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:

_____ _____

Qingyang Xiao Date

The development and application of advanced $PM_{2.5}$ exposure models driven by satellite data

By
Qingyang Xiao
Doctor of Philosophy
Environmental Health Sciences

_____
Yang Liu, PhD
Advisor

_____
Howard H. Chang, PhD
Committee Member

_____
Haidong Kan, PhD
Committee Member

_____
Mitchel Klein, PhD
Committee Member

_____
Matthew J. Strickland, PhD
Committee Member

Accepted:

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

_____
Date

The development and application of advanced $PM_{2.5}$ exposure models driven by satellite data

By

Qingyang Xiao
MPH, Emory University, 2014
BS, Peking University, 2012

Advisor: Yang Liu, PhD

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in Environmental Health Sciences
2018

Abstract

The development and application of advanced $PM_{2.5}$ exposure models driven by satellite data

By Qingyang Xiao

Introduction

Satellite aerosol optical depth (AOD) has been increasingly used to predict ground level $PM_{2.5}$ concentrations and assess $PM_{2.5}$ exposures. However, non-random missing AOD due to cloud/snow cover and the complex non-linear relationship between AOD and $PM_{2.5}$ concentration make this task highly challenging. Previous studies used ground $PM_{2.5}$ measurements to fill missing data and included predictors constructed from ground measurements to improve model performance; however, these strategies cannot be applied in developing regions where historical air quality measurements are unavailable. In this study, we developed an original gap-filling method that provided high-resolution complete-coverage $PM_{2.5}$ predictions (Aim 1). Then the maternal $PM_{2.5}$ exposure was assessed by satellite-based $PM_{2.5}$ predictions to estimate its associations with adverse birth outcomes in Shanghai, China (Aim 2). In Aim 3, an ensemble machine learning model was developed to hindcast historical $PM_{2.5}$ levels in China where routine air quality monitoring began only recently.

Methods

For Aim 1, we applied the Multiple Imputation (MI) method that combined the emerging high-resolution satellite retrievals with chemical transport model (CTM) AOD simulations and cloud fraction retrievals to fill missing AOD. Then we fitted a two-stage statistical model driven by gap-filled AOD, meteorology and land use information to estimate daily $PM_{2.5}$ concentrations in the Yangtze River Delta at 1-km resolution. For Aim 2, birth registration records of 132 783 singleton live births during 2011-2014 in Shanghai were obtained and maternal exposures were assessed with satellite predictions from Aim 1. Linear and logistic regressions were used to estimate associations with term birth weight and term low birth weight, respectively. Logistic and discrete-time survival models were used to estimate associations with preterm birth. For Aim 3, a clustering method was designed to control unobserved spatial heterogeneity in $PM_{2.5}$ prediction models. Regional models for each cluster were trained with various machine learning algorithms, including random forest, generalized additive model and extreme gradient boosting. Then we fitted a generalized additive model that fused predictions from these algorithms to improve hindcast accuracy and robustness.

Results

In Aim 1, our gap-filling method did not rely on ground $PM_{2.5}$ measurements and performed better than previous gap-filling methods with complete coverage and high accuracy. In Aim 2, we observed decreased term birth weight, increased risk of preterm birth, and increased risk of term low birth weight in association with maternal $PM_{2.5}$ exposure. We noticed that satellite-based exposure assessments without accounting for missing data led to attenuation of estimated health effects. In Aim 3, our ensemble model

provided more accurate $PM_{2.5}$ hindcasts at daily and monthly level compared with previous models. Cluster-based models outperformed corresponding national models.

Conclusions

We presented a gap-filling method that corrected the exposure bias due to missing satellite data and a machine learning-based ensemble model that provided reliable historical $PM_{2.5}$ predictions. Our methods can support epidemiological studies on the chronic and acute health effects of $PM_{2.5}$ in highly polluted regions with limited ground $PM_{2.5}$ monitoring.

The development and application of advanced $PM_{2.5}$ exposure models driven by satellite data

By

Qingyang Xiao
MPH, Emory University, 2014
BS, Peking University, 2012

Advisor: Yang Liu, PhD

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in Environmental Health Sciences
2018

Acknowledgement

I am grateful to all the faculty, staff, and friends I met during my study at Emory University. I cannot show enough gratitude to my dissertations committee members. This work would not have been possible without their guidance and support. I would especially like to thank Dr. Yang Liu, my advisor and committee chair, for the advice and encouragement he has provided on scientific research, professional career, and life. I appreciate the friendship and accompaniment from members of the Emory Remote Sensing Group.

I would like to thank my parents for their love and encouragement. I would like to thank my family members for sharing all the ups and downs of my life, though we are thousands of miles apart.

My work was supported by NASA Applied Sciences Program (Grant # NNX14AG01G, NNX16AQ28G), the Jet Propulsion Laboratory (Contract # 1363692), the EPA STAR program (Grant # 83586901), and the National Institutes of Health (Grant # R01ES027892). The contents of the publication are solely the responsibility of the grantee and do not necessarily represent the official views of the U.S. EPA. Further, the U.S. EPA does not endorse the purchase of any commercial products or services mentioned in this publication.

Contents

List of Tables

CHAPTER 1

CHAPTER 2

CHAPTER 3

List of Figures

CHAPTER 1

CHAPTER 2

CHAPTER 3

INTRODUCTION

Numerous studies have documented the associations between $PM_{2.5}$ (fine particulate matter with an aerodynamic diameter of 2.5 μm or less) and adverse health outcomes. The 2015 Global Burden of Diseases study identified ambient $PM_{2.5}$ as the fifth largest overall risk factor for global mortality and $PM_{2.5}$ exposure is responsible for 4.2 million deaths in 2015 [1]. This study also pointed out that the uncertainty of the estimated health burden was partly due to the lack of epidemiological evidence on the health effects of $PM_{2.5}$ in highly polluted regions. Most studies on the health effects of $PM_{2.5}$, especially chronic health effects of $PM_{2.5}$, are conducted in developed regions where historical monitoring records are available and exposure levels are low. Epidemiological studies in highly polluted regions are needed to further elucidate the magnitude of $PM_{2.5}$-associated health effects, and provide crucial information on the shape of $PM_{2.5}$ concentration-response curves at high exposure levels [2]. However, these studies are hindered by the lack of $PM_{2.5}$ measurements. For instance, in China, the annual average $PM_{2.5}$ exposure can be over 150 $\mu g/m^3$, but the national air quality monitoring network was established since 2013 so that $PM_{2.5}$ measurements before 2013 were unavailable.

