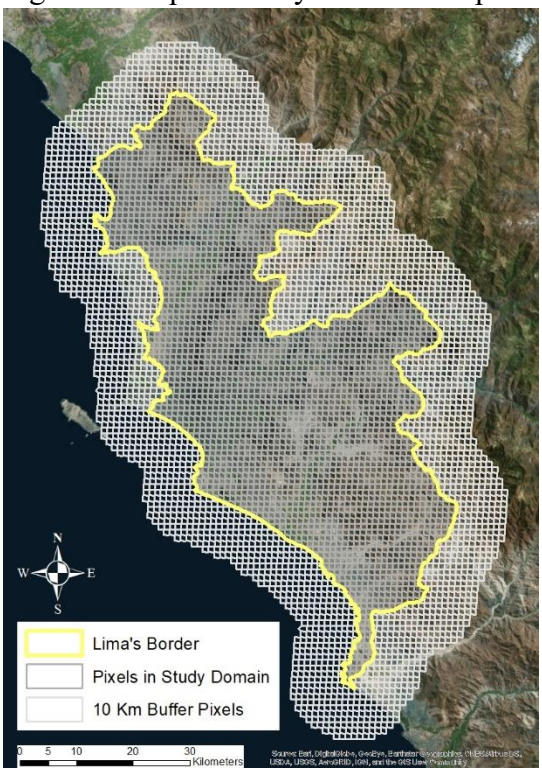


Table 1. Data availability of DIGESA monitors.

YEAR	DIGESA STATIONS					
	CALLAO	AUXILIADORA	LUZMILA	UNANUE	LA REPUBLICA	LINCE
2010	31	38	44	46	0	0
2011	14	3	15	18	0	0
2012	17	13	23	28	0	0
2013	40	37	26	54	9	0
2014	23	51	32	45	21	173
2015	18	52	15	44	22	34
2016	9	41	27	36	21	0
Total	152	235	182	271	73	207

Table 2. Data availability of SENAMHI monitors.

YEAR	SENAMHI STATIONS									
	ATE	SBJ	CDM	STA	VMT	HCH	SJL	SMP	CRB	PPD
2014	90	92	59	31	29	258	240	269	251	251
2015	263	293	300	341	158	344	361	304	312	327
2016	264	283	265	325	275	202	274	298	286	318
Total	617	668	624	697	462	804	875	871	849	896

Table 3. Data availability of Checkley sites.

Site	Monitor	N
Check_2	A2520, A2613, A2653, A2770	369
Check_7	A2760	455
Check_8	A2612	313
Check_9	A2210, A2497, A2628, A2723, A2977, P101	481
Check_10	A2715	313
Check_11	A2686, A2821	334

Figure 2. Map of study domain with location of network monitors.

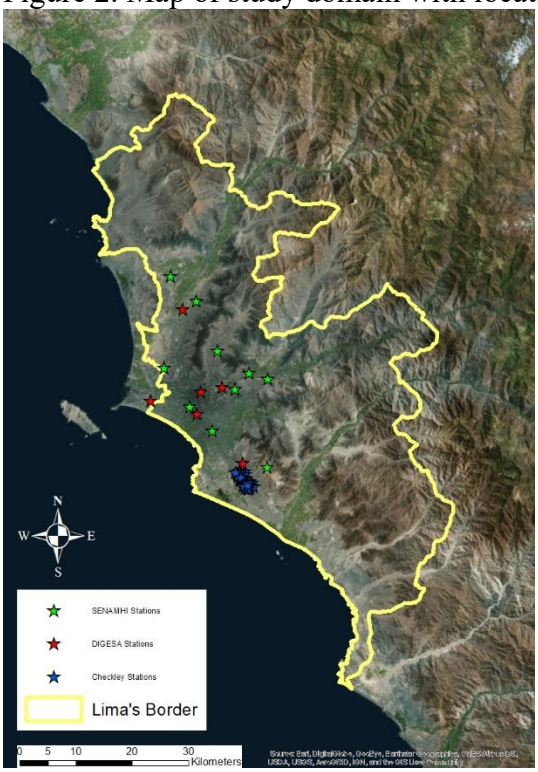


Figure 3. Time series of $PM_{2.5}$ concentrations in $\mu g/m^3$ at each monitoring station.

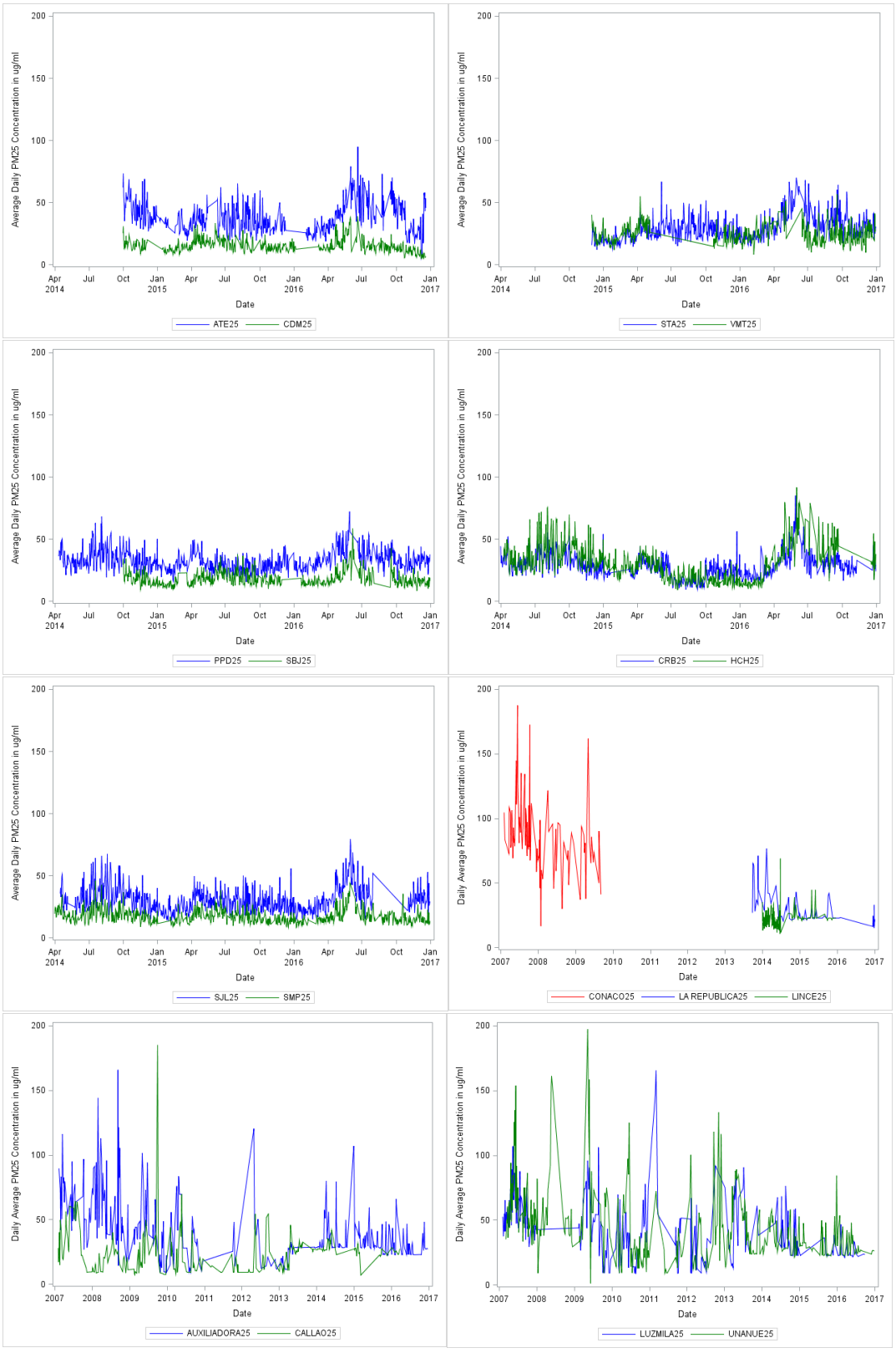


Figure 4. Mean annual gap-filled MAIAC AOD.

