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:

_____ _____

Sophia Le Date

Utilizing the GOCI Satellite to Estimate Hourly $PM_{2.5}$ Concentrations in South Korea from 2015 – 2018

By

Sophia Le
Master of Science in Public Health

Environmental Health – Epidemiology

Yang Liu, Ph.D.
Committee Chair

Utilizing the GOCI Satellite to Estimate Hourly $PM_{2.5}$ Concentrations in South Korea from 2015 – 2018

By

Sophia Le

Bachelor of Science
Arizona State University
2017

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
2020

Abstract

Utilizing the GOCI Satellite to Estimate Hourly $PM_{2.5}$ Concentrations in South Korea from 2015 – 2018

By Sophia Le

Air quality in South Korea has been deteriorating in recent years, due to both domestic sources and long-range transport from China. Ground PM2.5 measurements are primarily located in urban environments, resulting in limited spatial and temporal coverage. To address this limitation, AOD retrieved from GOCI was utilized in multiple random forest machine learning models to estimate $PM_{2.5}$ concentrations. We developed 8 separate random forest machine learning models with time-varying meteorology and spatially fixed land information parameters to be included in the model. A 6 km modeling grid with 6,307 pixels was used to match the parameters to the GOCI retrievals. The average GOCI AOD coverage in the study domain was 48%. The 10-fold cross validation R^2 ranged from 0.47 – 0.54 with RMSEs' ranged between 9.74 – 13.34 μg/m^3. The regression slope between observed and predicted hourly concentrations ranged between 1.2 – 1.3. Prediction maps of hourly PM2.5 levels indicate a higher concentration on the western coast of South Korea compared to the eastern coast. We further analyzed an episode of long-range transport from China to South Korea during March 10th and 11th, 2015. Results from those findings revealed evidence of long-range transport given that the western coast has higher concentrations. Overall, our hourly prediction models allow us to understand the spatiotemporal variations of $PM_{2.5}$ concentrations in South Korea that could not have been done by ground stations alone.

Utilizing the GOCI Satellite to Estimate Hourly $PM_{2.5}$ Concentrations in South Korea from 2015 – 2018

By

Sophia Le

Bachelor of Science
Arizona State University
2017

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
2020

Acknowledgements

I would like to express my deepest appreciation to my advisor, Dr. Yang Liu for his support and guidance throughout the entirety of my thesis.

Secondly, I would also like to thank Dr. Jhoon Kim's team at the Atmospheric Sciences department at Yonsei University – especially Seoyoung Lee and Dr. Jhoon Kim for providing me the data for my analyses. I would also like to express my gratitude to the members of the Environmental Remote Sensing Group at Emory University for their help whenever I faced an obstacle.

Finally, I would like to thank my family and friends for their constant encouragement as I pursued my academic endeavors. I would like to express my appreciation to Justin Sukernek for his never-ending support and love as I worked on this project. I would also like to thank my parents, Jimmy Le and Ngan Tran, from the bottom of my heart for their hard work and sacrifice to ensure that I receive a higher education.

Table of Contents

1. Introduction 1
    1.1 $PM_{2.5}$ Background 1
    1.2 South Korea's Air Quality and Meteorology 1
    1.3 Ground Measurements of $PM_{2.5}$ in South Korea 3
    1.4 Air Pollution Studies in Korea Utilizing GOCI AOD 3
    1.5 Study Objectives 4
2. Data and Methods 6
    2.1 Study Domain 6
    2.2 Dataset Description 6
        2.2.1 GOCI Satellite YAER Version 2 Data 6
        2.2.2 Ground $PM_{2.5}$ Measurements 7
        2.2.3. Meteorology Data 7
        2.2.4. Ancillary Data 8
        2.2.5. Data Integration 9
    2.3 Model Development 9
        2.3.1 Random Forest Machine Learning Model 10
        2.3.2 *k*-fold Cross Validation 11
3. Results 13
    3.1.Descriptive Statistics 13
    3.2 Spatiotemporal AOD Coverage 14
    3.3 Model Performance and Prediction 16
    3.4 Long Range Transport 18
4. Discussion and Conclusion 19
5. References 22
6. Tables and Figures 27

1. Introduction

1.1 $PM_{2.5}$ Background

Exposure to ambient air pollution has been associated with a higher risk of morbidity and mortality in numerous epidemiological studies [7, 25, 27, 32, 38]. Particulate matter with an aerodynamic diameter of less than 2.5 μm ($PM_{2.5}$) has been determined to have adverse health impacts, including increasing the risk for cardiovascular and respiratory diseases, and premature death [21]. These fine particulates can reach the smallest airways and alveoli in the pulmonary system, eventually perpetuating the alveolar-capillary membrane and spread into the circulatory system [10]. Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease, published by WHO in 2016, estimated 92% of the world's population are exposed to $PM_{2.5}$ concentrations above the recommended annual mean of 10 $\mu g/m^3$ [34]. People living in low and middle income countries are most disproportionately affected by the burden of ambient air pollution, with Southeast Asia and the Western Pacific experiencing the worst outcome of premature mortality [2]. In 2012, due to ambient air pollution, Southeast Asia and the Western Pacific region, including South Korea, accounted for 3 million of the projected 4.2 million deaths [19, 34].

1.2 South Korea's Air Quality and Meteorology

Due to the impact of fine particulate matter on adverse health effects, it is imperative to closely examine countries worst affected by ambient air quality – one of these countries being South Korea. In the Environmental Performance Index (EPI) 2018 published by Yale University, South Korea was ranked 119^{th} out of 180 countries in overall ambient air quality but ranked 174^{th} out of 180 and 169^{th} out of 180 in $PM_{2.5}$

exposure and exceedance respectively [33]. The index indicates that South Korea has a higher average $PM_{2.5}$ concentration and that exposure among Korea's population exceeds the WHO's annual threshold.

Regional air quality in the Seoul Metropolitan Area (SMA), consisting of Seoul, Incheon, and Gyeong-Gi, are urbanized regions within South Korea that experienced an increase in the frequency of haze in recent years [18]. Air quality in this region is affected by both domestic and long range transport of anthropogenic emissions from China, Mongolia, and Russia [18]. In a highly urbanized region such as the SMA, domestic emission sources of $PM_{2.5}$ can be attributable to traffic, soil dust, industrial factories, or energy producing plants [18, 29]. In addition, long range wind transport of anthropogenic and natural sources such as dust from the Gobi desert and wildfire smoke from Mongolia can also contribute to the degradation of South Korea's air quality [17, 18]. The Korea – United States Air Quality (KORUS–AQ) campaign, a collaboration between the United States' National Aeronautics and Space Administration (NASA) and South Korea revealed that domestic sources contribute to 26% of $PM_{2.5}$ levels and long–range transport (LRT) from China contribute to nearly 68% of $PM_{2.5}$ concentrations in surface air pollution in South Korea during extreme episodes [8].

Further studies have shown pollutants emitted by cities located in Eastern China are transported to the SMA by northwestern winds [15, 20]. As one of the primary contributors to pollution episodes and haze, wind is one of the main meteorological components that affect surface air pollution [15]. Wind speed and direction controls the source-receptor relation, initiates local emissions, and contributes to long range transport of pollutants which affects South Korea's air quality [15]. A study conducted by Jung et

al. (2010) found that when long range transport of pollution from China influences the Korean Peninsula, average $PM_{2.5}$ concentrations in Seoul, South Korea exceeded 100 ug/m^3, which is approximately four times the annual average of 27 ug/m^3 [16]. Another study by Jung et al. (2018) observed firecracker emissions during the Lunar New Year over the Korean Peninsula. Their results found evidence suggesting LRT of pollutants from China to Korea under Asian continental outflows [16].

1.3 Ground Measurements of $PM_{2.5}$ in South Korea

Previous studies have utilized ground monitor stations in cities to measure fine particulate matter exposure, but because these stations are primarily located in urban centers, total population exposure to $PM_{2.5}$ cannot be accurately estimated [19]. Ground monitor stations are usually placed at fixed sites, such as urban regions, and can provide accurate data only for the restricted area [6]. Past findings on $PM_{2.5}$ concentrations in South Korea are confined to the SMA, but there is little literature examining the entire country [18, 21, 30]. In light of this, Lennartson et al. (2018) acknowledges that Seoul's $PM_{2.5}$ levels may not indicative of all of South Korea [23].

1.4 Air Pollution Studies in Korea Utilizing GOCI AOD

Due to the variance in PM levels and the number of different factors that lead to surface air pollution, several models have been developed to closely examine the cause and severity of current $PM_{2.5}$ concentrations. A study conducted by Park et al. (2019) adopted the Random Forest machine learning model to estimate surface coarse and fine particulate matter concentration using AOD retrieved from GOCI in South Korea between 2015 and 2016. Their model produced R^2 values of 0.78 and 0.73 for PM_{10} and $PM_{2.5}$ and root mean square errors (RMSEs) of 17.08 μg/m^3 and 8.25 μg/m^3 respectively

[28]. In Park's (2019) proposed model, they identified AOD retrieved from the GOCI satellite as the most significant parameter for estimating ground-level $PM_{2.5,}$ followed by wind speed, solar radiation, and dew-point temperature [28]. In models proposed by other studies, meteorological variables such as planetary boundary layer height, wind speed, relative humidity, land use information, as well as AOD have been used as parameters to predict PM levels [11, 14, 36]. An air quality forecasting model done by Ha et al. (2019) assimilated GOCI AOD during the KORUS – AQ campaign in a three-dimensional variational data assimilation method coupled with WRF-Chem [12]. Their results found that GOCI AOD is comparable to other satellite observations such as the Moderate Resolution Imaging Spectroradiometer (MODIS). Additionally, their results are consistent with other studies that suggest the assimilation of surface $PM_{2.5}$ concentrations alone underestimates PM2.5 concentrations, whereas the assimilation of GOCI AOD and surface $PM_{2.5}$ improves forecasts up to 24 hours [12]. Another study utilizing GOCI AOD retrievals in South Korea was done by Lennartson et al. (2018). Their study observed the diurnal variations in AOD and surface $PM_{2.5}$ measurements and their findings revealed a difference in AOD diurnal variations between coastal and inland regions as well as a large diurnal variation of the AOD-$PM_{2.5}$ relationship. Lee et al. (2019) examined GOCI AOD measurements and found an association between high AOD episodes in the Shandong province in China and South Korea, indicating a strong influence on LRT [22]. While recent studies have utilized GOCI AOD retrievals in various methods, this study is, to the best of our knowledge, the first to estimate hourly $PM_{2.5}$ concentrations using 8 separate random forest models. Overall, literature regarding estimating $PM_{2.5}$ concentrations outside of the SMA (i.e. the entirety of South Korea) is limited.

1.5 Study Objectives

In this study, we aim utilize GOCI AOD products to predict hourly $PM_{2.5}$ concentrations in South Korea between 2015 – 2018. In addition to GOCI AOD, other parameters such as meteorology and land use information were included in the model to reduce prediction error.

Our estimated predictions will help elucidate $PM_{2.5}$ levels across the entire country, not just in the metropolitan areas. This allows us to gain a greater understanding of $PM_{2.5}$ in South Korea that cannot be done by ground stations alone. A secondary objective of this study is to investigate the significance and impact that long range transport from China may have on the deteriorating PM concentrations in South Korea. By examining long range transport from China, this study can highlight specific factors that could influence heavy pollution episodes in South Korea.