To extend ground air quality monitoring networks, satellite remote sensing retrieved aerosol optical depth (AOD) has been increasingly used for air pollution monitoring and exposure assessment in the past decade [3, 4]. Satellite data with broad coverage, a long historical record and high spatial resolutions can contribute to assessment of air pollution exposure levels in epidemiological studies. Specifically, for studies in the U.S., satellite predictions were employed to increase spatial coverage and resolution of ground measurements. For studies in developing regions where long-term monitoring of $PM_{2.5}$ is unavailable, satellite predictions can not only extend ground monitoring networks in space, but also provide valuable information on historical $PM_{2.5}$ levels. However, missingness in satellite data and degraded hindcast quality have raised

concerns regarding the usage of satellite predictions in epidemiological studies in these regions [5, 6]. Previously developed $PM_{2.5}$ prediction models in the North America and Europe used ground $PM_{2.5}$ measurements to fill missing satellite data [7-9], and included daily random effects or predictors constructed from $PM_{2.5}$ measurements to improve model performance [7, 10]. Unfortunately, these strategies cannot be applied to regions where historical air pollution measurements are unavailable. Thus, a gap-filling method without relying on ground measurements is needed in developing regions to ensure unbiased long-term exposure assessments aggregated during certain exposure windows. Similarly, although daily random effects controlled the unobserved temporal variations in associations between $PM_{2.5}$ concentrations and explanatory variables, applying these random effects outside the model fitting period imposes a strong and often unrealistic assumption that the estimated daily random effects during the model fitting period will remain constant during the hindcast period. Violation of this assumption can partly explain the decreased hindcast accuracy reported in previous models [11, 12]. Including smooth surfaces of $PM_{2.5}$ constructed from ground measurements better described the spatial auto-correlation in $PM_{2.5}$, but sacrificed the model hindcast ability. Thus, a $PM_{2.5}$ prediction model that provides high-accuracy $PM_{2.5}$ hindcast predictions is urgently needed to support environmental health studies in highly polluted regions.

DISSERTATION AIMS

The three aims of this dissertation are listed as follows. **Aim 1**: To develop a multi-stage model that fills the missing AOD values and predicts ground $PM_{2.5}$ concentrations with complete coverage at high resolution. **Aim 2**: To apply satellite predictions from Aim 1 for exposure assessment and estimate the associations between maternal $PM_{2.5}$ exposure and adverse birth outcomes in Shanghai, a highly polluted region. **Aim 3**: To train a machine learning based ensemble model that provides high-accuracy hindcast $PM_{2.5}$ predictions.

To address these aims, we conducted three studies in China. Aim 1 was conducted in Yangtze River Delta, where summer monsoon season with weeks of rainy and cloudy weather leads to more than 60% missing satellite data annually. We obtained the emerging satellite aerosol retrievals at 1-km resolution to reveal local scale variation in $PM_{2.5}$ distribution. We developed an original gap-filling method including satellite cloud products in order to account for aerosol-cloud interactions when filling missing AOD data. Thus, our model provides $PM_{2.5}$ predictions with complete coverage in space and time. To examine the potential benefits of employing fine-resolution satellite predictions in exposure assessment and the effects of missing satellite data on the estimated chronic health effects, we conducted an epidemiological study in Shanghai with three exposure metrics assessed from satellite predictions with missingness, gap-filled satellite predictions with complete coverage, and measurements from ground central monitors (Aim 2). In Aim 3, a national model was developed with satellite data at 10-km resolution. We aimed to extend $PM_{2.5}$ monitoring networks in time. Thus, we abandoned daily effects and predictors constructed from ground measurements. We also proposed a clustering method that improved model performance by controlling unobserved spatial heterogeneity. To improve accuracy and robustness of the satellite driven $PM_{2.5}$ predictions, we trained various machine learning models, including random forest, extreme gradient boosting, and generalized additive model, and fused predictions from these models by an ensemble model.

The methods developed in Aim 1 and Aim 3 allow researchers to estimate $PM_{2.5}$ levels in regions with limited $PM_{2.5}$ monitoring data and assess health effects of $PM_{2.5}$ in these regions. The findings in Aim 2 indicated that satellite predictions without accounting for missing data led to attunation of estimated chronic health effects of $PM_{2.5}$. Exposure assessed from high-resolution satellite predictions revealed local-scale spatial variations and increased the precision of estimated health effects.

REFERENCES

1. Forouzanfar, M.H., et al., *Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015.* The Lancet, 2016. **388**(10053): p. 1659-1724.
2. Tonne, C., *A call for epidemiology where the air pollution is.* The Lancet Planetary Health, 2017. **1**(9): p. e355-e356.
3. Sorek-Hamer, M., A.C. Just, and I. Kloog, *Satellite remote sensing in epidemiological studies.* Current opinion in pediatrics, 2016. **28**(2): p. 228-234.
4. Weng, Q., et al., *Use of earth observation data for applications in public health.* Geocarto International, 2014. **29**(1): p. 3-16.
5. Li, R., et al., *Estimating ground-level pm 2.5 using fine-resolution satellite data in the megacity of Beijing, China.* Aerosol Air Qual. Res, 2015. **15**: p. 1347-1356.
6. Ma, X., et al., *Can MODIS AOD be employed to derive PM2. 5 in Beijing-Tianjin-Hebei over China?* Atmospheric Research, 2016. **181**: p. 250-256.
7. Kloog, I., et al., *A new hybrid spatio-temporal model for estimating daily multi-year PM 2.5 concentrations across northeastern USA using high resolution aerosol optical depth data.* Atmospheric Environment, 2014. **95**: p. 581-590.
8. Kloog, I., et al., *Using new satellite based exposure methods to study the association between pregnancy PM 2.5 exposure, premature birth and birth weight in Massachusetts.* Environmental Health, 2012. **11**(1): p. 1.
9. Just, A.C., et al., *Using high-resolution satellite aerosol optical depth to estimate daily PM2. 5 geographical distribution in Mexico City.* Environmental science & technology, 2015. **49**(14): p. 8576-8584.
10. Di, Q., et al., *Assessing PM2. 5 exposures with high spatiotemporal resolution across the continental United States.* Environmental science & technology, 2016. **50**(9): p. 4712-4721.
11. Ma, Z., et al., *Satellite-based spatiotemporal trends in PM2. 5 concentrations: China, 2004-2013.* Environmental Health Perspectives (Online), 2016. **124**(2): p. 184.
12. Xiao, Q., et al., *Full-coverage high-resolution daily PM2. 5 estimation using MAIAC AOD in the Yangtze River Delta of China.* Remote Sensing of Environment, 2017. **199**: p. 437-446.

Chapter 1

Full-coverage high-resolution daily $PM_{2.5}$ estimation using MAIAC AOD in the Yangtze River Delta of China

Qingyang Xiao, Yujie Wang, Howard H. Chang, Xia Meng, Guannan Geng, Alexei Lyapustin,Yang Liu

Xiao Q, Wang Y, Chang HH, Meng X, Geng G, Lyapustin A, et al. 2017. Full-coverage high-resolution daily $PM_{2.5}$ estimation using maiac AOD in the Yangtze River Delta of china. *Remote Sensing of Environment* 199:437-446.