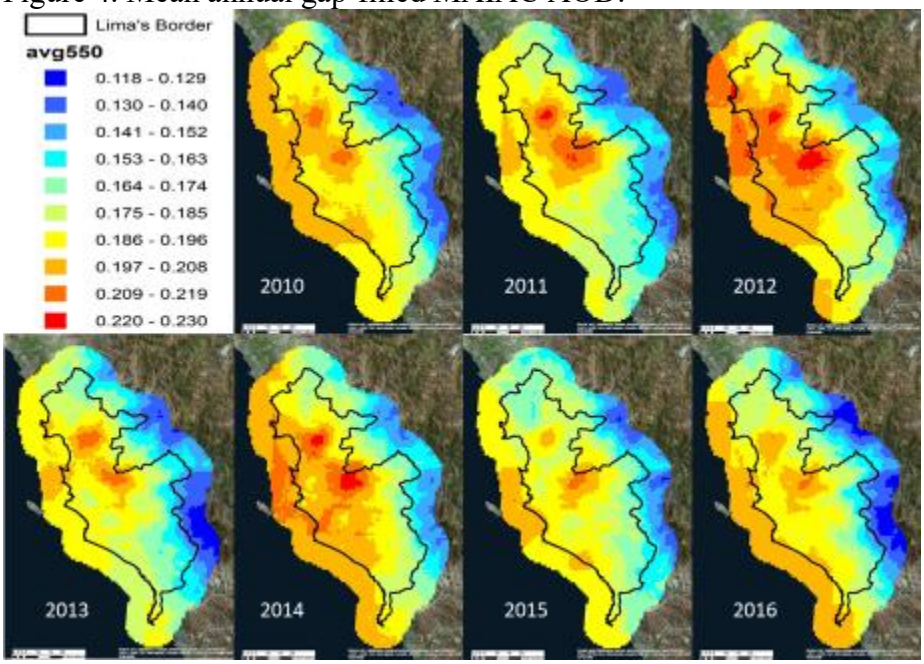
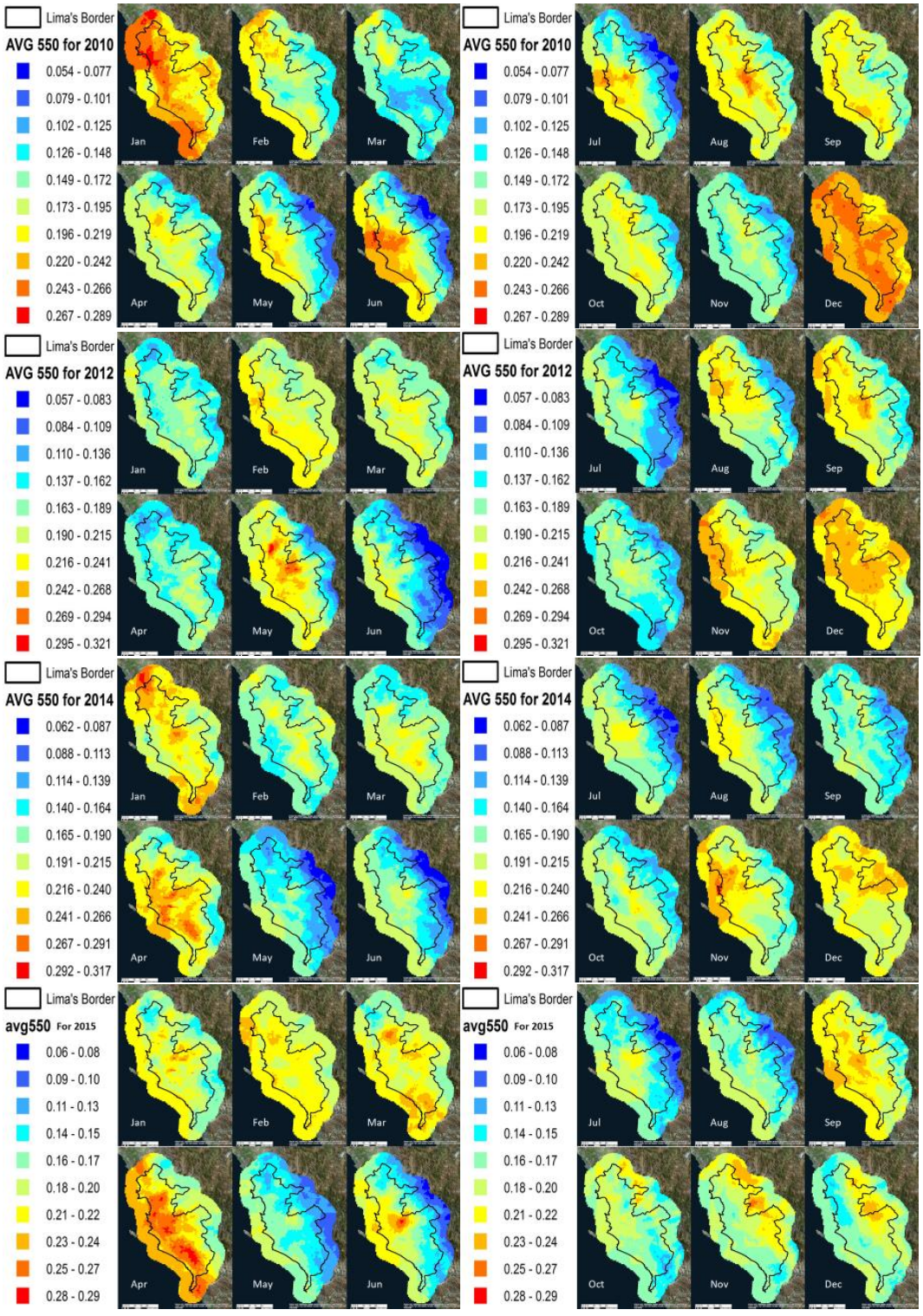


Figure 5. Monthly mean AOD for 2010, 2012, 2014, 2015, and 2016.



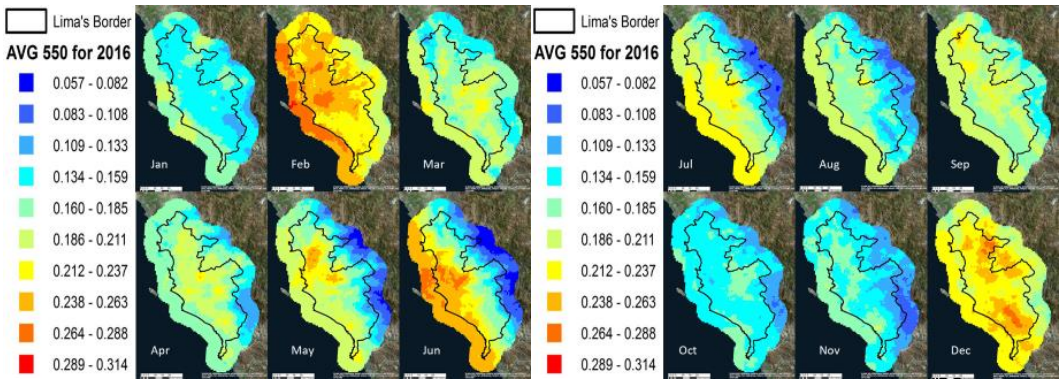


Figure 6. Maps of temperature (in Fahrenheit) on April 1 2015 before and after interpolation.

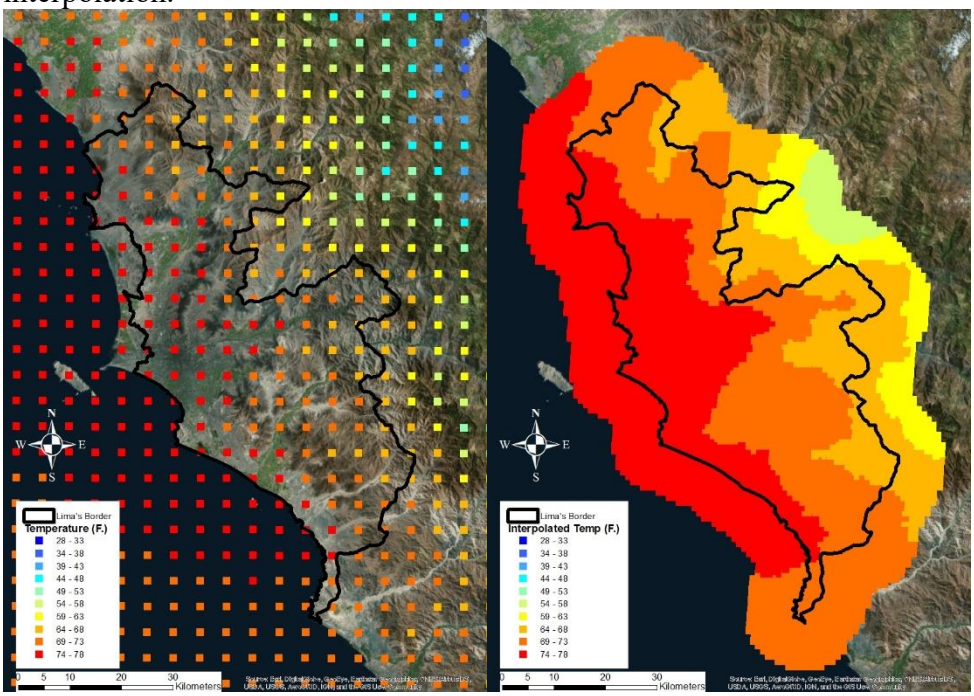


Table 4. Correlations between ground $PM_{2.5}$ measurements and simulated $PM_{2.5}$ from WRF-CHEM by SENAMHI station from 2014 to 2016.

Station	$PM_{2.5}$	N
ATE	0.01	321
SBJ	0.23	356
CDM	0.22	301
Station	0.23	310
VMT	0.09	141
HCH	0.21	542
SJL	0.32	535
SMP	0.21	515
CRB	0.15	498
PPD	0.18	523

Figure 7. Yearly average simulated $PM_{2.5}$ in $\mu g/m^3$ from WRF-CHEM in vertical layer 0.

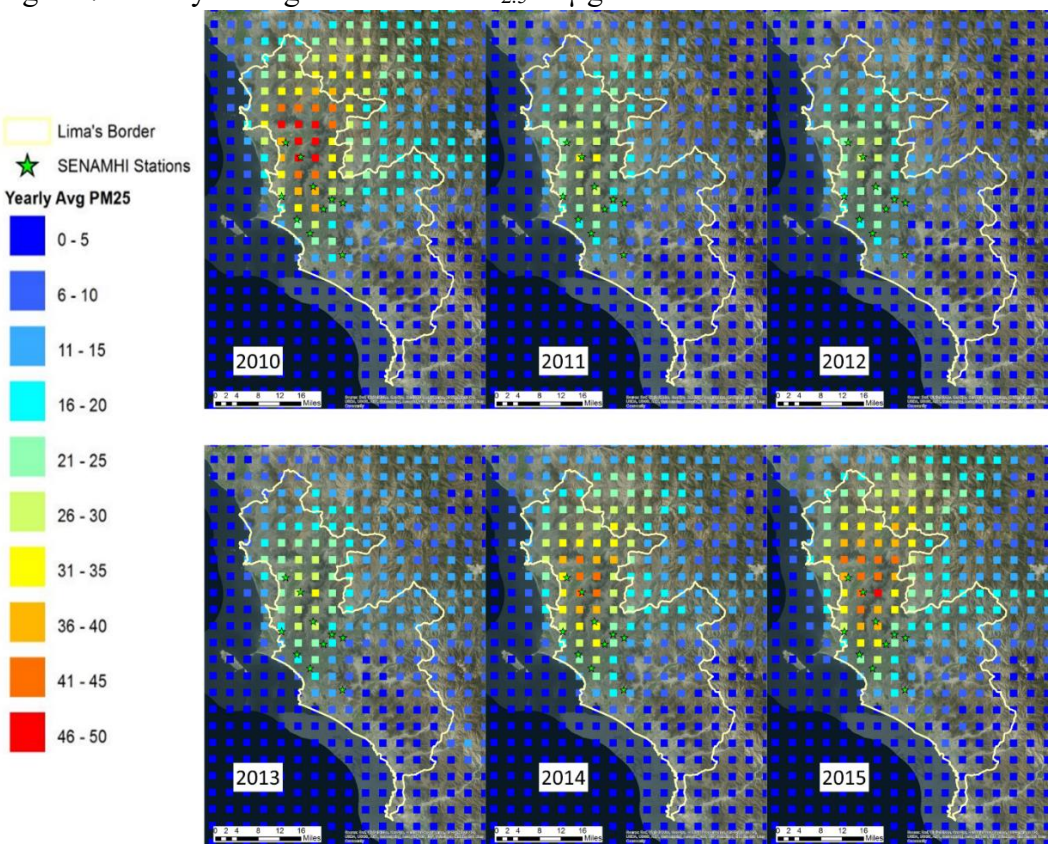


Figure 8. Average planetary boundary layer height in meters for 2014 at vertical layer 0.