2. Data and Methods

2.1 Study Domain

South Korea is located east of Beijing, situated in the middle of the Yellow Sea, the East China Sea, and the Sea of Japan. The distance between Seoul and Beijing is an estimated 957 km. The study domain ranges from 38°18 N, 33°09 S, 125°03 W, and 129°33 E. The map of South Korea was downloaded from [https://gadm.org/]. In total, there are 333 ground monitor stations in South Korea, located in 96 counties and cities throughout the country [1]. Under unique meteorological conditions, air quality in South Korea can deteriorate due to LRT from China because Korea is located downwind [22]. Figure 1 depicts South Korea's proximity to China's eastern coast.

2.2 Dataset Descriptions

2.2.1 GOCI Satellite YAER Version 2 Data

First launched in 2010, the Communication, Ocean, and Meteorology Satellite (COMS) was the first geostationary observation satellite in Northeast Asia [26]. The Geostationary Ocean Color Imager (GOCI) is one of the three instruments aboard the COMS satellite and acquires data in 8 spectral bands (6 visible, 2 NIR) with a spatial resolution of 500 meters [26]. GOCI has an observation density of eight retrievals per day (hourly observations between 09:30 to 16:30 Korean Standard Time) and monitors East Asia in sixteen slots at 2,500 meters by 2,500 meters spatial resolution [22].

The GOCI Yonsei aerosol retrieval (YAER) version one product was developed by Yonsei University to retrieve AOD at 550 nm and other supplemental aerosol optical features such as Fine Mode Fraction (FMF) and Angstrom Exponent [9]. Due to cloud coverage, approximately 40–70% of satellite data is missing on average in East Asia [37].

Despite the capabilities of the GOCI satellite, the version one algorithm prevents near-real-time processing because a database was required for the calculation for surface reflectance [9]. In addition, GOCI version one aerosol optical products resulted in a negative bias over land and positive bias over the ocean as a result of errors [9].

To adjust for the errors in GOCI version one, a version two algorithm was developed to process near real time data and modify surface reflectance and wind speed data [9]. A study conducted by Choi et al. (2018) analyzed aerosol optical products from the improved GOCI version two algorithm and found that AOD retrieval and values are akin to results from MODIS and Visible Infrared Imaging Radiometer Suite (VIIRS) AOD retrieval, surpassing the original GOCI version one [9]. In this study, GOCI AOD and FMF were used as parameters in the hourly $PM_{2.5}$ estimation model. Our study period ranges from January 2015 to December 2018.

2.2.2 $PM_{2.5}$ Ground Measurements

Hourly $PM_{2.5}$ concentrations in 2015–2018 were measured at approximately 325 ground monitoring stations across South Korea (Fig. 1). The $PM_{2.5}$ data was downloaded from Air Korea [https://www.airkorea.or.kr/], Korea's national real time air pollution data that provides information about air quality levels to the public [1].

2.2.3 Meteorology Data

Meteorological data was retrieved from the GEOS-FP Atmospheric Data Assimilation System at a 0.25° latitude × 0.3125° longitude spatial resolution [ftp://rain.ucis.dal.ca/ctm/GEOS_0.25x0.3125_CH.d/GEOS_FP/]. GEOS-FP was a collaborative effort by NOAA's National Centers for Environmental Protection and NASA's Global Modeling and Assimilation Office. Meteorology parameters such as

evaporation, surface incident shortwave flux, and 10 meter and pressure level (500 hPa) wind vector (U and V component) were included in the model. The temporal resolutions of the GEOS-FP were collected every 3 hours, and linear interpolation was used to estimate hourly data.

2.2.4 Ancillary Data

Land cover data for the study domain was downloaded from the Climate Change Initiative (CCI) of the European Space Agency [https://www.esa-landcover-cci.org/?q=node/164]. The 1992 – 2015 annual time series global land cover has a spatial resolution at 300 meters. The 2015 CCI land cover product includes 22 parameters such as grassland, urban areas, water bodies, and cropland at a 300 – meter spatial resolution. The Advanced Spaceborne Thermal Emission and Radiometer Global Digital Elevation Model (ASTER GDEM) version 3 elevation data for the South Korea domain was downloaded from NASA's EARTHDATA [https://search.earthdata.nasa.gov/search/]. ASTER GDEM is a satellite image produced from a collaboration between Japan's Ministry of Economy, Trade, and Industry and the United States National Aeronautics and Space Administration (NASA) [3]. ASTER GDEM's version 3 dataset was released on August 5th, 2019; the newest version maintains the same GeoTIFF format, grid structure, and 1° tiles as previous versions [3]. LandScan data, the Oak Ridge National Laboratory's global population distribution data, for the years 2015 – 2018 were downloaded from the East View Geospatial LandScan Global Archive [http://wms.cartographic.com.proxy.library.emory.edu/landscan/portal.aspx]. The LandScan population data represents the ambient population distribution averaged over

24 hours [4]. Global road network data was downloaded as a shapefile from the open source website, OpenStreetMap [https://www.openstreetmap.org/#map=4/38.01/-95.84].

2.2.5 Data Integration

A 6 km modeling grid (6,307 pixels) was produced from the centroid of the GOCI covering the study region. A 50 km buffer was implemented around the study domain to ensure the accuracy of the GOCI AOD. The inverse distance weighting (IDW) method was used to spatially interpolate the meteorology and land use information to match the 6 km grid resolution. Land use information such as the road and land cover variables were reclassified in ArcGIS Pro using the unsupervised reclassification technique to condense the categories for the respective datasets. The road network was reclassified into 2 categories: primary and secondary roads. The land cover parameters were reclassified from 22 into 6 classes: farmland, forest, grass, water, urban, and other. Distance from road type and percentage of each land cover group was calculated within each grid pixel.

2.3 Model Development

A preliminary analysis was conducted to examine the correlation between the 8 AOD hours. The correlation matrix revealed a high correlation between all 8 AOD hours from 2015 – 2018. Each hour was more highly correlated with the hours closest to it as depicted in Figure 2a –d. Over the four years, hour 0 and hour 7 are the least correlated among all 8 hours.

To avoid the high correlation between AOD hours, 8 separate hourly random forest models were produced. All the models utilized the same predictor variables at their associated times to estimate hourly $PM_{2.5}$ concentrations. There are 20 total predictors,

such as relative humidity, wind speed and direction (10 m and 500 mb), and GOCI fine mode fraction at 550 nm.

2.3.1 Random Forest Machine Learning Model

A random forest model was utilized to fit predictors to the ground $PM_{2.5}$ observations in South Korea. Random forest is an "ensemble learning" method that expands upon its predecessors by adding random tree construction to a previous method known as bagging [24]. Random forest constructs numerous classification or regression trees, varying the bootstrap data sample and predictors used in each tree. Along with the predictors, random forest requires two parameters to determine the branching: n_{tree} and m_{try}. The number of total trees is specified by n_{tree}, with each tree including a subgroup of predictors; m_{try} defines the number of predictors included. The random forest algorithm constructs classification or regression decision trees dependent on the number of bootstrap samples selected in the n_{tree} parameter [24]. In each of the random forest trees, each node is split using the best subset of predictors that are randomly selected at the node in the m_{try} parameter. Additionally, the model calculates prediction error in each bootstrap iteration by aggregating out-of-bag (OOB) predictions—predictions of data outside of the bootstrap. In turn, a variable importance measure is derived from the observed increases in prediction error associated with specific variables. The model's final prediction is calculated by aggregating every trees' prediction [24]. The models in this study set n_{tree} at 500 and m_{try} at 10. Two models were created: a full model with all the predictors and a restricted model that removed some predictors with low variable importance. The restricted model was selected for use in estimating hourly $PM_{2.5}$ concentrations.

We narrowed down the full model from 32 parameters in the dataset to 20 total predictors. GEOS-FP meteorology fields were reduced to 13 variables including surface shortwave incidence flux (SWGDN), total precipitation (PRECTOT), surface pressure (PS), and evaporation (EVAP). The selected parameters for land use information were farmland and water. Elevation, road network data, and population density predictors were kept in the model. These variables along with hourly GOCI retrievals (AOD and fine mode fraction) and surface $PM_{2.5}$ observations were inputted into 8 random forest models at their associated time.

2.3.2 k-fold Cross Validation

A *k*-fold cross validation is often utilized for evaluating the performance of regression or classification algorithms [35]. The *k*-fold cross validation is typically used to estimate the accuracy of a model's predictions, since a model will usually overfit its training data—resulting in predictions that are too high. The procedure randomly divides a dataset into a specified number (*k*) of folds of approximately equal size. Each fold is then used to test the model induced from the other *k* -1 folds in the regression algorithm [35]. The performance of the model is evaluated by the average of the *k* accuracies resulting from the *k*-fold cross validation.

A 10-fold cross validation was performed on each of the random forest models to validate the predictions that were produced from the models. The prediction dataset for each model was split into 10 randomly divided folds, with each fold containing 10% of the data. Nine out of the 10 folds acted as the training dataset, with the 10^{th} fold acting as the testing dataset to produce predictions. This *k*-fold cross validation is repeated 10

times, with each cross validation partitioning the prediction data at different intervals to ensure that the folds are not repeated.

3. Results

3.1 Descriptive Statistics

The full dataset for each model contains between 20,000 to 34,000 observations. Table 1 shows the descriptive statistics of some of the time-varying and spatially fixed parameters used in the random forest models. The mean $PM_{2.5}$ concentration in South Korea between 2015 – 2018 is 27 $\mu g/m^3$, with a low of 0 $\mu g/m^3$ and the highest level over the 4-year study period was 185 $\mu g/m^3$. Most of the ground $PM_{2.5}$ measurements had lower concentrations as evidenced by the mean of 27 $\mu g/m^3$. However, throughout the prediction data there were occasional spikes over 150 $\mu g/m^3$. The mean AOD value was 0.27, ranging from a minimum of -0.05 to a maximum of 3.59. No negative AOD values were removed from the prediction dataset because of the predictor's importance in the model. The highest elevation in the study domain is approximately 1304.77 m and the lowest point is at sea level (0 m). Wind speed 10 m and 500 m above the ground ranged between 0.08 – 51.42 m/s, and a mean of 4.83 m/s and 0.04 – 57.00 m/s, and a mean of 12.98 m/s respectively. The average planetary boundary layer height, surface pressure, and 2 m air temperature during the study period was 888.74 m, 1002.07 Pa, and 290.49 K respectively. The predictors' range varied from 59.15 – 4305.28 m, 903.44 – 1036.40, and 238.19 – 311.25 K.

Figure 3 depicts the average monthly ground $PM_{2.5}$ measurements during the study period. The time series illustrates the fluctuating trend of ground measurements over the course of 4 years. At the beginning of the year, average $PM_{2.5}$ concentrations are the highest, but appear to decrease in the middle of the year before increasing again. This trend is better illustrated in Figure 4, which depicts seasonal ground measurement averages. Over the study period, $PM_{2.5}$ concentrations tend to be highest in the spring and

winter, and lowest in summer. As seen in Figures 3 and 4, the increases start in the fall, jumping from 20.50 – 22.41 μg/m^3 to 25.21 – 30.61 μg/m^3 in the winter from 2015 – 2018. The average monthly $PM_{2.5}$ measurements' highest peak is 36.14 μg/m^3, which falls in March 2017, and the lowest peak falls in September 2018 at 12.10 μg/m^3.