ABSTRACT

Satellite aerosol optical depth (AOD) has been used to assess population exposure to fine particulate matter ($PM_{2.5}$). The emerging high-resolution satellite aerosol product, Multi-Angle Implementation of Atmospheric Correction (MAIAC), provides a valuable opportunity to characterize local-scale $PM_{2.5}$ at 1-km resolution. However, non-random missing AOD due to cloud/snow cover or high surface reflectance makes this task challenging. Previous studies filled the data gap by spatially interpolating neighboring $PM_{2.5}$ measurements or predictions. This strategy ignored the effect of cloud cover on aerosol loadings and has been shown to exhibit poor performance when monitoring stations are sparse or when there is seasonal large-scale missingness. Using the Yangtze River Delta of China as an example, we present a Multiple Imputation (MI) method that combines the MAIAC high-resolution satellite retrievals with chemical transport model (CTM) simulations to fill missing AOD. A two-stage statistical model driven by gap-filled AOD, meteorology and land use information was then fitted to estimate daily ground $PM_{2.5}$ concentrations in 2013 and 2014 at 1 km resolution with complete coverage in space and time. The daily MI models have an average R^2 of 0.77, with an inter-quartile range of 0.71 to 0.82 across days. The overall model 10-fold cross-validation R^2 (root mean square error) were 0.81 (25 $\mu g/m^3$) and 0.73 (18 $\mu g/m^3$) for year 2013 and 2014, respectively. Predictions with only observational AOD or only imputed AOD showed similar accuracy. Comparing with previous gap-filling methods, our MI method presented in this study performed better with higher coverage, higher accuracy, and the ability to fill missing $PM_{2.5}$ predictions without ground $PM_{2.5}$ measurements. This method

can provide reliable $PM_{2.5}$ predictions with complete coverage that can reduce bias in exposure assessment in air pollution and health studies.

KEYWORDS

$PM_{2.5}$, MAIAC, Chemical Transport Model (CTM), multiple imputation, gap filling, cloud fraction

INTRODUCTION

Ambient air pollution, mostly $PM_{2.5}$ (fine particulate matter with an aerodynamic diameter of 2.5 μm or less), is responsible for more than 3 million premature deaths per year around the world in 2010 [1]. The highest per capita mortality is reported in the Western Pacific region where persistent high $PM_{2.5}$concentrations together with extremely high population density have raised serious public health concerns [2]. However, accurately assessing air pollution exposure in this region is challenging due to limited air pollution monitoring. To support exposure assessment for epidemiological studies and risk analysis, satellite aerosol optical depth (AOD) with global coverage, relatively high resolution, and a long data record has been employed to predict air pollution levels in the past decade [3-5]. Previous studies indicated that satellite data can effectively extend ground air quality monitoring networks, but are challenged by non-random missingness due to cloud/snow cover, high surface reflectance, and extremely high aerosol loading that can be misclassified as cloud [6, 7].The non-random missingness in AOD retrievals may lead to bias in exposure assessment due to potential systematic differences in $PM_{2.5}$ concentrations when AOD is missing or retrieved. Zheng, Zhang [8] reported that the accuracy of annual $PM_{2.5}$ predictions was lower than daily $PM_{2.5}$ predictions due to missingness in AOD, even after correcting annual $PM_{2.5}$ predictions with ground measurements. Other researchers raised concerns that large-scale seasonal missingness in satellite AOD will limit its usage in exposure assessment [9, 10].

To improve the coverage of $PM_{2.5}$ predictions and reduce bias in exposure assessment, various gap-filling methods have been proposed recently. One strategy is to develop regional retrieval algorithms that are more suitable for local geographic conditions and atmospheric characters to

retrieve more AOD pixels. For example, Li, Chen [11] improved the AOD retrieval algorithm and successfully retrieved AOD over bright targets in urban areas of north China during winter time where the MODIS Dark Target algorithm has failed. Van Donkelaar, Martin [7] relaxed the cloud screening criteria of MODIS Dark Target algorithm when studying the Moscow fire event in 2010, leading to a 21% increase in AOD coverage. Although this strategy can significantly increase the coverage and potentially improve the accuracy of satellite AOD, it is restricted to specific study regions and cannot fill missing AOD with true cloud coverage. Another strategy is to use spatial statistical models to estimate missing retrievals from the spatiotemporal autocorrelation of $PM_{2.5}$. For example, Just, Wright [12] used regional daily average $PM_{2.5}$ concentration and spatial smooth function to fill in missing $PM_{2.5}$ predictions. Kloog, Nordio [3]used inverse probability weighting to address the non-random missingness when fitting prediction models, and then interpolated the missing $PM_{2.5}$ predictions using $PM_{2.5}$ predictions or measurements in surrounding grid cells with spatial smoothing. This method can improve the prediction coverage, but by relying on measurements from monitoring stations, it cannot fill missing data when predicting historical $PM_{2.5}$ concentrations before the establishment of air quality monitoring network, and it may exhibit poorer performance if the monitoring networks are sparse or when data over large geographical regions are missing. For example, in Southeastern China, monsoon season leads to several months of rainy and cloudy weather covering several provinces. Additionally, since the spatial pattern was normally fitted monthly or seasonally, it may underestimate the variance of $PM_{2.5}$. Moreover, complex cloud-aerosol interaction has been reported by previous studies [13, 14]. Cloud cover is associated with meteorological conditions that affect aerosol production and deposition [15], thus $PM_{2.5}$ concentrations may not be spatially similar under versus outside a cloud and filling $PM_{2.5}$ concentrations from nearby predictions may introduce error.

In addition to remote sensing techniques, chemical transport models (CTM), such as GEOS-Chem [16] and CMAQ [17], have also been widely used to characterize atmospheric aerosol distribution, including $PM_{2.5}$ concentrations, $PM_{2.5}$ composition, and AOD. However, the accuracy of CTM simulations depend on the emissions inventory, meteorological input data as well as parameterization of chemical and physical processes included in the model[18]. Previous studies reported that the prediction error of CTM model varied spatially and seasonally [19, 20], and biased health effect estimates in epidemiological study [21]. For example, Appel, Chemel [20] reported that CMAQ overestimated $PM_{2.5}$ by more than 30% over North America and underestimated $PM_{2.5}$ by up to 55% in winter in Europe. Quennehen [22] evaluated seven models' performance in predicting ozone and aerosols over East Asia. They showed an overestimation in black carbon and sulfate aerosols in urban regions in China, as well as a general underestimation in scattering aerosols in the boundary layer, due to errors in emissions inventory and physical processing. Although fusing ground measurements and model simulations could improve prediction accuracy [23], the error in CTM simulations may not be fully corrected in the fusion results. Moreover, although the spatial resolution of some CTM simulations can be as high as 4 km in regional studies, typical model simulations are at a relatively low spatial resolution (> 10 km), thus can hardly detect local-scale pollution variability that may be critical for some epidemiological studies [24].