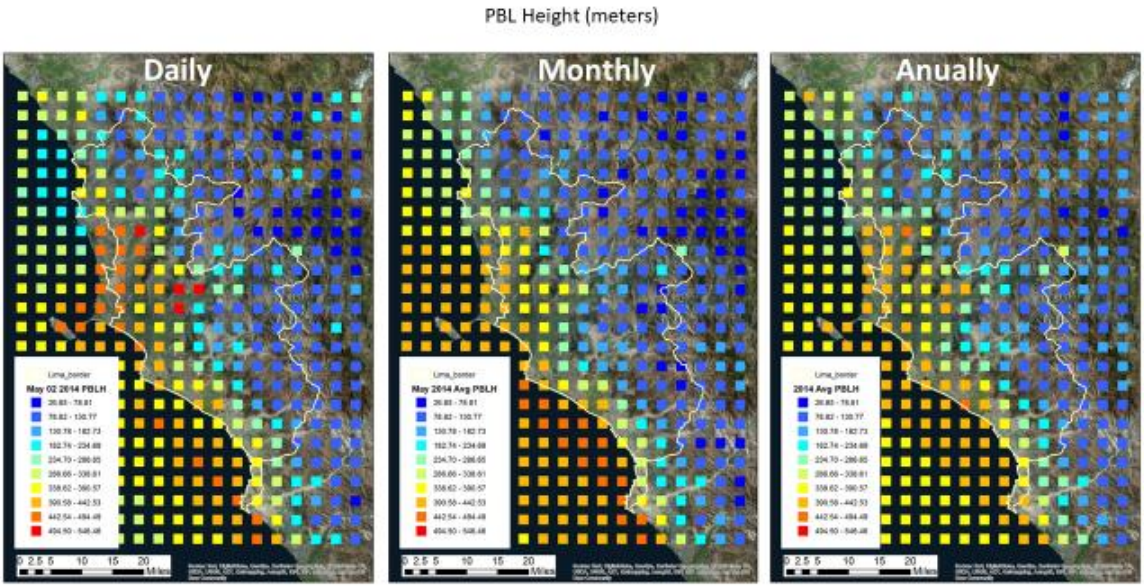


Figure 9. Monthly average concentrations of $PM_{2.5}$ in $\mu g/m^3$ in 2014 for vertical layer 0.

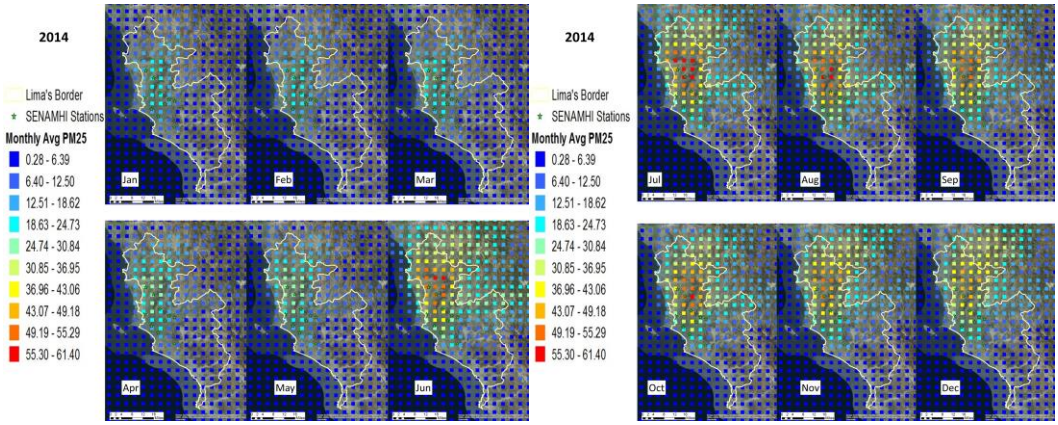


Figure 10. Monthly average concentrations of $PM_{2.5}$ in $\mu g/m^3$ in 2014 for vertical layer 4.

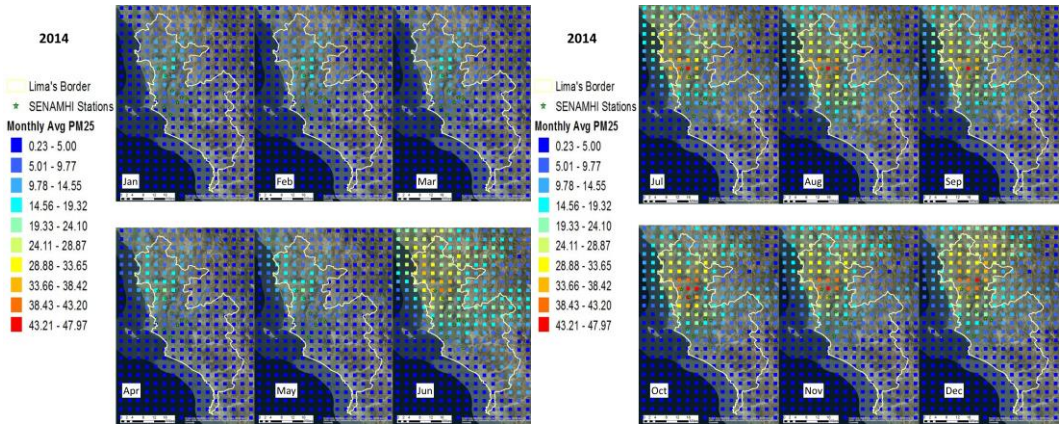


Figure 11. Monthly average concentrations of $PM_{2.5}$ in µg/m3 in 2014 for vertical layer 9.

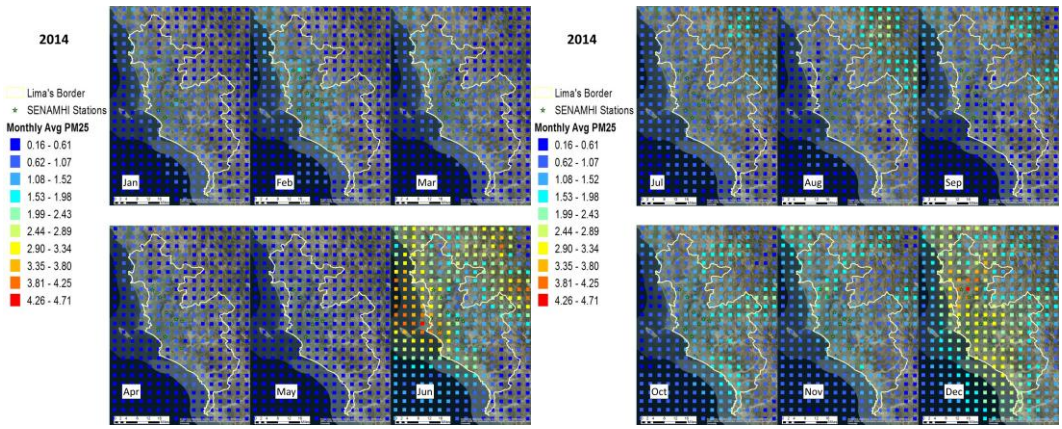
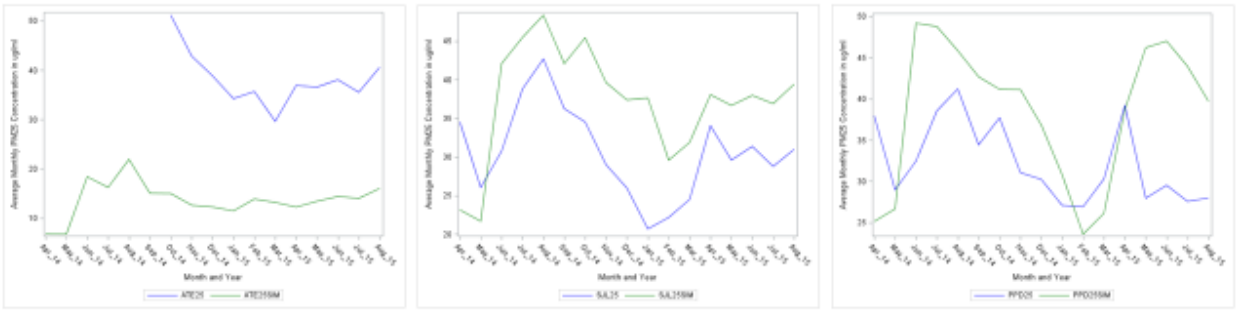


Figure 12. Time series a SENAMHI station located in the low-, middle-, and high-regions of corresponding WRF-CHEM $PM_{2.5}$ values in $\mu g/m^3$



Low Point Mid Point High Point

Figure 13. Maps of 2-Meter temperature (in Fahrenheit) of ECMWF data on April 1 2015 before and after interpolation.