3.2 Spatiotemporal AOD Coverage

Figure 5 depicts the average monthly AOD coverage in the study domain between 2015 – 2018. On average, monthly AOD coverage ranges between 30% – 80% in 2015, 25% – 70% in 2016, 30% – 77% in 2017, and 30% to 74% in 2018. The mean AOD coverage across all 4 years is 48%, indicating that 52% of the full dataset is missing AOD values. While the mean AOD coverage seems low, results suggest that the GOCI satellite outperforms more established satellites such as MODIS, with an annual AOD coverage around 10% [13]. Overall, the trend in the temporal AOD coverage indicates that there is higher coverage in the spring (March – May) ranging from 63% – 73% across the four years. Furthermore, the trend in coverage appear to suggest a similar pattern year by year. In the first half of the year, generally AOD coverage is higher and then begins to decrease and has a small spike during the fall before decreasing again. The lowest AOD coverage generally occurs in December, and the January afterwards. During the study period, AOD coverage in these months range between 30 – 36%. In the summer (June – August, AOD coverage, ranging from 37% – 51%. In October 2015 and 2018, AOD coverage spikes at 80% and 59% respectively, which is unusual compared to the overall trend in the time series. In August 2016 and July 2018, AOD coverage has a small spike around 60%, a unique trend considering that the other years in the plot illustrate AOD coverage decreasing in the summer. The lowest AOD coverage occurs in November 2015 with a

23% temporal coverage and the highest in October 2015. Out of all the years in the study period, 2015 appears to have the most fluctuation in AOD coverage as evidenced in Figure 5. Unlike the other years, 2015 has more unusual minimum and maximum AOD coverage trends.

Figures 5a-d illustrate the hourly AOD spatial coverage in South Korea between 2015 – 2018. The spatial coverage in the figures represent AOD coverage per pixel in the modeling grid. The spatial AOD coverage patterns suggests that the GOCI satellite has more coverage in the pixels in the northern region of the country compared to the southern half. In addition, hour 0—the beginning of the observation period—appears to have a higher spatial AOD coverage compared to the rest of the hours during the study period. With the exception of hour 5 in 2016, hours 5 and 6 from 2015 – 2018 display high AOD pixel coverage. This is evident by the darker shading on the hourly maps in the figures. The hour 0 maps in 2015 and 2017 have the highest spatial AOD coverage in the pixels; in 2015, the spatial coverage is concentrated on the eastern coast of the country whereas in 2017, AOD coverage is concentrated on the northwestern region. From 2015 – 2018, hours 5 (except 2016) and 6's spatial coverage is spread out throughout the country rather than concentrated on a specific region like the hour 0 coverage. In the hour 5 coverage in 2016, there is less spatial coverage compared with the other years. The spatial coverage in the pixels range between 0% – 40%, with 2018 having the smallest coverage range of 0% – 25%. An unusual pattern is seen in the spatial AOD coverage that occurs in hour 7 in 2015 and 2016, as illustrated in bottom right map on the figures. Unlike the other years, the map of hour 7 in 2015 shows high coverage on the east and west coasts of Korea, but low coverage in the middle. In 2016, spatial AOD

coverage in hour 7 shows almost the opposite of hour 7 in 2015; there is coverage in the middle of the country, but less coverage on the coasts.

3.3 Model Performance and Predictions

Eight random forest models were used to fit predictors of $PM_{2.5}$ to the ground observation $PM_{2.5}$ data. Given that GOCI has 8 hourly retrievals a day, 8 random forest models were produced using the AOD hour as an anchor. Time varying parameters were temporally matched to the GOCI AOD hour and both parameters with spatially matched to the grid cells in the modeling grid. The models were set with a m_{try} of 10 and n_{tree} at 500. The restricted model contained predictors such as GOCI AOD, road type, land use, wind speed and direction 10 m and 500 m above ground. The "out of bag" R^2 from the random forest models ranged from 0.48 – 0.55 with RMSEs' ranged between 9.64 – 13.23 $\mu g/m^3$. Figure 6 shows the density plots of the correlations between the predicted and observed PM2.5 measurements across the 8 individual models. The cross-validated R^2 ranged from 0.47 – 0.54 with RMSEs' ranged between 9.74 – 13.34 $\mu g/m^3$. Given that the cross validated R^2 is comparable to the out of bag R^2, this indicates that our model accurately fits the data. The regression slope in the cross validated plots ranged between 1.2 – 1.3. Due to the variance in the models, the predicted $PM_{2.5}$ values tended to be lower than the observed $PM_{2.5}$ measurements. Table 2 demonstrates the top 5 important predictors in the 8 random forest models. The threshold for importance ranking in a random forest model is determined by a permutation test. The higher the importance or the mean squared error (MSE) is in the model – this indicates the higher predictive accuracy of the parameter in the model [24]. Across the 8 models, GOCI AOD, relative humidity, wind speed and direction 500 m above ground are highly significant predictors

of $PM_{2.5}$ measurements. Other highly significant parameters are GOCI fine mode fraction and meteorology variables such as surface shortwave incidence flux, total precipitation, and total cloud fraction. The least important variables in the model are road networks, population density, land use information. Given that ground monitor stations are located in urban regions (e.g. the Seoul metropolitan area), it is expected that spatial invariant parameters would not be important predictors in the model. Variables such as road networks and population density do not fluctuate over time and the insignificance of these predictors is evident in the variable importance ranking.

Prediction maps from the random forest models are shown in Figures 7a-d. Hourly maps over 2015 – 2018 illustrate the spatiotemporal variations in $PM_{2.5}$ concentration between 9:00 – 16:00 KST time in South Korea. Mean hourly predicted $PM_{2.5}$ concentrations ranged from 22.8 – 30.5 $\mu g/m^3$. However, the overall $PM_{2.5}$ predictions from the random forest models underestimates values at some hours. The prediction maps suggest that the western coast and the northern region of South Korea experience higher $PM_{2.5}$ concentrations compared to the southern region of the country. As depicted in Figures 7a-b, the predicted hourly maps show that there are higher $PM_{2.5}$ concentrations near the western coast of the domain in 2015 and 2016. This trend is most prominent in Figure 7b (2016); across 8 hours, the entire western coast displays higher concentrations, with hours 1 and 2 being the darkest shade. In Figures 7c-d, the predicted maps illustrate high $PM_{2.5}$ concentrations in the northern region of Korea. The hourly maps suggest that as time passes, the concentrations in the northern region moves towards the eastern coast. Hourly predictions from the model also seem to suggest that eastern coast tends to have lower $PM_{2.5}$.

3.4 Long-Range Transport

Potential long-range transport from China to South Korea was assessed from the $PM_{2.5}$ predictions produced from the random forest model as depicted in Figure 8. Figure 8 illustrates hourly $PM_{2.5}$ predictions during a long-range transport episode on March 10th – 11th, 2015. An episode is classified as long-range transport if the wind direction for winds 500 m above the ground are blowing from west to east (China to Korea) and this event occurs between 2 – 3 consecutive days. The average $PM_{2.5}$ concentrations on March 10th and 11th are 22.94 μg/m^3 and 36.47 μg/m^3 respectively. The highest predicted concentration of 46.95 μg/m^3 occurred in hour 6 (15:00) on March 11th. Across the study hours on March 10th, $PM_{2.5}$ concentrations were comparatively lower than the concentrations on March 11th. Hourly average $PM_{2.5}$ levels on the 10th ranged from 19.73 μg/m^3 – 24.56 μg/m^3. Hourly mean $PM_{2.5}$ concentrations on the 11th ranged from 30.45 μg/m^3 – 46.95 μg/m^3.

The maps in Figure 8 depict higher concentrations near South Korea's western coast, indicating potential long-range transport from China. The eastern coast of the country has lower concentrations in comparison. On the hour 6 (15:00) map on March 11th, the shading illustrates extremely high concentrations over 70 μg/m^3 in the Southwestern portion of the country. This unusual pattern is not seen in the other hourly maps on March 10th or 11th. Overall, the hourly maps suggest that long-range transport from China is partly responsible in exacerbating South Korea's deteriorating air quality.

4. Discussion and Conclusion

To the best of our knowledge, this study was the first of its kind to use the geostationary satellite to estimate hourly $PM_{2.5}$ concentrations in South Korea. The study is also novel in that we examined the impact of long-range transport on South Korea's air quality. Our estimated hourly predictions will help shed a light on $PM_{2.5}$ levels across the entire country, not just in the metropolitan areas where ground stations are primarily located. As such, this study allowed us to gain a greater understanding of the spatiotemporal patterns of $PM_{2.5}$ in South Korea. In addition, by investigating the impact of long range transport on Korea's regional air quality, we can study the factors that shape heavy air pollution episodes and implement policies to reduce ambient pollution.

We developed 8 separate random forest models to predict hourly $PM_{2.5}$, which allowed us to better estimate and observe the spatiotemporal variations of $PM_{2.5}$ in the study domain. The out of bag R^2 from the 8 models ranged from 0.48 – 0.55 with RMSEs' ranged between 9.64 – 13.23 $\mu g/m^3$. This indicates in the predictors in the model explain about half of the variance in the dataset. While the R^2 is not incredibly high, it does not necessarily mean that the mean predictions are inaccurate. It just means that we have a large spread in the dataset. The cross-validated R^2 ranged from 0.47 – 0.54 with RMSEs' ranged between 9.74 – 13.34 $\mu g/m^3$. The comparable R^2s between the models and the cross-validations indicate that overfitting of the prediction data is not likely. That said, there is still a possibility that the model is not accurately estimating $PM_{2.5}$ concentrations.

Looking at variable importance, GOCI AOD was consistently the most important predictor in all of the random forest models. The least important predictors in the model

were spatially fixed parameters such as road networks, elevation, and land use. It was expected that these predictors were not highly ranked due to the fixed position of the ground monitor stations. Unlike meteorological parameters, spatially fixed variables do not have any temporal variation which is the reason they are not highly significant in a model that was developed to estimate hourly predictions. Wind speed and wind direction at 500 m above the ground were also highly significant predictors in the models. These predictors are extremely important when observing long-range transport from China to South Korea.

Figure 8 depicts the spatiotemporal variation of long-range transport of pollutants from China to South Korea. As evidenced in the maps, the western coast of Korea has much higher $PM_{2.5}$ concentrations compared to the eastern coast during March 10^{th} and March 11^{th}, 2015. This evidence of long-range transport is corroborated with the findings from Lee et al. (2019). Study findings from Bae found that Chinese contribution to South Korea's deteriorating air pollution is likely caused by the inflow of air from the northwesterly winds in the spring and winter [5] . On the other hand, we see less contribution in the summer and fall due to the weakening winds.

A notable finding in the results was the seasonal trend in ground $PM_{2.5}$ measurements throughout 2015 – 2018. PM levels tend to be lower in the summer and fall and increase in the spring and winter. A possible reason as to why $PM_{2.5}$ concentrations in South Korea are higher during this period is because the primary wind direction makes it favorable to transport pollutants from China to Korea [5]. Other reasons as to why $PM_{2.5}$ concentrations are higher in the winter and spring could be due to the increase usage of biomass burning during the colder seasons. As the winds weaken

in the fall and summer, domestic sources make a greater contribution to the regional air quality in South Korea [5].