In this study, we propose a method that brings together the emerging satellite aerosol product and CTM simulations by multiple imputation to fill missing AOD. Taking advantage of the high resolution and high accuracy of the latest MAIAC satellite AOD and the complete coverage of CTM AOD, our model provide high-accuracy $PM_{2.5}$ predictions with complete coverage at a 1-km resolution. By filling in the missingness in AOD rather than in $PM_{2.5}$ predictions, this model

also reduced systemic prediction error by including the meteorology and land use information of all the grid cells in model development. This method is generalizable and can provide high-quality $PM_{2.5}$ predictions in other regions, especially in regions with large-scale missingness in AOD.

METHODS

Study Region

The study region (about 200,000 km^2) covers the Yangtze River Delta of China including Jiangsu Province, Zhejiang Province and Shanghai Metropolitan area (Figure 1.1). It is one of the most populated regions on earth with approximately 156 million residents in 2010. This region is affected by summer monsoon with rainy and cloudy weather. A 50-km buffer was used in data collection and model development to ensure that gap-filled AOD and estimated $PM_{2.5}$ concentrations are of the same accuracy near the boundary as in the rest of the study domain.

Datasets

MAIAC AOD data

The latest AOD data retrieved by the Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm from measurements of the Aqua (crossover at 1:30 pm local time) and Terra (crossover at 10:30 am local time) Moderate Resolution Imaging Spectroradiometer (MODIS) instruments were used in this study [25, 26]. The fine spatial resolution (1 km) and high accuracy of MAIAC AOD makes it possible to characterize local-scale aerosol heterogeneity [27]. MAIAC algorithm uses time series analysis to characterize spectral surface reflectance which is required for aerosol retrievals. The combination of the time series and spatial analysis helps improve quality of cloud and snow detection. Because high AOD levels over China often trigger false cloud detection thus limiting aerosol retrievals, specific cloud tests and thresholds have been

regionally adjusted to ensure good MAIAC performance. MAIAC AOD retrievals have been used to estimate ground $PM_{2.5}$ distributions and support air pollution epidemiological studies in the US and Mexico [12, 28-30]. In the current study, MAIAC data from January 1, 2013 to April 30, 2015 were obtained from the MAIAC team.

MAIAC provides quality assurance (QA) flags indicating the retrieval quality, including cloud mask, land/water/snow mask and adjacency mask (i.e., proximity to cloud or snow). Data cleaning was conducted based on the QA codes after calibrating MAIAC AOD against ground AOD from the Aerosol Robotic Network (AERONET). MAIAC pixels that were cloud contaminated or covered with snow were excluded [31]. To improve the coverage of MAIAC retrievals, a linear regression between daily Aqua and Terra MAIAC AOD was fitted and the regression coefficients were used to estimate missing Aqua/Terra AOD when only one of them is present. Then the observed and predicted AOD values were averaged to reflect daily aerosol loadings [32]. The 1-km grid of the MAIAC data was used for data integration.

AERONET data

AERONET measurements have been widely used as "ground truth" in satellite retrieval calibration and aerosol characterization [33]. AERONET AOD at 550 nm, interpolated from AOD at 500 and 675 nm, from two stations in our study region (Figure 1. 1), i.e., the Taihu station and the Xuzhou-CUMT station, were downloaded from the Goddard Space Flight Center (http://aeronet.gsfc.nasa.gov/).

$PM_{2.5}$ measurements

There are 204 air quality monitoring stations in the study region (Figure 1. 1). Hourly $PM_{2.5}$ measurements from these stations are published in real time by the China National Environmental Monitoring Center (CNEMC, http://www.cnemc.cn/). Measurements were downloaded from PM25.in (http://pm25.in/), a direct mirror of data from CNEMC. Repeated identical

measurements for at least three continuous hours were removed because these measurements are likely caused by instrument malfunction [34]. Hourly measurements less than 1μg/m^3 were also removed because it is below the instruments' limit of detection. Daily average $PM_{2.5}$ concentrations, calculated from hourly concentrations during 0:00-23:00 local time, were used as the dependent variable of our statistical model. For 2014 and 2015, days with less than 18 (75%) valid hourly measurements were excluded from the analysis. Due to lack of hourly data, ground $PM_{2.5}$ measurements of year 2013 included all the daily average data [5].

Cloud, meteorology and land use data

Cloud fraction (CF) data were obtained from Aqua and Terra Collection 6 level 2 cloud products (MYD06_L2 and MOD06_L2), at 5-km spatial resolution, downloaded from the LAADS website (https://ladsweb.nascom.nasa.gov/index.html). Daily CF was calculated as the average of Aqua and Terra CF. Other meteorological data including planetary boundary layer height (PBLH), mean air temperature, relative humidity, and wind speed in the planetary boundary layer, surface incident shortwave flux, relative humidity and air temperature at 2 m, and total precipitation during the previous day, were extracted from the Goddard Earth Observing System Data Assimilation System GEOS-5 Forward Processing (GEOS 5-FP) at a spatial resolution of 0.25° × 0.3125°. Normalized Difference Vegetation Index (NDVI) data were obtained from Terra MODIS 16-day global NDVI dataset at 500m resolution (MOD13A1).The elevation data were obtained from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM) version 2 at 30 m resolution. Population density data were obtained from the LandScan Global Population Database at 1km resolution [35].

CMAQ simulation

The community multi-scale air quality (CMAQ) model version 5.1 was utilized to simulate $PM_{2.5}$ component concentrations over China during 2013-2015 with a spatial resolution of 36 km. The

model used CB05 as the gas-phase mechanism, AERO6 as the aerosol module, and Regional Acid Deposition Model (RADM) model as the aqueous-phase chemistry. CMAQ was driven by meteorological fields simulated by the Weather Research and Forecasting model (WRF) v3.5.1 (http://www.wrf-model.org/) and emission inventory data from the Multi-resolution Emission Inventory of China (MEIC, http://www.meicmodel.org/). Following Zhang, Vijayaraghavan [36], hourly AOD columns were calculated from the simulated $PM_{2.5}$ component concentrations using empirical equation suggested by Chameides, Luo [37]. Daily AOD values were calculated from hourly AOD values between 9:00 am and 3:00 pm.