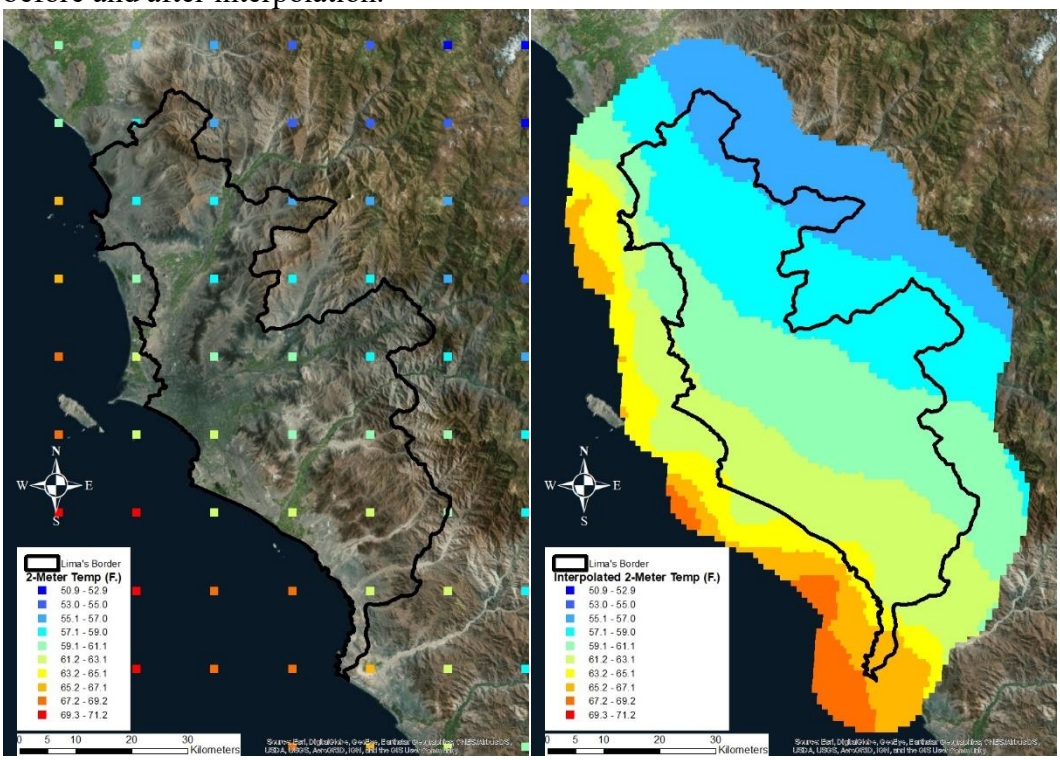


Figure 14. Elevation of study domain in meters.

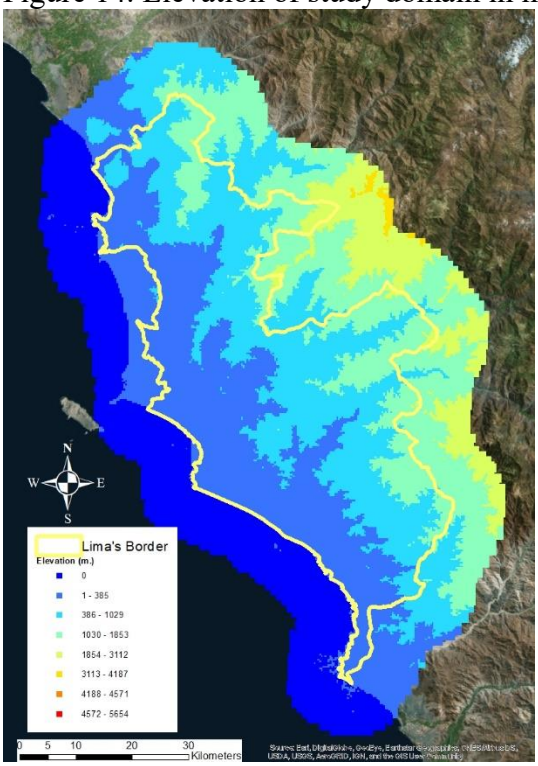


Figure 15. Lima's population density (number of people per square kilometer) in 2010.

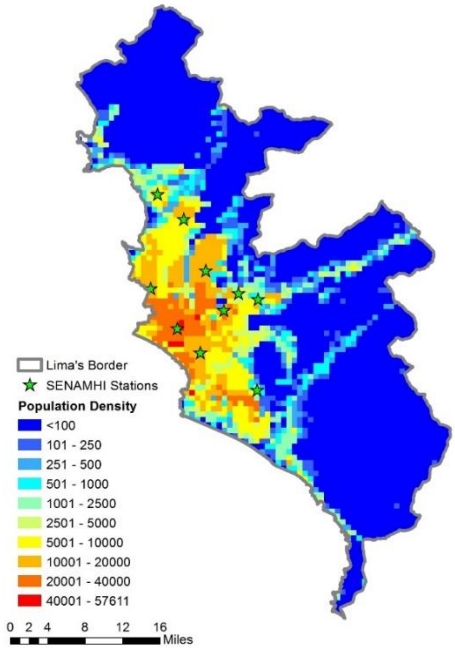


Table 5. Names of Weather Underground station and code for each personal weather station located within the study domain with their operational start and end date.

Station	Start Date	End Date
Jorge Chaves Int. (Airport)	May 30 1995	Present
Miraflores (ILIMALIM7)	April 25 2013	Present
Miraflores (ILIMAMIR3)	February 23 2012	Present
Santiago de Surco (ILIMALIM15)	February 27 2016	Present
El Remanso, La Molina (ILIMALIM12)	April 28 2015	Present

Figure 16. Map of reclassified land use categories.

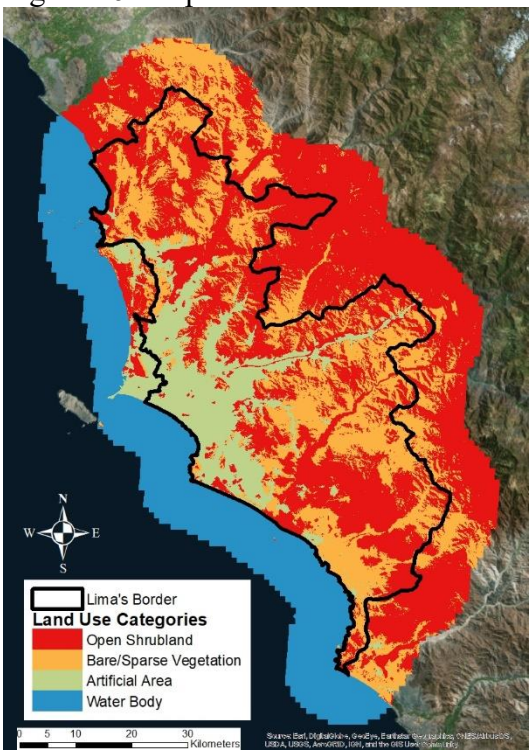


Figure 17. Map of percent urbanization calculated from reclassification of land use categories.

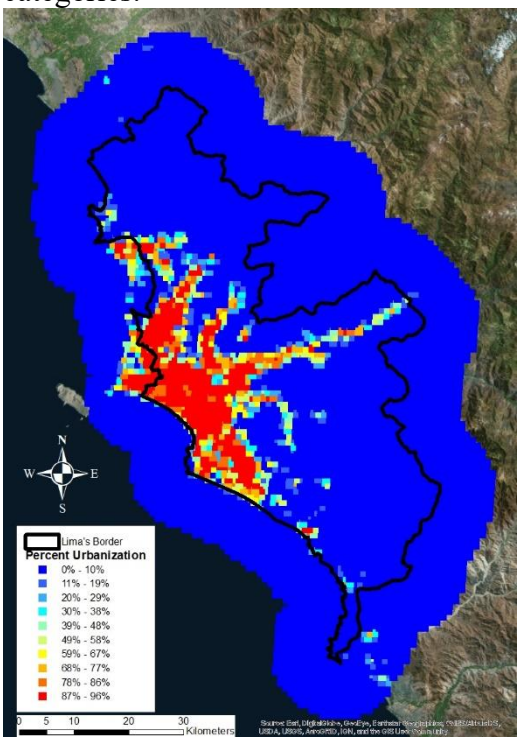


Figure 18. Map of NDVI values for December 26th 2016.

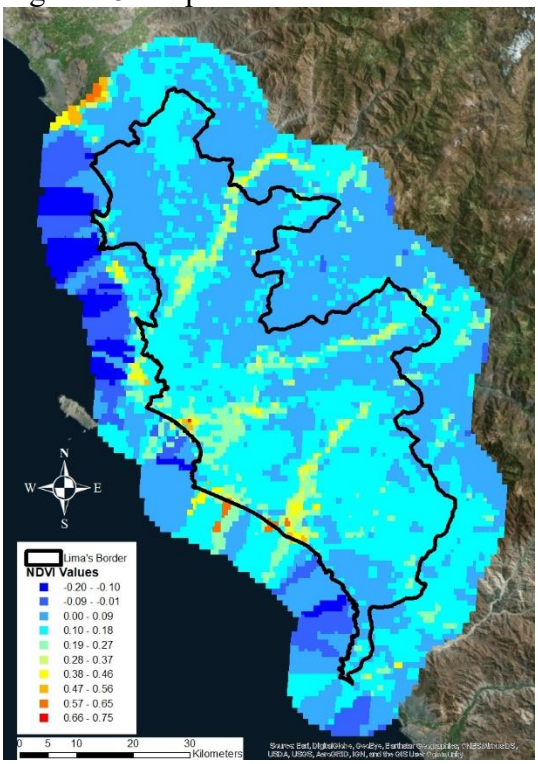


Figure 19. Map of primary and trunk roads with corresponding distance values in meters of each pixel to the nearest road segment.