One of the limitations in this study was the possibility of inaccurate predictions of $PM_{2.5}$ concentrations. The low R_2 produced from the random forest models limits our ability to make more general conclusions. This may be due to the coarser resolution of the modeling grid. Previous studies have found that a 6 km resolution is unable to reflect finer spatial contrasts in a grid cell [31]. Another limitation in this study was that we did not gap-fill GOCI AOD data, which resulted in an overall missingness of 52% in the Korean Peninsula. In addition, ground monitor stations are primarily located in urban settings, which limits our ability to assess the models' performance in rural areas.

Future directions of this research could involve interpolating missing $PM_{2.5}$ measurements using a statistical model in conjunction with a random forest model to obtain a larger prediction dataset for more accurate estimates. Since ground stations are fixed in cities, it would be prudent to focus on assessing the model's performance in rural regions to gain a better understanding of the spatiotemporal variations. Another direction to take on this study relates to long-range transport, specifically observing events in China that have heavy pollution days (e.g. Chinese New Year) and the impact of LRT on Korea's air quality.

5. References

1. *Air Korea*. Available from: https://www.airkorea.or.kr/web.
2. *Air Pollution*. Available from: http://www.who.int/airpollution/en/.
3. *ASTER: Advanced Spaceborne Thermal Emission and Reflectance Radiometer*. Available from: https://asterweb.jpl.nasa.gov/gdem.asp.
4. *LandScan Geographic Information Science & Technology*. Available from: https://landscan.ornl.gov/.
5. Bae, M., et al., *A Multiscale Tiered Approach to Quantify Contributions: A Case Study of PM2.5 in South Korea During 2010–2017.* Atmosphere, 2020. **11**(2).
6. Borghi, F., et al., *Miniaturized Monitors for Assessment of Exposure to Air Pollutants: A Review.* Int J Environ Res Public Health, 2017. **14**(8).
7. Chen, C.C., P.S. Chen, and C.Y. Yang, *Relationship between fine particulate air pollution exposure and human adult life expectancy in Taiwan.* Journal of Toxicology and Environmental Health-Part a-Current Issues, 2019. **82**(14): p. 826-832.
8. Choi, J., et al., *Impacts of local vs. trans-boundary emissions from different sectors on PM2.5 exposure in South Korea during the KORUS-AQ campaign.* Atmospheric Environment, 2019. **203**: p. 196-205.
9. Choi, M., et al., *GOCI Yonsei aerosol retrieval version 2 products: an improved algorithm and error analysis with uncertainty estimation from 5-year validation over East Asia.* Atmospheric Measurement Techniques, 2018. **11**(1): p. 385-408.

10. Du, Y., et al., *Air particulate matter and cardiovascular disease: the epidemiological, biomedical and clinical evidence.* J Thorac Dis, 2016. **8**(1): p. E8-E19.

11. Guo, J.P., et al., *Impact of diurnal variability and meteorological factors on the PM2.5 - AOD relationship: Implications for PM2.5 remote sensing.* Environmental Pollution, 2017. **221**: p. 94-104.

12. Ha, S., et al. *Improving air quality forecasting with the assimilation of GOCI AOD retrievals during the KORUS-AQ period*. 2019.

13. Hua, Z., et al., *A Full-Coverage Daily Average PM2.5 Retrieval Method with Two-Stage IVW Fused MODIS C6 AOD and Two-Stage GAM Model.* Remote Sensing, 2019. **11**(13).

14. Huang, K., et al., *Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain.* Environ Pollut, 2018. **242**(Pt A): p. 675-683.

15. Jeong, U., et al., *Assessing the effect of long-range pollutant transportation on air quality in Seoul using the conditional potential source contribution function method.* Atmospheric Environment, 2017. **150**: p. 33-44.

16. Jung, J., et al., *Chemical characterization of the long-range transport of firework/firecracker emissions over the Korean Peninsula: A novel indicator of Asian continental outflows.* Atmospheric Environment, 2018. **178**: p. 223-230.

17. Kim, H., H. Kim, and J.T. Lee, *Effects of ambient air particles on mortality in Seoul: Have the effects changed over time?* Environ Res, 2015. **140**: p. 684-90.

18. Kim, H.C., et al., *Recent increase of surface particulate matter concentrations in the Seoul Metropolitan Area, Korea.* Sci Rep, 2017. **7**(1): p. 4710.

19. Kim, J.H., et al., *Premature Deaths Attributable to Long-term Exposure to Ambient Fine Particulate Matter in the Republic of Korea.* J Korean Med Sci, 2018. **33**(37): p. e251.

20. Koo, Y.-S., et al., *Performance evaluation of the updated air quality forecasting system for Seoul predicting PM10.* Atmospheric Environment, 2012. **58**: p. 56-69.

21. Lee, H.M., et al., *PM2.5 source attribution for Seoul in May from 2009 to 2013 using GEOS-Chem and its adjoint model.* Environ Pollut, 2017. **221**: p. 377-384.

22. Lee, S., et al., *Analysis of long-range transboundary transport (LRTT) effect on Korean aerosol pollution during the KORUS-AQ campaign.* Atmospheric Environment, 2019. **204**: p. 53-67.

23. Lennartson, E.M., et al., *Diurnal variation of aerosol optical depth and PM2.5 in South Korea: a synthesis from AERONET, satellite (GOCI), KORUS-AQ observation, and the WRF-Chem model.* Atmospheric Chemistry and Physics, 2018. **18**(20): p. 15125-15144.

24. Liaw, A. and M. Wiener *Classification and Regression by randomForest*. 2002. **2/3**.

25. Liu, C., et al., *Ambient Particulate Air Pollution and Daily Mortality in 652 Cities.* New England Journal of Medicine, 2019. **381**(8): p. 705-715.

26. NASA Goddard Space Flight Center, O.E.L., Ocean Biology Processing Group. *Geostationary Ocean Color Imager (GOCI).* [cited 2019 12/18]; Available from: https://oceancolor.gsfc.nasa.gov/data/goci/.

27. Park, J.-O., et al., *The Effects of Air Pollution on Mortality in South Korea.* Procedia Environmental Sciences, 2015. **26**: p. 62-65.

28. Park, S., et al., *Estimation of ground-level particulate matter concentrations through the synergistic use of satellite observations and process-based models over South Korea.* Atmospheric Chemistry and Physics, 2019. **19**(2): p. 1097-1113.

29. Ryou, H.G., J. Heo, and S.Y. Kim, *Source apportionment of PM10 and PM2.5 air pollution, and possible impacts of study characteristics in South Korea.* Environ Pollut, 2018. **240**: p. 963-972.

30. Seo, S., et al., *Estimation of PM10 concentrations over Seoul using multiple empirical models with AERONET and MODIS data collected during the DRAGON-Asia campaign.* Atmospheric Chemistry and Physics, 2015. **15**(1): p. 319-334.

31. She, Q., et al., *Satellite-based estimation of hourly PM2.5 levels during heavy winter pollution episodes in the Yangtze River Delta, China.* Chemosphere, 2020. **239**: p. 124678.

32. Wang, J., et al., *Associations between ambient air pollution and mortality from all causes, pneumonia, and congenital heart diseases among children aged under 5 years in Beijing, China: A population-based time series study.* Environmental Research, 2019. **176**.

33. Wendling, Z.A., et al., *2018 Environmental Performance Index*. 2018, Yale Center for Environmental Law & Policy: New Haven, CT.

34. WHO, *Ambient air pollution: a global assessment of exposure and burden of disease*. 2016, World Health Organization.

35. Wong, T.-T., *Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation.* Pattern Recognition, 2015. **48**(9): p. 2839-2846.

36. Xiao, Q., et al., *An Ensemble Machine-Learning Model To Predict Historical PM2.5 Concentrations in China from Satellite Data.* Environ Sci Technol, 2018. **52**(22): p. 13260-13269.

37. Xiao, Q., et al., *Evaluation of VIIRS, GOCI, and MODIS Collection 6 AOD retrievals against ground sunphotometer observations over East Asia.* Atmospheric Chemistry and Physics, 2016. **16**(3): p. 1255-1269.

38. Xing, Y.-F., et al., *The impact of PM2.5 on the human respiratory system.* Journal of thoracic disease, 2016. **8**(1): p. E69-E74.

6. Tables and Figures

Figure 1. Study domain showing the proximity of South Korea to China and includes ground monitor stations. The two monitor stations that appear located in the middle of water are on small islands.

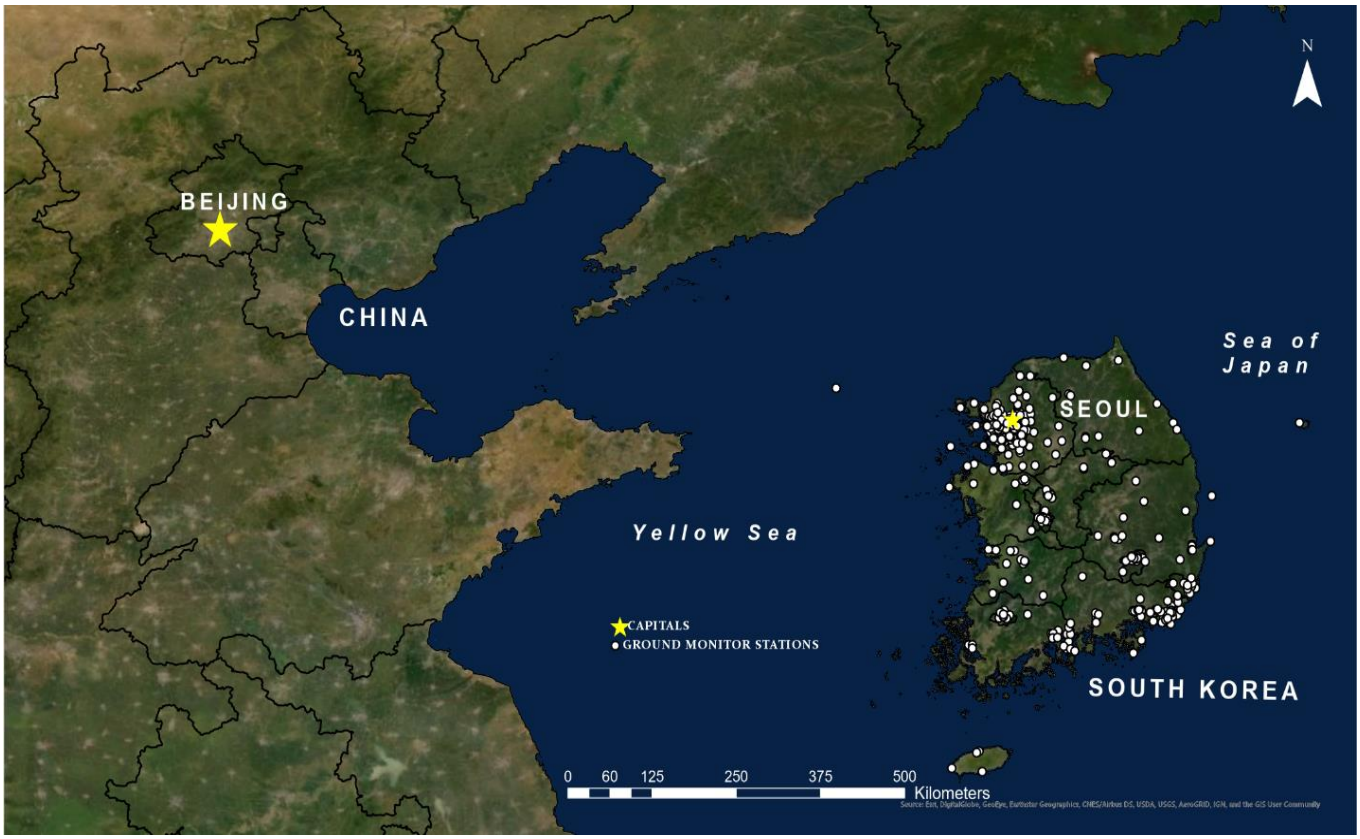


Figure 2a-d. Correlation matrix showing the correlation coefficients between hourly GOCI AOD for the study period