Data processing

All these datasets were integrated to the 1-km MAIAC grid covering the study region. The cloud data were matched to the centroid of the MAIAC grid cells by nearest neighbor approach to avoid artificial smoothing of the cloud fraction data. Elevation data were averaged to the 1-km grid. Meteorological data and NDVI data were matched to grid cell centroids by inverse distance weighting to create smooth surfaces of these parameters. CMAQ AOD data were matched to the 1-km grid cells whose centroids were within a given CMAQ grid cell. ArcMap 10.3.1. was used to calculate the road length (km) and green space area (km^2) within each grid cell.

Methods

The workflow of this study is shown in Figure 1. 2. First, we conducted 10-time imputation to fill the missing AOD with an additive imputation model including CMAQ AOD simulation, elevation, MODIS CF, as well as temperature and humidity data from GEOS FP as predictor variables. Second, a two-stage hierarchical statistical model was developed to predict ground $PM_{2.5}$ concentrations.

Multiple imputation

Imputation methods have been developed to substitute missing data with values estimated from other observed parameters, assuming that the variability in the missing data can be fully explained by these parameters. Imputation leads to extra variability due to random error [38]. To address this extra variability, we conducted multiple imputation that imputed missing data multiple times with plausible values. The multiple imputation method properly addresses the uncertainty of the imputation model and the random error in drawing the imputation value [39].

In this study, we employed a flexible statistical model, including cloud fraction, elevation, humidity, temperature, and spatiotemporal trends, to impute the missing AOD. By including cloud fraction and meteorological information, our imputation model also considered the aerosol-cloud interaction. We used a bootstrap method, by repeatedly sampling the original dataset with replacement, to fit this imputation model in order to account for uncertainty in the imputation procedure [40]. Smoothing splines of the X and Y coordinates of grid cell centroids were fitted to represent the spatial trend of AOD. On a given day with no or very few AOD retrievals over the study region, we took advantage of the temporal autocorrelation of AOD and included data on two days prior to and two days after that day for model fitting. Thus, by assuming that the spatial pattern of the relationship between MAIAC AOD and CTM AOD remains constant during each rolling 5-day period controlling for daily variation in cloud fraction, temperature, humidity, and elevation, we predicted the missing AOD on the 3rd day of this 5-day period from the imputation model. Dummy variables of day, ranging between 1 and 5, were included in the model to account for temporal differences of AOD. Smooth functions of quadratic polynomials of X and Y coordinates of the grid cell centroid, as well as the interaction term between X and Y, were included in the model to allow a flexible spatial surface of AOD. Each missing AOD value was

imputed 10 times to generate 10 complete datasets that were used in the following analyses. The additive model is shown as Equation 1:

$$AOD_{jt} = s(X_j) + s(Y_j) + s(X_j^2) + s(Y_j^2) + s(X_j \times Y_j) + \beta_1 CF_{jt} + \beta_2 CMAQ\_AOD_{jt} + \beta_3 Temp_{jt} + \beta_4 RH_{jt} + \beta_5 SH_{jt} + \beta_6 Elev_j + D_t + \varepsilon_{jt} \quad (1)$$

where AOD_{jt} is the average daily MAIAC AOD at cell j on day t; X_j and Y_j are the coordinates (km) of the centroid of grid cell j; CF_{jt} is the daily average cloud fraction at grid cell j on day t; $CMAQ\_AOD_{jt}$ is the daily average CMAQ AOD at grid cell j on day t; $Temp_{jt}$, RH_{jt}, and SH_{jt} are the daily average air temperature (K), average relative humidity, and average specific humidity (kg/kg) under the boundary layer at grid cell j on day t; $Elev_j$ is the elevation (m) at grid cell j; D_t is the dummy variable of five levels that representing the day of period index, and s() represents a smoothing spline with 10 knots specific to the 5-day period. We also considered precipitation in the MI model, but it did not significantly improve the model performance, thus we excluded it from the final MI model.

LME-GAM prediction model

A two-stage statistical model was developed to calibrate the spatiotemporal relationships between $PM_{2.5}$, AOD, meteorological parameters, and land use [41, 42]. The ten datasets from multiple imputation were used to fit the two-stage model separately, and then the $PM_{2.5}$ predictions from these ten models were averaged as final $PM_{2.5}$ predictions. The first stage model is a linear mixed-effects (LME) model that allows the AOD-$PM_{2.5}$relationship to vary daily. Quadratic terms for AOD and interactions between AOD and PBLH were added in the model to account for the non-linear relationship between AOD and $PM_{2.5}$. We also explored other variables, including evaporation from turbulence, wind speed and wind direction at 10 m and at 500 m above the ground, and surface pressure. Including these parameters did not significantly improve the model

performance. Thus we excluded them from the final model. The LME model structure can be expressed as Equation 2:

$$PM_{2.5_{jt}} = (\beta_0 + \theta_0) + (\beta_1 + \theta_1)AOD_{jt} + (\beta_2 + \theta_2)AOD_{jt}^2 + \beta_3 PBLH_{jt} \times AOD_{jt} + (\beta_4 + \theta_4)PBLH_{jt} + (\beta_5 + \theta_5)SH_{jt} + \beta_6 Temp_{jt} + \beta_7 SWGDN_{jt} + \beta_8 SWGDN_{jt}^2 + \beta_9 Wind_{jt} + \beta_{10} NDVI_{jt} + \beta_{11} PRECTOT_{jt} + \varepsilon_{1_{jt}}(\theta_0, \theta_1, \theta_2) + \varepsilon_{2_{jt}}(\theta_4, \theta_5)$$

$$\varepsilon_{1_{jt}} \sim N(\mathbf{0}, \boldsymbol{\psi_1}) \varepsilon_{2_{jt}} \sim N(\mathbf{0}, \boldsymbol{\psi_2}) \quad (2)$$

where β_0 is the fixed intercept; β_1 and β_2 are the fixed slopes of square polynomials for AOD; β_3 is the slope of the interaction between PBLH and AOD; β_7 and β_8 are the fixed slopes of square polynomials for surface incident shortwave flux (SWGDN); β_4, β_5, β_6, β_9, β_{10}, and β_{11} are the fixed slopes of PBLH, specific humidity at 2 m, temperature at 2 m, wind speed under PBL, NDVI, and total precipitation during the previous day; θ_0 is the daily random intercept; θ_1 and θ_2 are the daily random slopes of square polynomials for AOD; θ_4 andθ_5 are the monthly random slope of PBLH and SH.