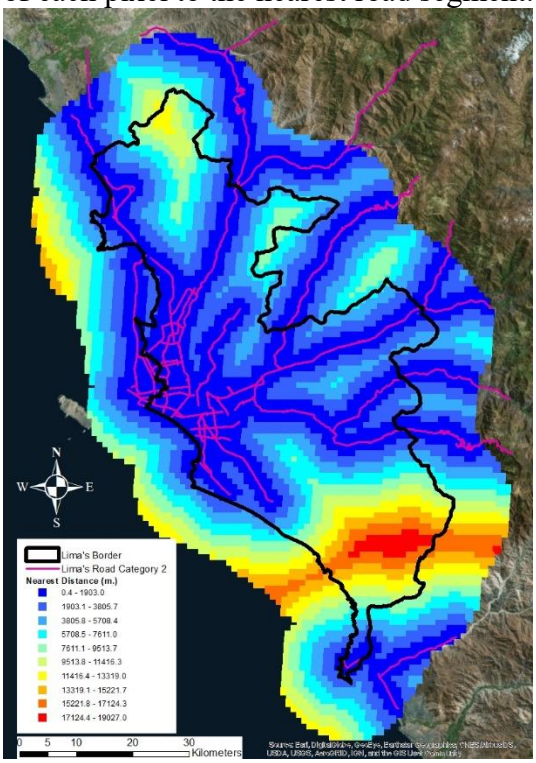
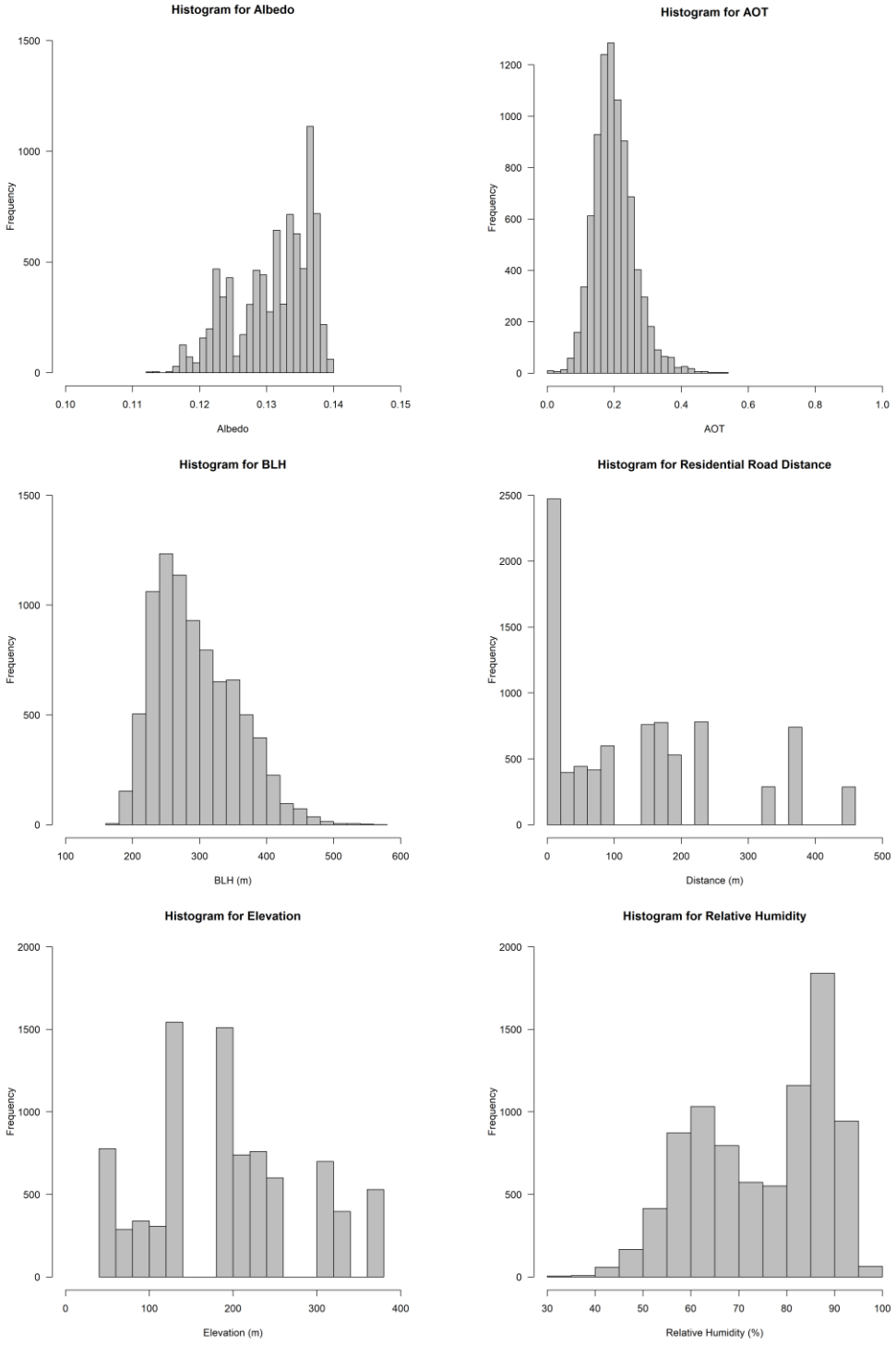
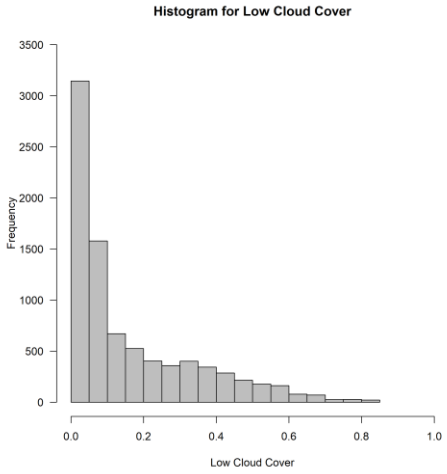
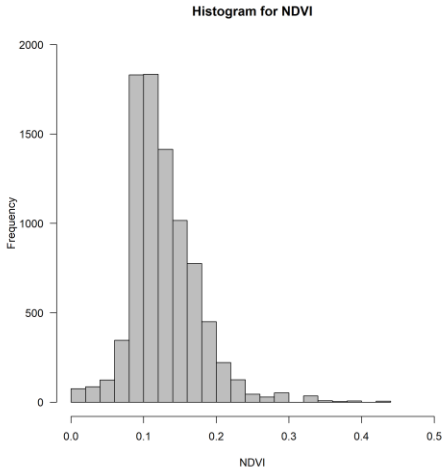
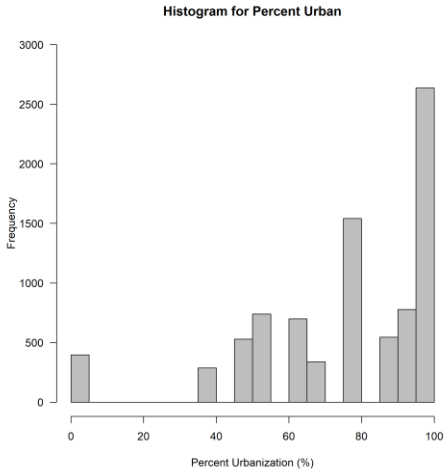


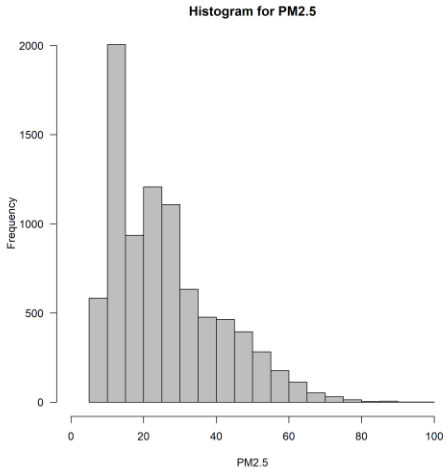
Figure 20. Histogram of each predictor used in both the LME and Random Forest Model.

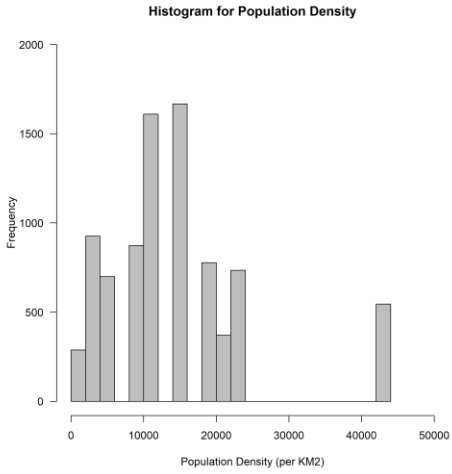


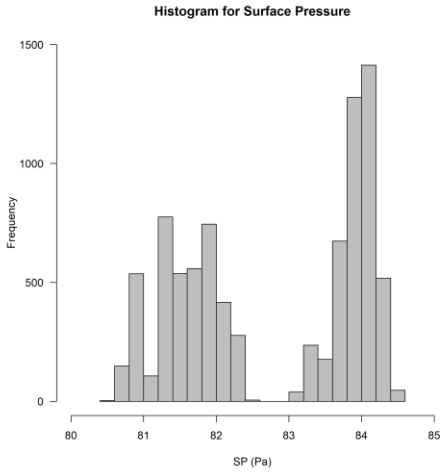

Histogram for Low Cloud Cover
Frequency
Low Cloud Cover

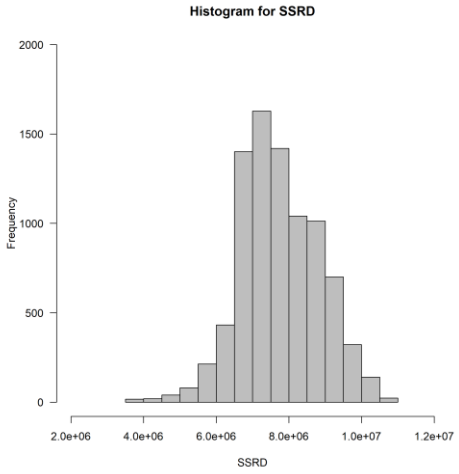

Histogram for NDVI
Frequency
NDVI

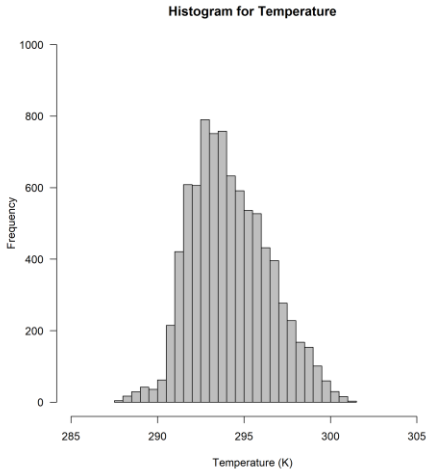

Histogram for Percent Urban
Frequency
Percent Urbanization (%)