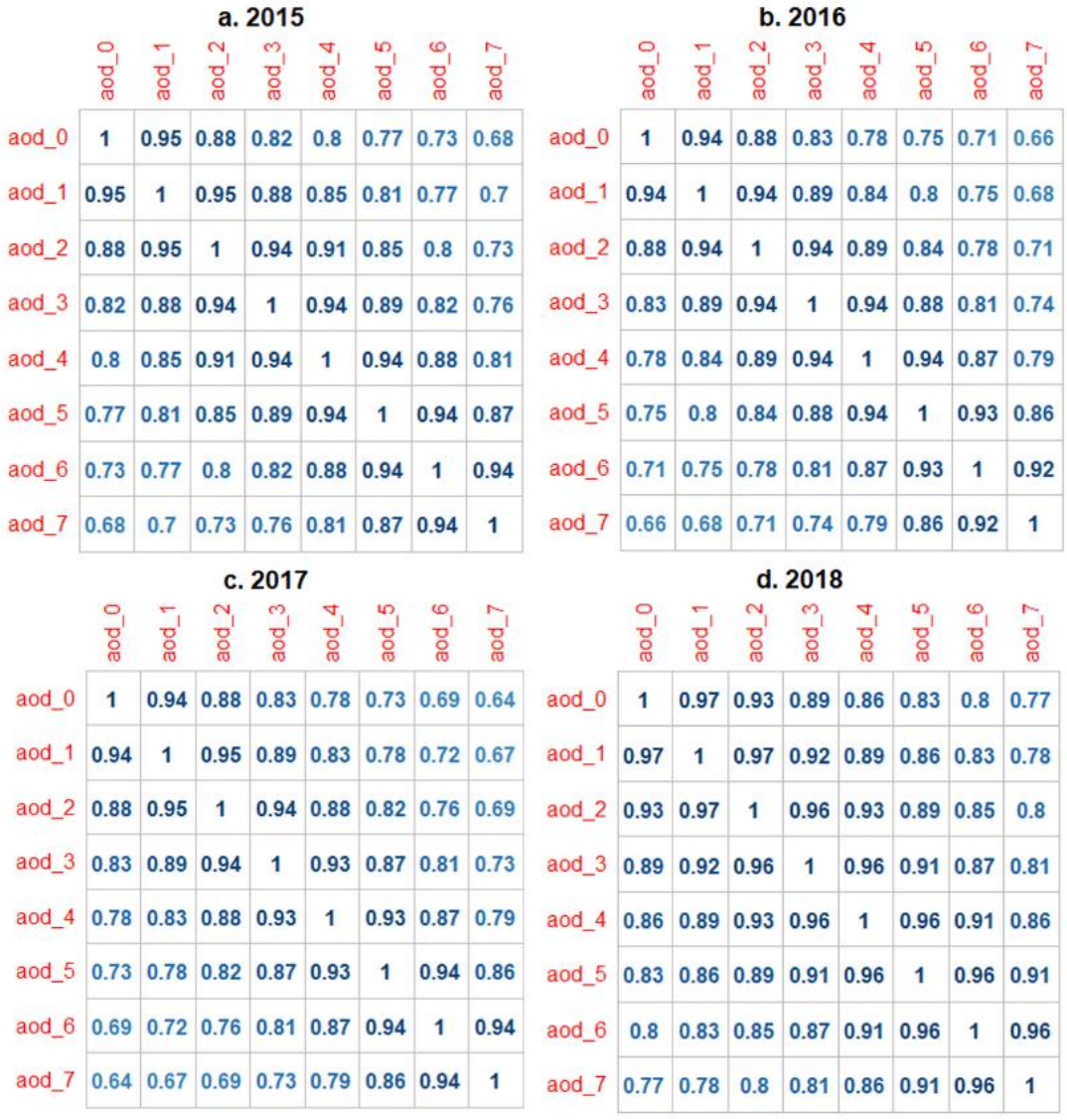


Table 1.

Descriptive statistics of time-varying and spatially fixed parameters

Parameters	Minimum	Mean (Std. Dev.)	Maximum
GOCI Products			
AOD	-0.05	0.27 (0.26)	3.59
FMF 550 nm	0.16	0.44 (0.05)	0.96
Ground Stations			
$PM_{2.5}$ ($\mu g/m^3$)	0	27 (16.50)	185
Ancillary Data			
Elevation (m)	0	139.08 (212.00)	1304.77
Meteorology			
10 m Wind Speed (m/s)	0.08	4.83 (2.25)	51.42
500 mb Wind Speed (m/s)	0.04	12.98 (8.49)	57.00
Surface Pressure (Pa)	903.44	1002.07 (20.61)	1036.40
2 m Air Temperature (K)	238.19	290.49(9.74)	311.25
Planetary Boundary Layer Height (m)	59.14	888.74 (354.15)	4305.28
Relative Humidity	$7.62*10^{-5}$	0.00022 ($1.6*10^{-4}$)	0.00307
Surface Incident Shortwave Flux (W/m^2)	17.22	424.92 (172.02)	903.16
Total Cloud Fraction	0	0.36 (0.31)	1
Total Precipitation (kg/m^2s^2)	0	0.000024 ($2.0*10^{-4}$)	0.01849

Figure 3. Time series illustrating the average monthly $PM_{2.5}$ concentrations (including standard deviation) from 2015 – 2018

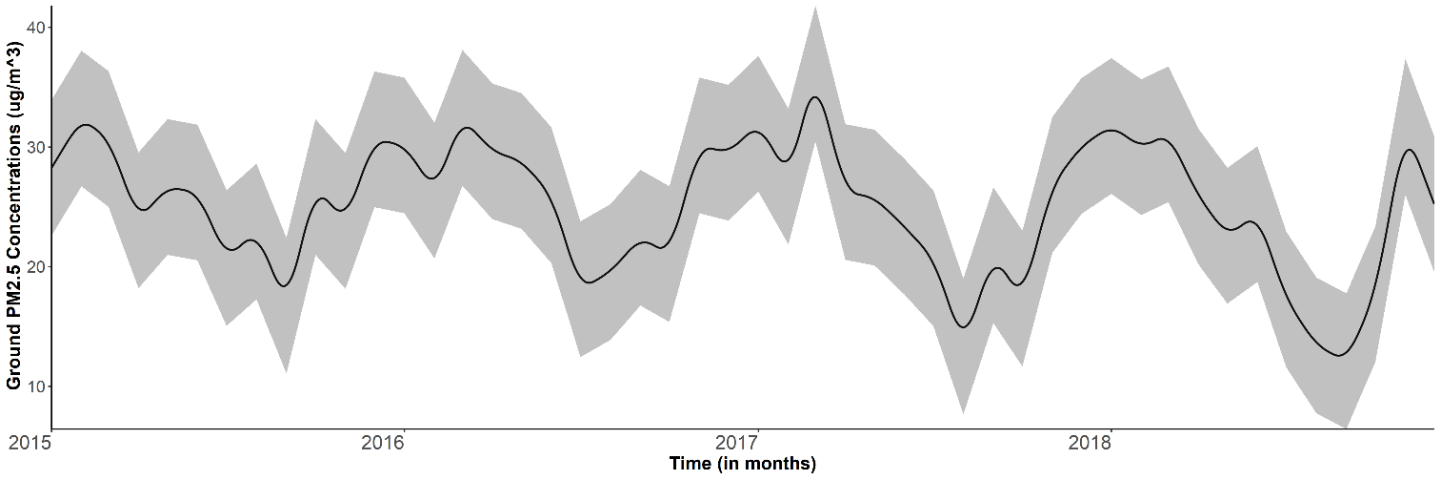


Figure 4. Time series illustrating the average seasonal $PM_{2.5}$ concentrations from 2015 – 2018

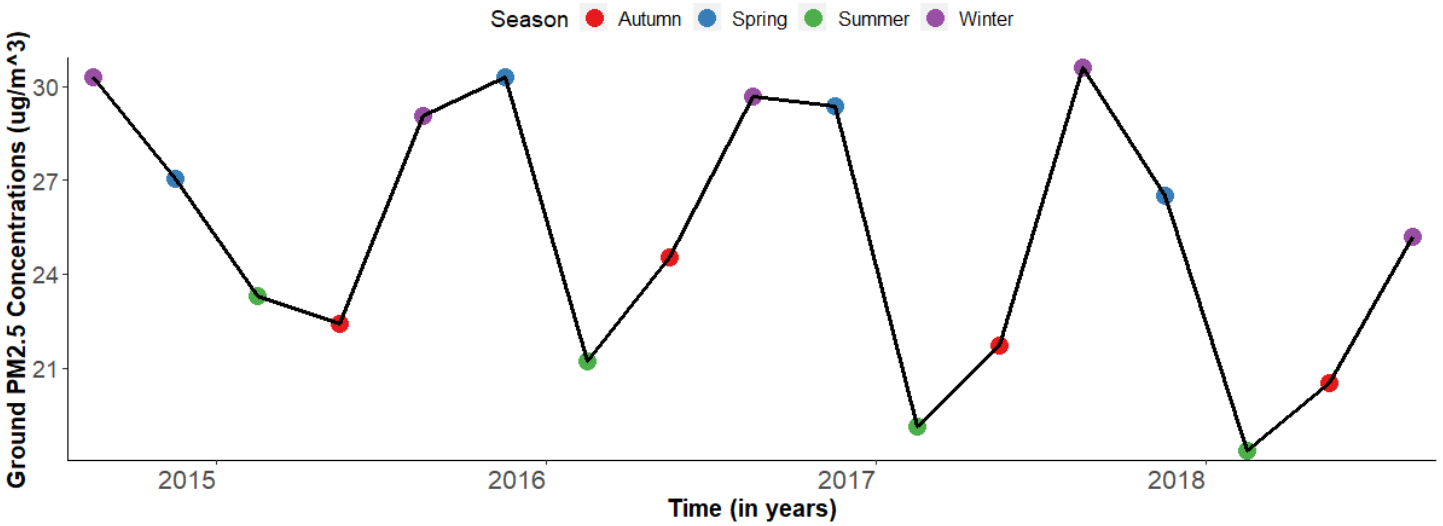


Figure 5. Time series depicting the monthly temporal GOCI AOD coverage in South Korea from 2015 – 2018