Then we modeled the residuals of the LME model using a second stage GAM with land use and population density. This GAM was fitted monthly to account for temporal variability, which has the following general structure:

$$PM_{2.5}Resid_{jt} = \mu + s\big((X,Y)_j\big) + s\big(GreenSpace_j\big) + s\big(RoadLength_j\big) + s\big(Population_j\big) + \beta\big(GasStation_j\big) + \varepsilon_{jt} \quad (3)$$

where $s((X,Y)_j)$ is the smooth function of the X and Y coordinates of the centroid of cell j, $s(GreenSpace_j)$ is the smooth function of the area of green space of cell j, $s(RoadLength_j)$ is the smooth function of road length of cell j, $s(Population_j)$ is the smooth function of population density of cell j, $GasStation_j$ is the number of gas stations in cell j, ranging between 0 and 4. Other

land use parameters explored, such as number of railway stations and area of water body, did not contribute significantly to the model and were excluded from the final model.

Ten-fold cross-validation was conducted to detect overfitting of this two-stage prediction model. To detect the over fitting in spatial interpolation, we also conducted spatial ten-fold cross-validation by dropping ten percent of grid cells when fitting the model and using the model to predict $PM_{2.5}$ of the dropped grid cells. For model validation, we fitted a linear regression between measured and predicted $PM_{2.5}$. The linear regression R^2, slope, and intercept, as well as root mean square error (RMSE) and relative prediction error (RPE), were used to evaluate model performance.

Model prediction

Using a $PM_{2.5}$ prediction model with daily random effects to estimate $PM_{2.5}$ concentrations outside the model fitting period tends to generate larger prediction errors, but does not affect point estimate for normal outcomes. However, one advantage of employing the satellite remote sensing data is its long data record that provides information of air pollution before the establishment of ground monitoring networks. To evaluate the model's ability of predicting historical $PM_{2.5}$ concentrations, we used data during a separate period, 2015 January to April, for prediction and evaluation. Since the fitted daily random effects and the second stage residual model may not be valid when predicting $PM_{2.5}$ levels outside the modeling period, we adjusted the LME model by removing the daily random effect and dropped the residual predictions from the GAM model. The average of the ten imputed AOD together with the AOD observations were used to provide complete-coverage $PM_{2.5}$ prediction. Thus, the $PM_{2.5}$ concentrations outside the modeling period were predicted from the LME model with the estimated fixed effects and random effects at monthly level. Two models that were fitted from data of year 2013 and data of year 2014, separately, were used to predict $PM_{2.5}$ levels in 2015.

Case study using other gap-filling methods

To quantitatively compare the performance of the MI gap-filling method with previously reported methods, we selected three representative studies by Lv, Hu [43], Just, Wright [12], and Kloog, Koutrakis [41] that presented various gap filling methods. We used data over Xuzhou and Wuxi in 2014, two cities where AERONET stations are located, as a case study. We compared the AOD filled by the method of Lv, Hu [43] to AERONET AOD and compared the $PM_{2.5}$ filled by the methods of Just, Wright [12] and Kloog, Koutrakis [41] to ground $PM_{2.5}$ measurements.

Following the method presented by Lv, Hu [43], we fitted linear regressions to fill the AOD missingness in grid cells with $PM_{2.5}$ monitoring stations (Equation 4). Linear regressions were fitted separately for each city during the warm (April 16 – October 15) season or the cold season. Then we used ordinary kriging (OK) to interpolate daily AOD surfaces and fill missing AOD. The gap filled AOD in the grid cells with AERONET stations were compared with AERONET AOD. We also compared our filled AOD in the same grid cells with AERONET AOD for evaluation.

$$\mathrm{AOD_{jt}} = \beta_0 + \beta_1 \frac{\mathrm{PM_{jt}}}{\mathrm{PM_{js}}} \mathrm{AOD_{js}} + \varepsilon \quad (4)$$

where AOD_{jt} and PM_{jt} are the AOD value and $PM_{2.5}$ concentration at grid cell j on day t, respectively; AOD_{js} and PM_{js} are the seasonal average AOD value and $PM_{2.5}$ concentration at grid cell j and season s that containing day t, respectively; and B_0 and β_1 are the city- and season-specific intercept and slope, respectively.

Similarly, following the gap-filling methods presented by Just, Wright [12] and Kloog, Chudnovsky [44], we used GAM (Equation 5) and GAM with random effects (Equation 6) to interpolate $PM_{2.5}$ surfaces from $PM_{2.5}$ measurements. The $PM_{2.5}$ monitoring stations in Xuzhou were highly clustered: the furthest distance between two stations in Xuzhou is only 16 km. Thus,

to better evaluate the accuracy of these spatial smoothing methods, we used data over Wuxi, where 13 $PM_{2.5}$ monitoring stations distributed as three clusters, for the case study (Figure 1. S1). To evaluate the spatial inference of the GAM methods, we also used $PM_{2.5}$ measurements from one cluster to fill the missing $PM_{2.5}$ predictions at stations that are not in this cluster but within a 60 km buffer of the center of the cluster. The filled $PM_{2.5}$ concentrations in grid cells with $PM_{2.5}$ monitoring stations were compared with $PM_{2.5}$ measurements.

$$\sqrt{\mathrm{PredPM_{jt}}} = \beta_0 + \beta_1\sqrt{\mathrm{MPM_t}} + s\big((X,Y)_j\big) + \varepsilon_{jt} \quad (5)$$

$$\mathrm{PredPM_{jt}} = (\beta_0 + \theta_0) + (\beta_1 + \theta_1)\mathrm{MPM_t} + s((X,Y)_j) + \varepsilon_{jt}(\theta_0, \theta_1) \;\; \varepsilon_{jt} \sim N(\mathbf{0}, \boldsymbol{\psi}) \quad (6)$$

where $PredPM_{jt}$ and $\sqrt{\mathrm{PredPM_{jt}}}$ are the predicted $PM_{2.5}$ concentration and its square root at grid cell j on day t, respectively. $PredPM_{jt}$ was estimated from observed MAIAC AOD, using the model developed with gap-filled AOD, thus the sampling bias of model fitting was corrected. MPM_t and $\sqrt{\mathrm{MPM_t}}$ are the regional mean measured $PM_{2.5}$ concentration and its square root on day t, respectively; B_0 and β_1 are the fixed intercept and slope, respectively; θ_0 and θ_1 are the daily random intercept and slope, respectively; $s((X,Y)_j)$ in Equation 5 is a monthly tensor product of cubic spline of the X and Y coordinates of the centroid of grid cell j; and $s((X,Y)_j)$ in Equation 6 is a monthly thin plate spline of X and Y.

RESULTS AND DISCUSSION

Coverage of satellite data and multiple imputation

Figure 1. 3 shows the coverage of daily MAIAC AOD after combining Aqua and Terra data. On average, for each cell more than 60% of days are missing. During the summer monsoon season, about 75% of days are missing. The southern region of our modeling domain showed more missingness than the northern region, and elevated areas showed more missingness.