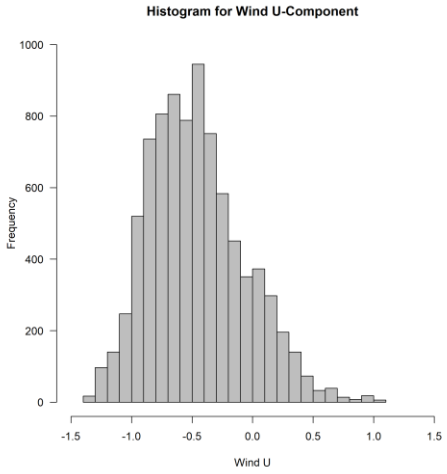

Histogram for PM2.5
Frequency
PM2.5


Histogram for Population Density
Frequency
Population Density (per KM2)


Histogram for Surface Pressure
Frequency
SP (Pa)


Histogram for SSRD
Frequency
0
500
1000
1500
2000
2.0e+06
4.0e+06
6.0e+06
8.0e+06
1.0e+07
1.2e+07
SSRD


Histogram for Temperature
Frequency
0
200
400
600
800
1000
285
290
295
300
305
Temperature (K)


Histogram for Wind U-Component
Frequency
0
200
400
600
800
1000
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Wind U

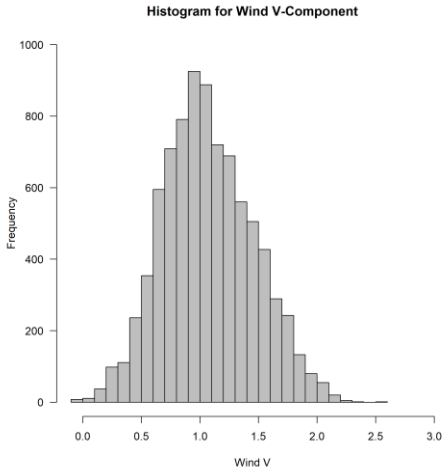

Histogram for Wind V-Component
Frequency
0
200
400
600
800
1000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Wind V

Table 6. Average $PM_{2.5}$ in $\mu g/m^3$ at each monitor station in the model fitting dataset.

Monitor	Average $PM_{2.5}$
Check_2	18.1
Check_7	19.8
Check_8	18.8
Check_9	19.9
Check_10	19.5
Check_11	16.8
ATE	38.3
CDM	15.2
CRB	28.0
HCH	30.9
PPD	32.8
SBJ	18.2
SJL	31.1
SMP	17.2
STA	29.0
VMT	24.4

Figure 21. Density plot of correlations between predicted and measured $PM_{2.5}$ values from the cross-validation of the LME model.

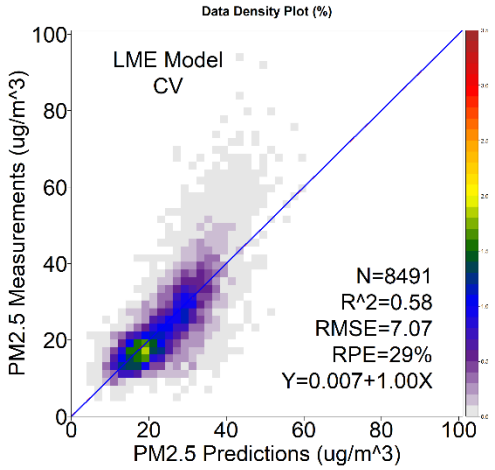


Table 7. All variables used in LME model along with beta values, standard error, degrees of freedom, t-values, and p-values.

Variables	Beta-Coefficients	Std.Error	DF	t-value	p-value
(Intercept)	24.84	0.87	8099	28.44	0.00
I(cent_AOT)	0.94	0.17	8099	5.65	0.00
I(cent_pm25)	2.08	0.16	8099	13.18	0.00
I(cent_u10)	-0.33	0.18	8099	-1.85	0.06
I(cent_v10)	-1.32	0.14	8099	-9.51	0.00
I(cent_temp)	-0.16	0.22	8099	-0.73	0.47
I(cent_NDVI)	-0.10	0.09	8099	-1.18	0.24
I(cent_int_rh)	-1.20	0.77	8099	-1.55	0.12
I(cent_blh)	2.55	0.20	8099	12.69	0.00
I(cent_sp)	-0.96	0.24	8099	-4.05	0.00
I(cent_al)	-1.51	0.17	8099	-8.81	0.00
I(cent_lcc)	-1.06	0.23	8099	-4.63	0.00
I(cent_ssrd)	-1.22	0.17	8099	-7.39	0.00
I(cent_dist3)	-0.69	0.10	8099	-6.85	0.00
I(cent_Elev)	3.99	0.21	8099	18.62	0.00
I(cent_pop)	-1.90	0.10	8099	-18.16	0.00
I(cent_perurb)	0.77	0.10	8099	7.30	0.00

Figure 22. Density plot of correlations between predicted and measured $PM_{2.5}$ values from the cross-validation of the Random Forest model.

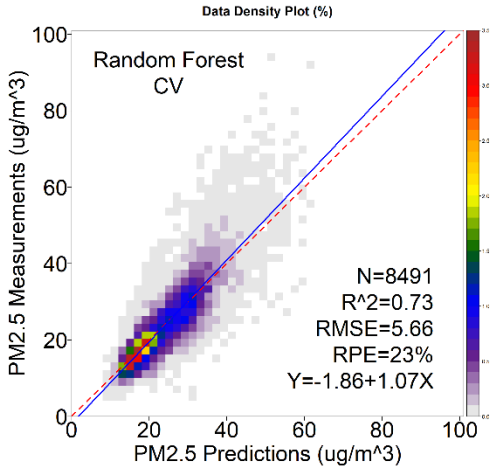


Table 8. All variables used in Random Forest model along with importance.

Parameters (Importance)		
WRF	**ECM**	**Misc**
PM25 (57)	RH (50)	AOT550 (68)
Temperature (80)	Surface Pressure (55)	NDVI (62)
	Wind_U (59)	% Urbanization (26)
	Wind_V (51)	Elevation (30)
	Albedo (78)	Cat3 Road Dist (25)
	Low Cloud Cover % (58)	Population Density (58)
	PBL (56)	
	Surf. Solar Radiation Downwards (72)	

Figure 23. Time-series comparing predicted $PM_{2.5}$ using the random forest model and ground PM2.5 concentrations in $\mu g/m^3$ for each Checkley monitor in 2012 starting from top left to right: Check_2, Check_7, Check_8, Check_9, Check_10, and Check_11.