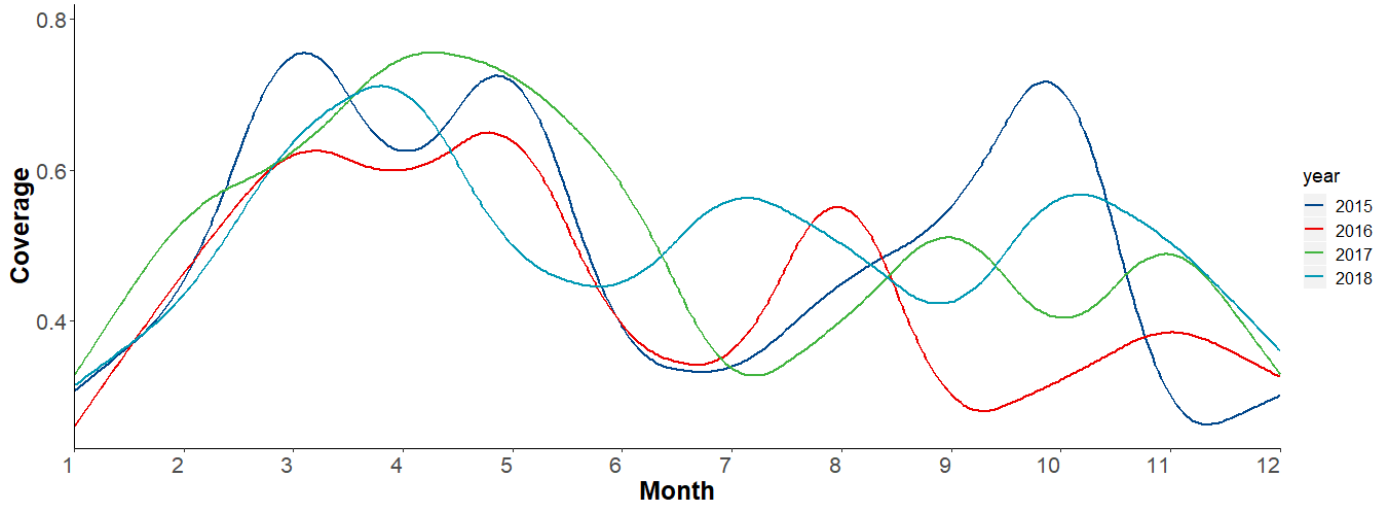
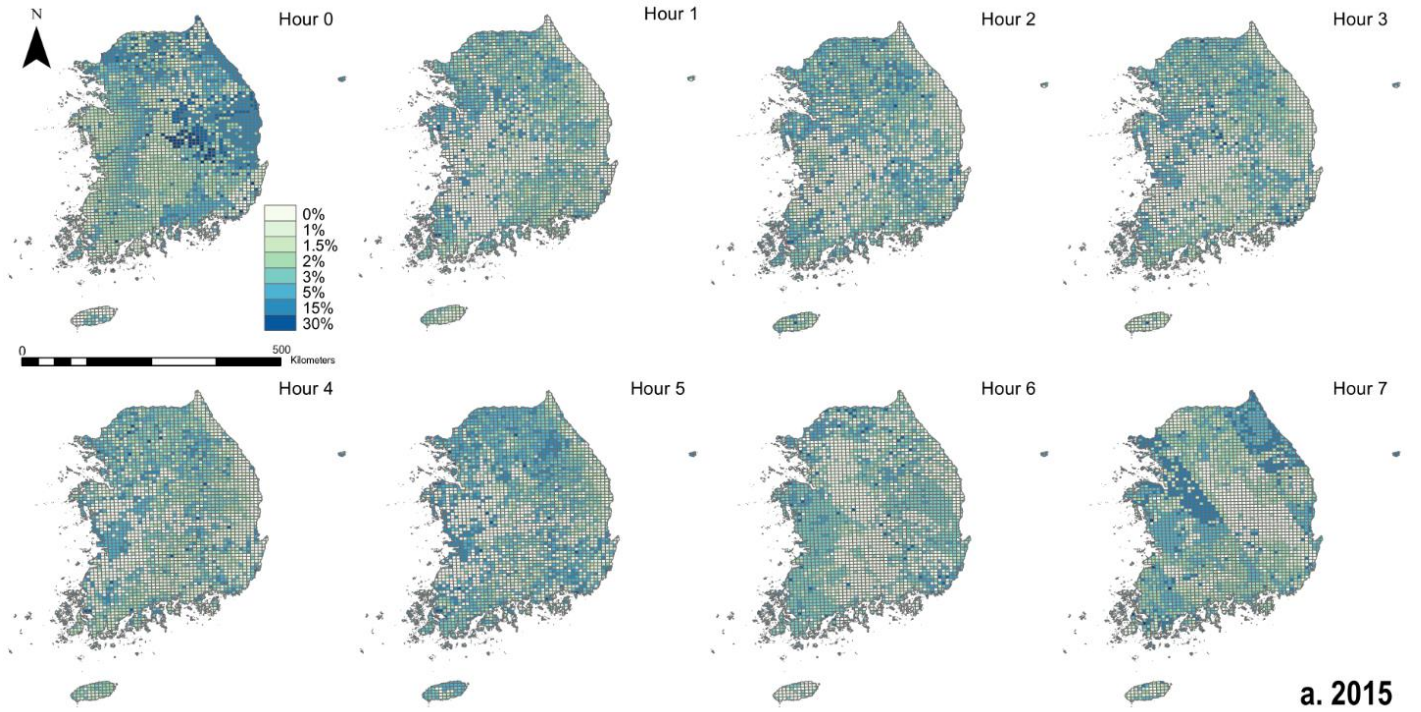
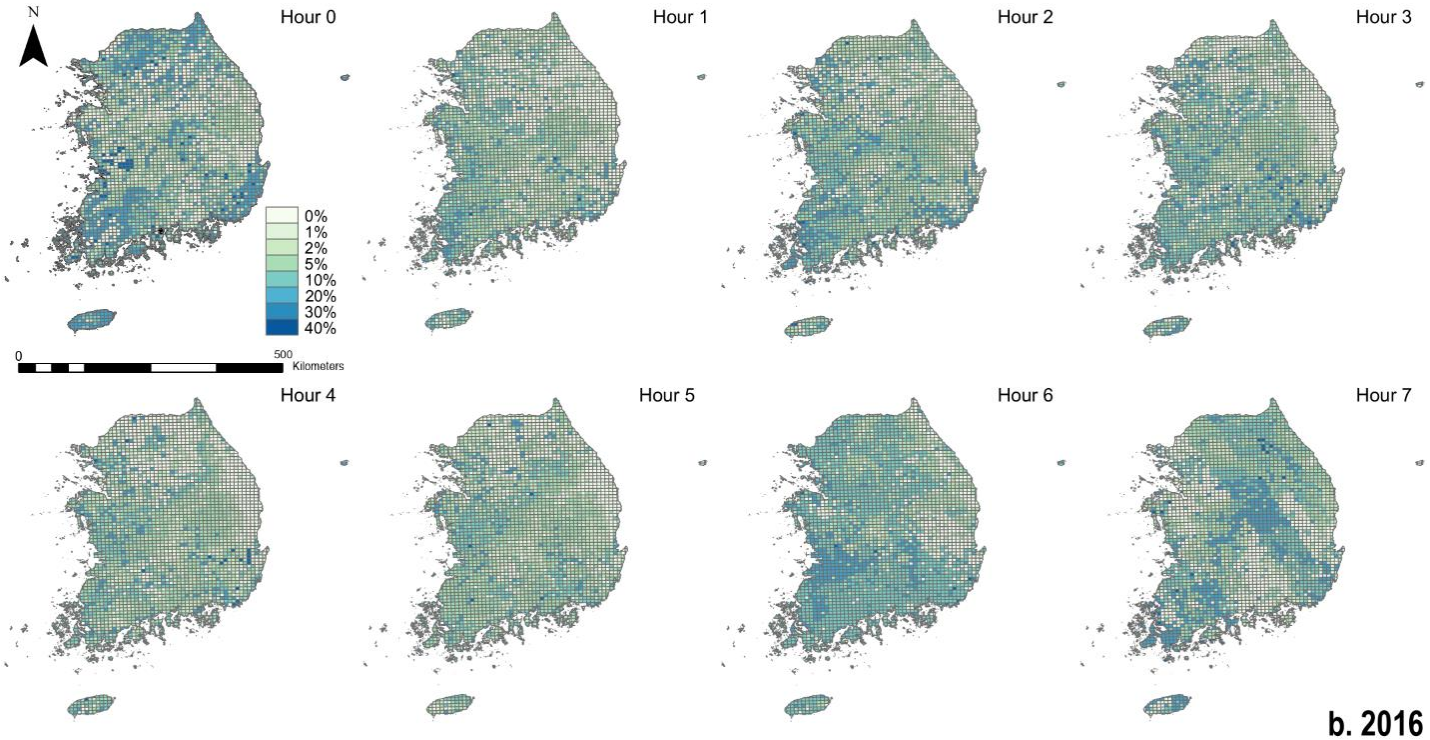
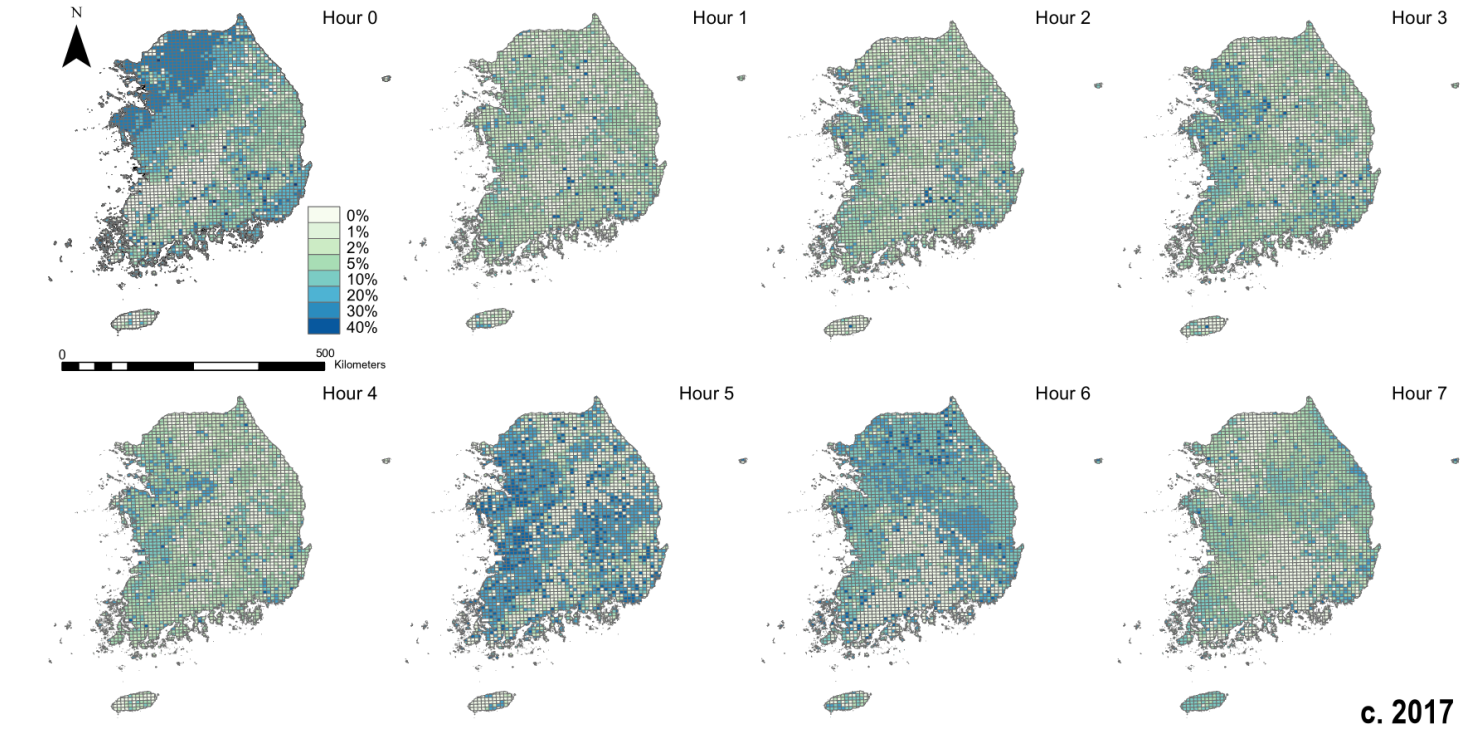