The daily MI model had an average model fitting R^2 as 0.77, ranging between 0.48 and 0.97, with an inter-quartile range of 0.71 to 0.82 across days. In general, days with high coverage also had high model fitting R^2 values, and days with large-scale missingness tended to have lower R^2 values. The imputation method increased data coverage to 100% by filling all the missing AOD values. The mean annual AOD distributions before and after imputation are shown in Figure 1. 4. The spatial patterns of AOD were similar before and after imputation: the AOD value increased from south to north in our modeling region, with relatively high values occurred at urban centers. However, the annual average AOD after imputation is higher than that of retrieved AOD by approximately 0.1. Since AERONET uses a specific procedure for cloud detection [45], it provides AOD observations sometimes when MAIAC had missing AOD. We conducted a student t-test with AERONET AOD on cloudy days when satellite AOD is missing and clear days when satellite AOD is present. The comparison results indicated that AERONET AOD values when satellite AOD is missing were 0.16 higher than those when satellite AOD is present (p-value< 0.01). Previous studies reported that high AOD associated with high cloud fraction from March to August and when AOD is larger than 0.4 [14, 46], because cloud cover leads to increased humidity that favors the hygroscopic growth of aerosols [13]. Since the missingness in satellite AOD in Yangtze River Delta is mainly due to cloud cover, filling the AOD gap led to a higher annual average AOD. We also conducted a student t-test with specific humidity at 2 m, average specific humidity under boundary layer as well as relative humidity under boundary layer of cloudy pixels that has missing AOD and those of clear pixels that has successfully retrieved MAIAC AOD in the modeling dataset. The t-test indicated that all three parameters showed significantly higher values on cloudy days. The results of comparing daily average MAIAC AOD, with and without imputation, and average AERONET AOD are shown in the Figure 1. S2. The average MAIAC AOD before imputation slightly overestimated the average AERONET

AOD. The gap-filled MAIAC AOD agreed with AERONET, with the slope of 0.91. The imputation model underestimated AOD at high aerosol loading.

Performance of the prediction model

Table 1. 1 shows the summary statistics of AOD and $PM_{2.5}$ in 2013 and 2014 model fitting datasets. The average $PM_{2.5}$ level in 2013 was higher than that in 2014 by more than 10 $\mu g/m^3$. There were more precipitation events and higher humidity levels in 2014 than in 2013. On average, the year of 2014 had 8% more missing days year round and 25% more missing days in summer than the year of 2013.

Figure 1. 5 shows the performance of our two-stage prediction model from ten-fold cross-validation. The cross-validation results indicated that $PM_{2.5}$ predictions matched well with observations, with the fitted linear regression line having a slope near unity. Our $PM_{2.5}$ prediction model provided higher accuracy than the previous high-resolution $PM_{2.5}$ prediction model in the YRD. The cross validation R^2 of a 3-km $PM_{2.5}$ prediction model developed by Ma, Liu [47] was 0.67 in 2013, while our model had the cross validation R^2 of 0.81. Our model fitted with 2013 data had a higher R^2 (0.81) than the model fitted with 2014 data (R^2 as 0.73), but the 2013 model had a higher RMSE (25 $\mu g/m^3$) and RPE (34%) than the 2014 model (RMSE as 18 $\mu g/m^3$ and RPE as 29%). This may be partly due to the relatively higher $PM_{2.5}$ level in 2013. When comparing model predictions with measurements, the R^2 of predictions from observed AOD and the R^2 of predictions from AOD imputation were both 0.80 for the year of 2013 and 0.76 vs. 0.69 for the year of 2014. This suggests that the imputation process did not or slightly decrease model accuracy. No overfitting was detected since the model performed similarly in model fitting and in cross-validation. The model fitting R^2 and the cross-validationR^2 was 0.82 and 0.81 for the year of 2013, and 0.75 and 0.73 for the year of 2014 (Figure 1. S3). The ten-fold spatial cross-

validation had R^2 as 0.80 and 0.72 in 2013 and 2014, respectively, indicating that the spatial interpolation of this two-stage model is validated.

By filling the AOD gap, we corrected the bias in annual average $PM_{2.5}$ predictions that was reported in Zheng, Zhang [8]. Zheng showed that their model had leave-one-out cross-validation R^2 of 0.76 when predicting annual average $PM_{2.5}$, lower than the leave-one-out cross-validation R^2 (0.8) when predicting daily average $PM_{2.5}$. In our model, when predicting the annual average $PM_{2.5}$ in 2013 and 2014, the 10-fold cross-validation R^2 was 0.94 and 0.87, respectively, with the relative prediction errors of 7% and 6%, respectively.

$PM_{2.5}$ prediction

In general, predicted $PM_{2.5}$ concentrations from AOD imputation was lower than that from observational AOD. In 2013, the average predicted $PM_{2.5}$from imputed AOD was 56 $\mu g/m^3$ while the average predicted $PM_{2.5}$ from observational AOD was 69 $\mu g/m^3$. Similarly, in 2014, the average predicted $PM_{2.5}$ from imputed AOD and from observational AOD was 50 $\mu g/m^3$ and 62 $\mu g/m^3$, respectively. In other words, $PM_{2.5}$ levels on cloudy pixels were lower relative to sunny pixels. A student t-test comparing ground $PM_{2.5}$ measurements on cloudy days (i.e., satellite AOD is missing) and clear days (i.e., satellite AOD is observed) indicated that $PM_{2.5}$ concentrations on cloudy days were lower than those on clear days by 20 $\mu g/m^3$ (p-value < 0.01) in the study region. This agrees with previous findings in the Southeastern US that $PM_{2.5}$ levels were negatively associated with cloud fraction [15]. We noticed that the AOD values on cloudy days were higher than on clear days, even though in general AOD was positively associated with $PM_{2.5}$ concentrations. One explanation is that in the Yangtze River Delta, a majority of AOD missingness is due to precipitation and cloud cover. Cloud cover leads to reduced photochemical reaction-related $PM_{2.5}$ production and precipitation removes $PM_{2.5}$ from the atmosphere, leading to decreased $PM_{2.5}$ dry mass concentration. However, cloud cover is also associated with

increased humidity that favors the hygroscopic growth of aerosols and leads to higher AOD values. Since $PM_{2.5}$ concentrations are measured at ground monitoring stations with controlled constant humidity and temperature, the increased humidity does not affect $PM_{2.5}$ measurements as significantly as AOD. We also noticed higher average humidity, higher average AOD value, but lower average $PM_{2.5}$ concentrations in 2014 comparing with 2013 (Table 1. 1) likely due to the same reason. Our results indicated that cloud cover and missing AOD was associated with lower in $PM_{2.5}$ measurements but higher AOD loading. Thus, cloud cover and humidity modified the association between AOD and $PM_{2.5}$, and this effect need to be considered when filling missing data.