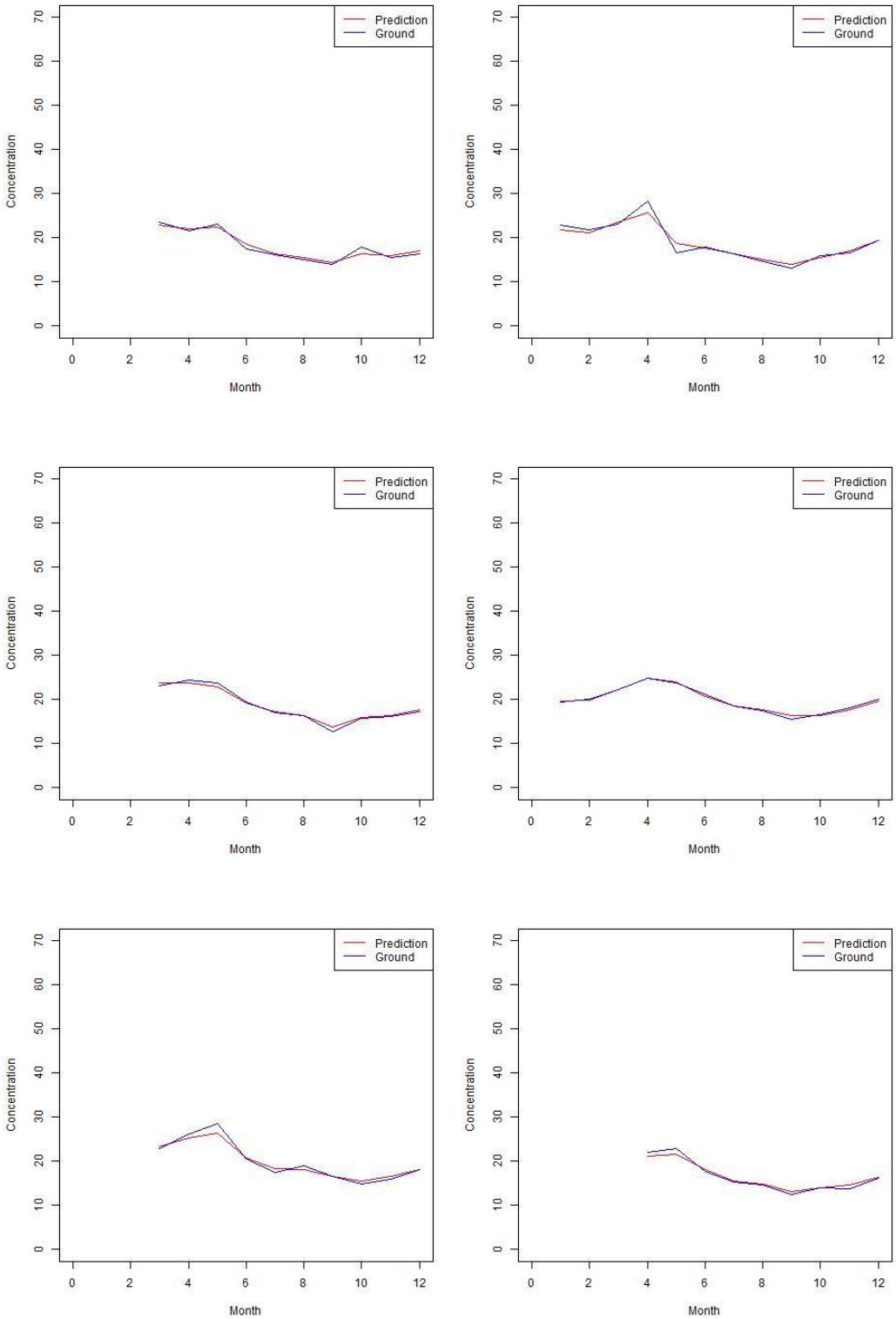
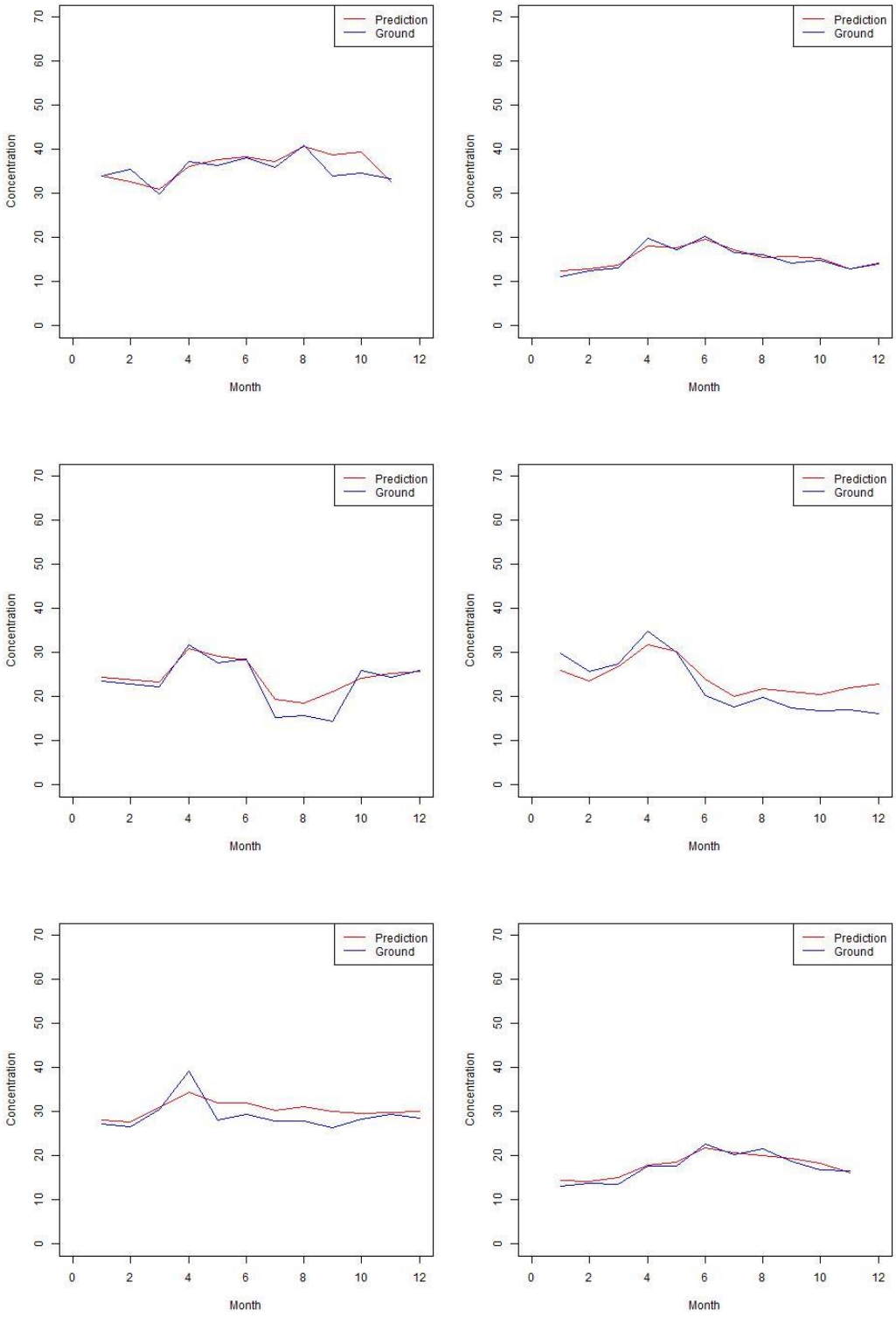


Figure 24. Time-series comparing predicted $PM_{2.5}$ using the random forest model and ground PM2.5 concentrations in $\mu g/m^3$ in 2015 for each monitor starting from top left to right: ATE, CDM, CRB, HCH, PPD, SBJ, SJL, SMP, STA, and VMT.



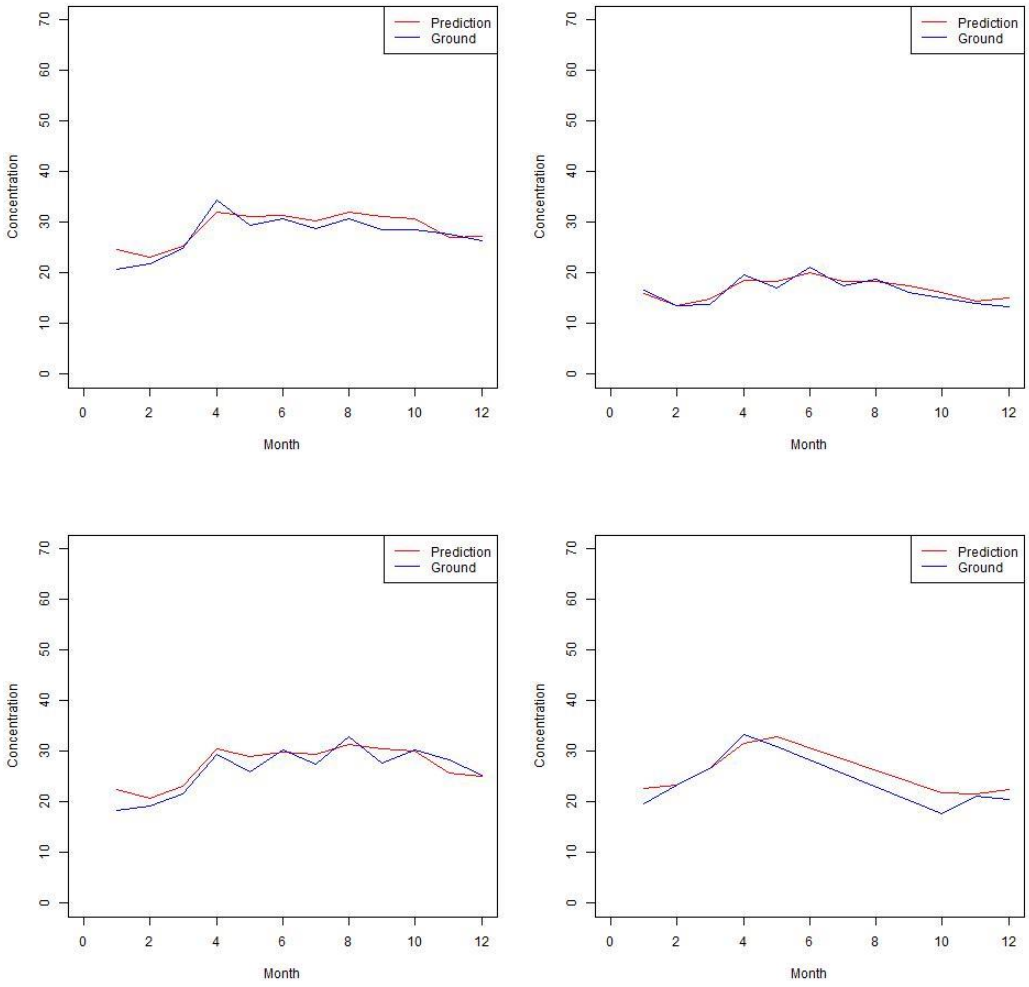


Figure 25. Maps comparing the mean concentration of each monitor station in the study domain on the left, to the mean estimated concentration in µg/m^{3}of each monitor from the cross-validation results on the right.