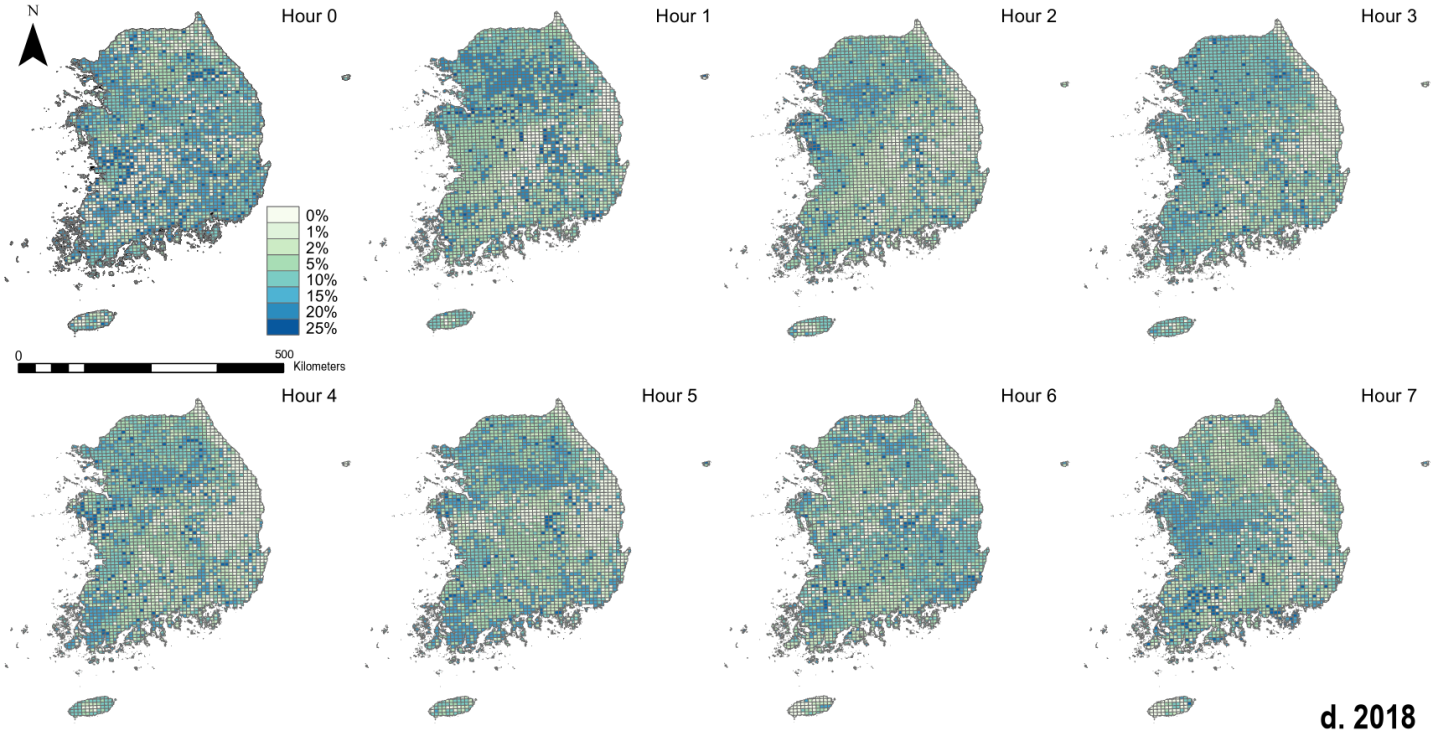
Figure 5a-d. Spatial coverage of AOD per hour over the 4-year study period






N
Hour 0
Hour 1
Hour 2
Hour 3
0%
1%
2%
5%
10%
20%
30%
40%
0
500
Kilometers
Hour 4
Hour 5
Hour 6
Hour 7

c. 2017


N
Hour 0
Hour 1
Hour 2
Hour 3
0%
1%
2%
5%
10%
15%
20%
25%
0
500
Kilometers
Hour 4
Hour 5
Hour 6
Hour 7

d. 2018

Figure 6. Density plot of correlations between the cross validated predicted and measured PM2.5 concentrations of each hour during the study period (2015 – 2018)

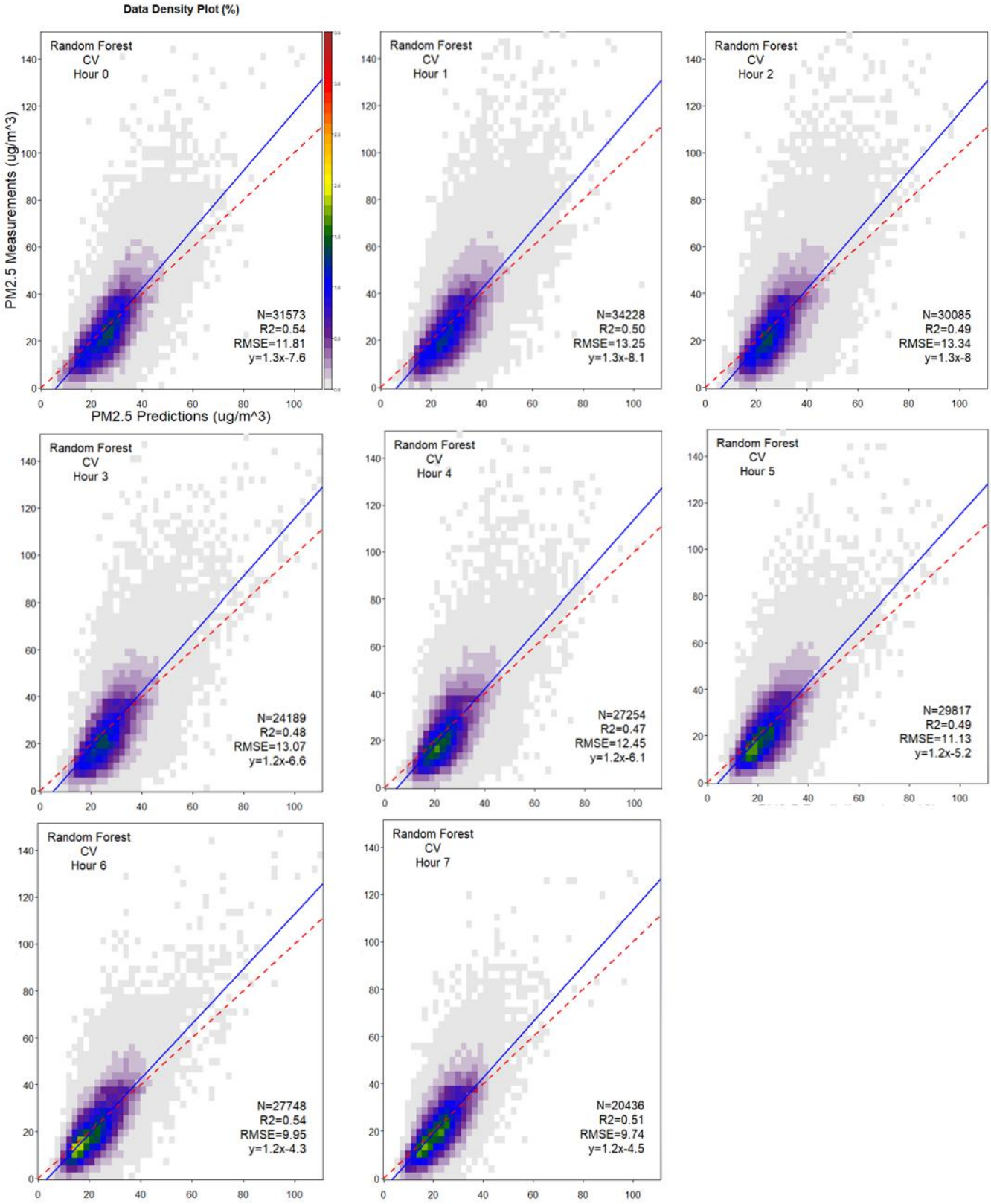


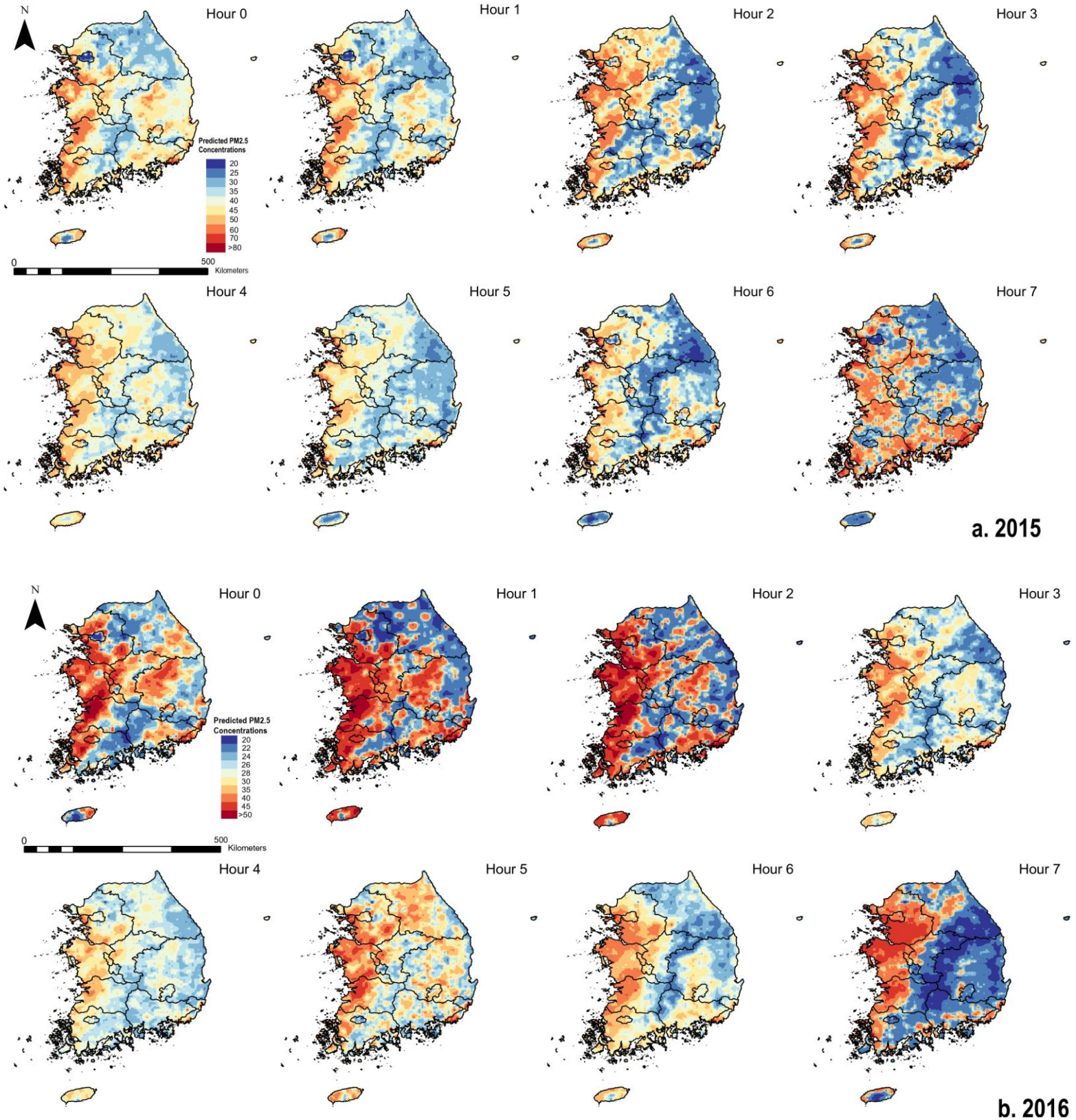
Table 2.