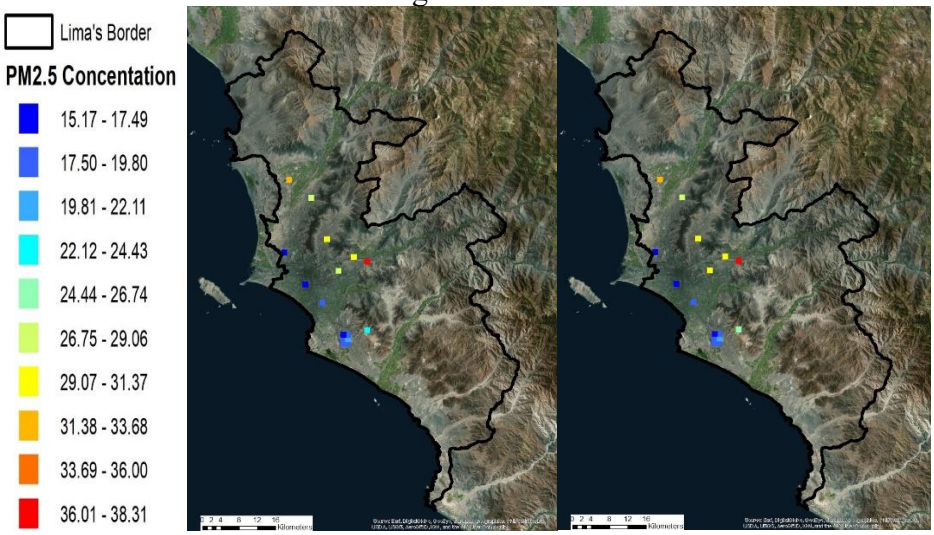


Figure 26. Annual prediction maps of $PM_{2.5}$ concentrations in $\mu g/m^3$ using the random forest model.

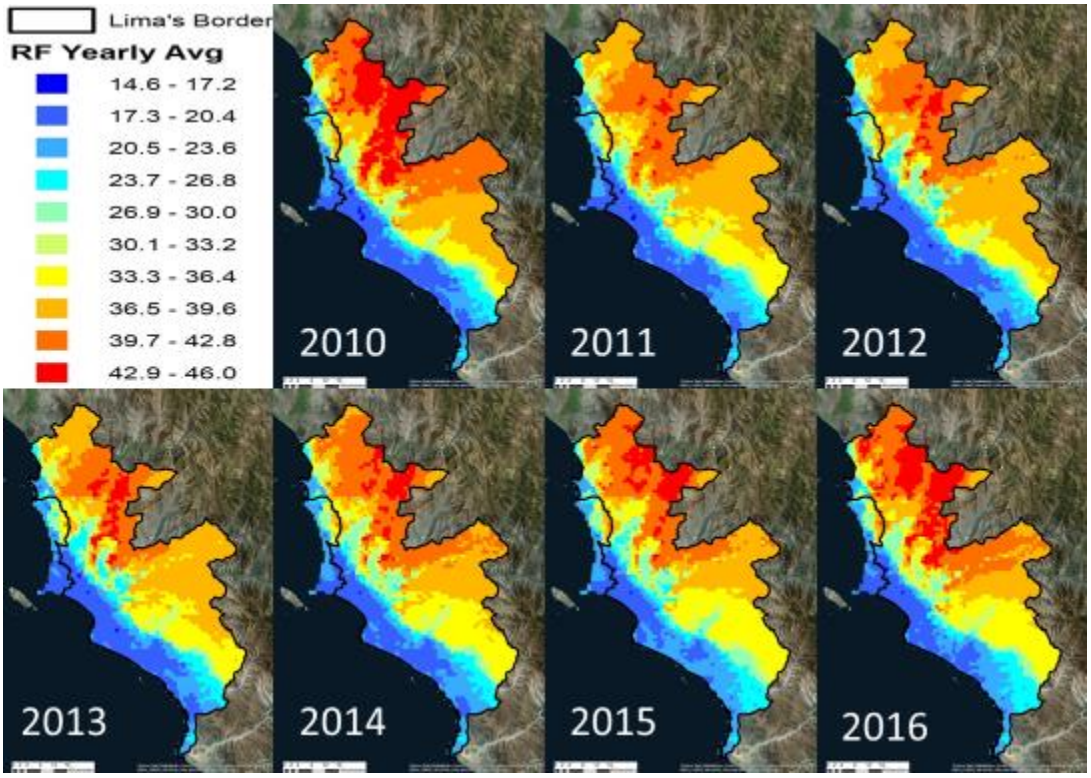
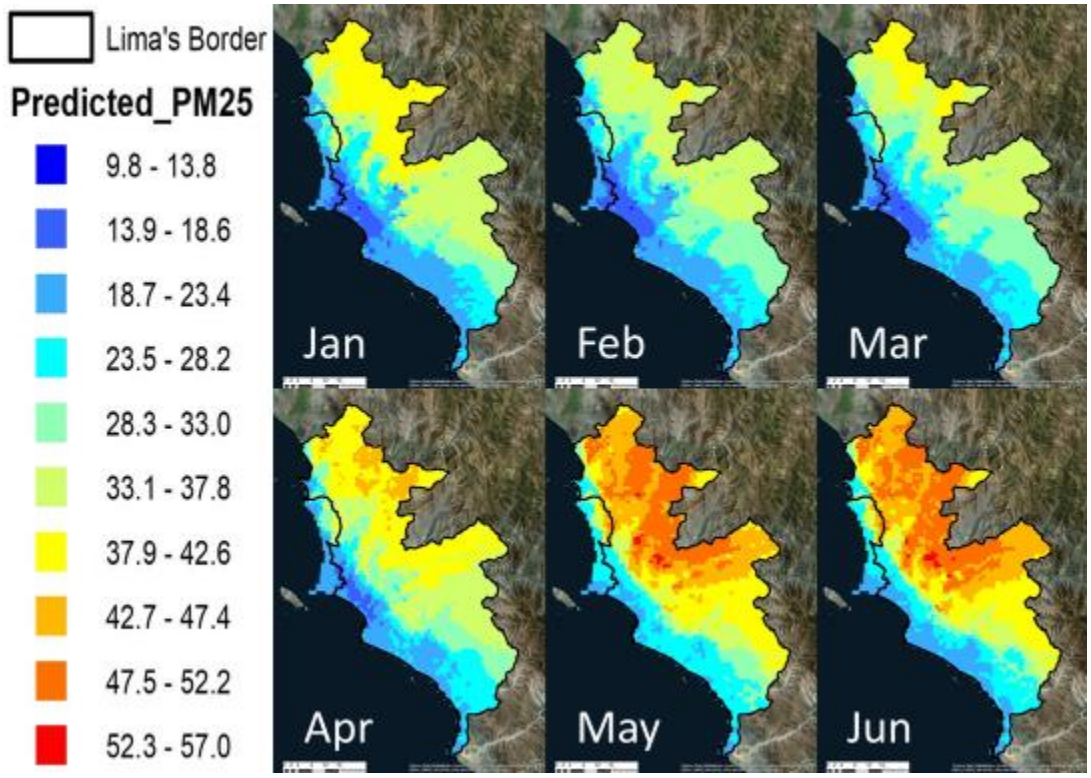
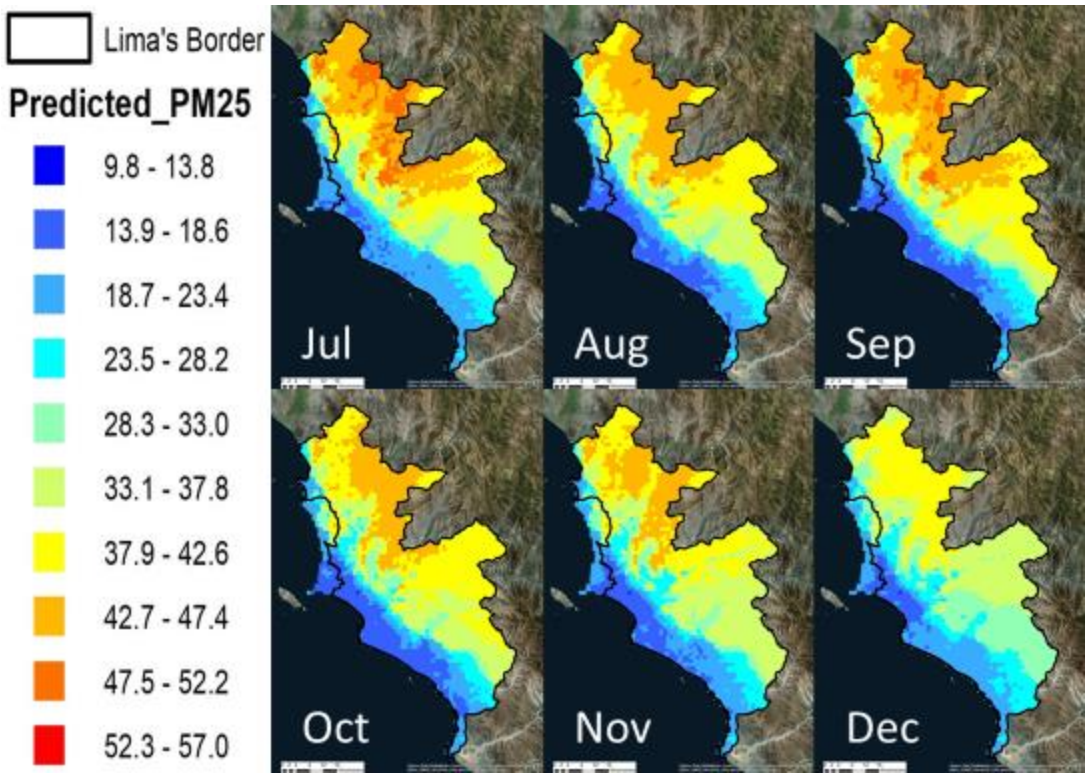


Figure 27. Monthly prediction maps of $PM_{2.5}$ concentrations in $\mu g/m^3$ for 2015 using the random forest model.




Lima's Border
Predicted_PM25
9.8 - 13.8
13.9 - 18.6
18.7 - 23.4
23.5 - 28.2
28.3 - 33.0
33.1 - 37.8
37.9 - 42.6
42.7 - 47.4
47.5 - 52.2
52.3 - 57.0
Jul
Aug
Sep
Oct
Nov
Dec