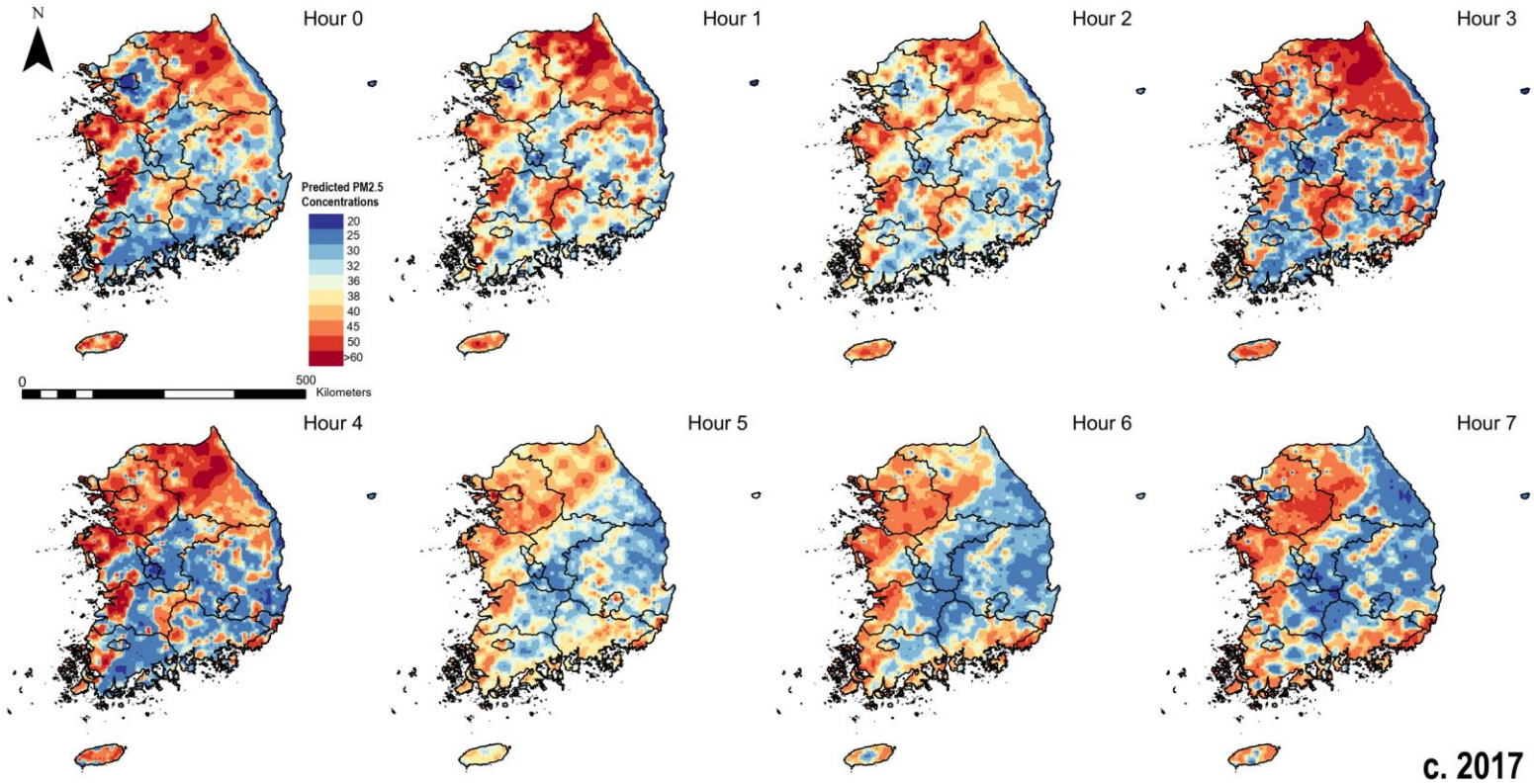
Variable importance in the random forest models (top 5)

Hour 0	Hour 1	Hour 2	Hour 3
AOD	AOD	AOD	AOD
500 mb wind speed	Relative humidity	Relative humidity	Fine mode fraction
Relative humidity	500 mb wind speed	Fine mode fraction	Relative humidity
500 mb wind direction	500 mb wind direction	500 mb wind speed	500 mb wind speed
Surface incident shortwave flux	10 m wind speed	500 mb wind direction	500 mb wind direction

Hour 4	Hour 5	Hour 6	Hour 7
AOD	AOD	AOD	AOD
Fine mode fraction	Fine mode fraction	Relative humidity	Relative humidity
Relative humidity	Relative humidity	Fine mode fraction	500 mb wind speed
500 mb wind speed	500 mb wind speed	500 mb wind speed	500 mb wind direction
500 mb wind direction	500 mb wind direction	Total cloud fraction	Total precipitation

Figure 7a-d. Hourly prediction maps of $PM_{2.5}$ concentrations ($\mu g/m^3$) from the random forest model in South Korea between 2015 – 2018




N
Hour 0
Hour 1
Hour 2
Hour 3
Predicted PM2.5 Concentrations
20
25
30
32
36
38
40
45
50
>60
0
500
Kilometers
Hour 4
Hour 5
Hour 6
Hour 7
c. 2017

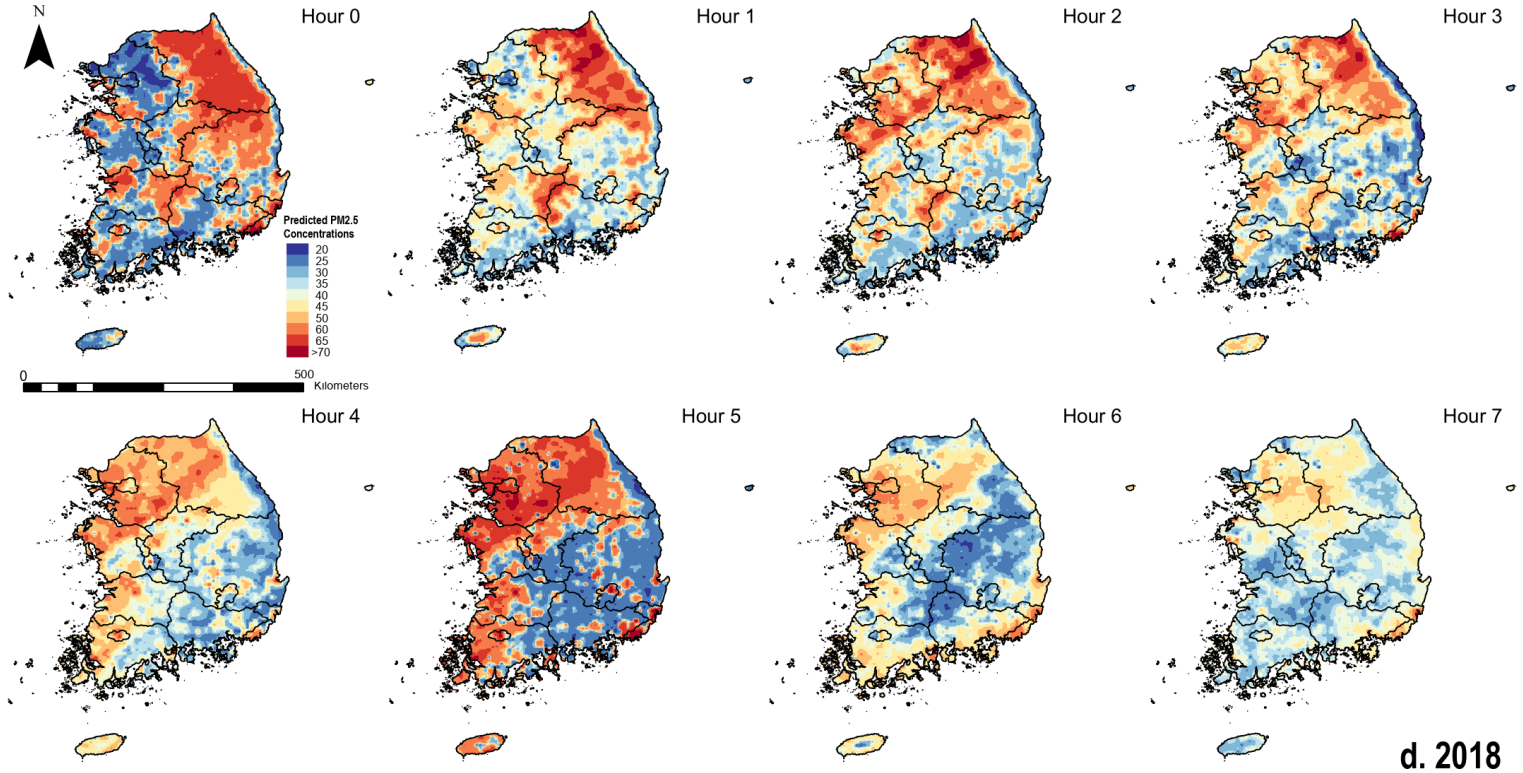

N
Hour 0
Hour 1
Hour 2
Hour 3
Predicted PM2.5 Concentrations
20
25
30
35
40
45
50
60
65
>70
0
500
Kilometers
Hour 4
Hour 5
Hour 6
Hour 7
d. 2018

Figure 8. Predicted hourly spatiotemporal variations of long range transport episode from China to South Korea (March 10th – 11th, 2015) **a.** hourly predictions on March 10th, 2015 **b.** hourly predictions on March 11th, 2015