Figure 1. 6 shows the annual average $PM_{2.5}$ distribution in 2013 and 2014. Over most regions except southern Zhejiang province, the annual $PM_{2.5}$ concentration was higher than the annual National Ambient Air Quality Standard of China (35 $\mu g/m^3$). The highest $PM_{2.5}$ values occurred in urban centers in Jiangsu province, including Taizhou, Changzhou, and Nanjing city. In Zhejiang province, cities in the Jin-Qu Basin and on the coast also had relatively high $PM_{2.5}$ concentrations due to higher population density and associated anthropogenic emissions. The high-resolution prediction map successfully shows local-scale variability in $PM_{2.5}$ concentrations. For example, in Figures 1. 6C and 1. 6D, regions covered by forest (dark green in Figure 1. 6D) had lower $PM_{2.5}$ concentrations relative to their surrounding regions; while urbanized regions, such as Town of Jurong (the blue dot in the lower right corner), had higher $PM_{2.5}$ concentrations relative to their surrounding regions. Temporally, the $PM_{2.5}$ levels decreased from 2013 to 2014 by 7 $\mu g/m^3$ (11%) on average. The largest decrease occurred at urban centers, such as Huai'an, Changzhou, and Taizhou (Figures 1. 6A and 1. 6B).

We used a separate time period to validate our model's prediction ability. Models fitted with data of year 2013 and year 2014 were used to predict weekly and monthly average $PM_{2.5}$

concentrations of year 2015 (Figure 1. 7). At the weekly level, the R^2 value, RMSE, and RPE were 0.45, 32 μg/m^3, and 48% using the 2013 model and 0.48, 19μg/m^3, and 28% using the 2014 model. At the monthly level, the R^2 value, RMSE, and RPE were 0.70, 25 μg/m^3, and 38% using the 2013 model and 0.71, 11μg/m^3, and 17% using the 2014 model. The 2013 model overestimated $PM_{2.5}$ levels in 2015. This may result from the effect of a few severe $PM_{2.5}$ pollution episodes in 2013 that affect the model coefficients when AOD value was high. For example, in the LME model, the fixed slopes of AOD^2 in 2013 and 2014 were 5.19 and -9.56, respectively (Table 1. 1). When the AOD value is high, the model fitted by 2013 data will predict higher $PM_{2.5}$ concentrations relative to the model fitted by 2014 data, and such difference widens with increase in AOD.

Comparisons with other gap-filling methods

A limitation of previously reported gap-filling methods is that they rely on $PM_{2.5}$ measurements. As a result, these methods are not suitable for prediction of historical $PM_{2.5}$ concentrations when $PM_{2.5}$ measurements were sparse or nonexistent. For example, in China, this method cannot be used to fill missingness in $PM_{2.5}$ predictions before 2013. In our case study in Xuzhou and Wuxi, the gap-filling methods of Lv, Hu [43], Just, Wright [12] and Kloog, Koutrakis [41] were able to increase data coverage to 94% in 2013, with 6% missingness due to lack for valid ground measurements; our MI method increased the data coverage to 100% (Table 1. 2). Another limitation of previous gap-filling methods is that they do not consider the aerosol-cloud interactions, leading to bias in spatial inference. When comparing filled AOD with AERONET AOD, the MI method provided higher R^2 (0.44) than the Lv, Hu [43] method (R^2 = 0.18), indicating that the filled AOD from our MI method had higher accuracy. When comparing filled $PM_{2.5}$ concentrations with $PM_{2.5}$ measurements, the MI method provided comparable R^2 (0.78) with the Just, Wright [12] (R^2 as 0.84) and Kloog, Koutrakis [41] (R^2 = 0.79) methods. When

using $PM_{2.5}$ measurements from stations in one cluster to fill missing $PM_{2.5}$ predictions over stations in other clusters, the R^2 of the Just, Wright [12] method and Kloog, Koutrakis [41] method dropped to 0.73 and 0.68, respectively, indicating that the spatial smoothing method is more likely to be negatively affected by the spatial distribution of ground monitors than ours. Since our MI method does not rely on ground measurements, it is more robust in spatial inference.

CONCLUSIONS

In this study, we developed a multiple imputation model using satellite-retrieved cloud fraction, CMAQ-simulated AOD, and meteorological parameters to fill the gaps of MAIAC AOD. A two-stage statistical model was then used to predict ground $PM_{2.5}$ concentrations from the gap-filled AOD, meteorological parameters, and land use information. This method improved the coverage of $PM_{2.5}$ prediction by about two-fold per year and provided predictions with high accuracy at 1-km resolution. By including all the pixels of all days into model development, this method can correct the sampling bias in exposure assessment due to non-random missingness in AOD, especially in regions with large-scale seasonal missingness. Comparing with previously reported gap-filling methods, the MI method has the strength of not relying on ground $PM_{2.5}$ measurements, therefore allows the prediction of historical $PM_{2.5}$ levels prior to the establishment of regular ground monitoring networks. This study advanced our capabilities to integrate ground observations, satellite data, model simulations, and land cover information in $PM_{2.5}$ exposure modeling, and will support epidemiological studies on the air pollution related health burden in China as well as other regions in the world with limited air pollution monitoring.

ACKNOWLEDGMENTS

The work of Y. Liu and Q. Xiao is partially supported by the NASA Applied Sciences Program (Grant # NNX14AG01G, PI: Liu) and the Jet Propulsion Laboratory (Contract # 1363692, PI: Liu). The work of G. Geng is partially supported by the EPA STAR program (Grant # 83586901). The contents of the publication are solely the responsibility of the grantee and do not necessarily represent the official views of the U.S. EPA. Further, the U.S. EPA does not endorse the purchase of any commercial products or services mentioned in this publication. We thank Yixuan Zheng of Tsinghua University for providing CMAQ simulation results.

REFERENCES

1. Lim, S.S., et al., *A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010.* The lancet, 2013. **380**(9859): p. 2224-2260.
2. Lelieveld, J., et al., *The contribution of outdoor air pollution sources to premature mortality on a global scale.* Nature, 2015. **525**(7569): p. 367-371.
3. Kloog, I., et al., *Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal PM2. 5 exposures in the Mid-Atlantic states.* Environmental science & technology, 2012. **46**(21): p. 11913-11921.
4. Liu, Y., et al., *Estimating ground-level PM2. 5 in the eastern United States using satellite remote sensing.* Environmental science & technology, 2005. **39**(9): p. 3269-3278.
5. Ma, Z., et al., *Satellite-based spatiotemporal trends in PM2. 5 concentrations: China, 2004-2013.* Environmental Health Perspectives (Online), 2016. **124**(2): p. 184.
6. Tao, M., et al., *Satellite observation of regional haze pollution over the North China Plain.* Journal of Geophysical Research: Atmospheres, 2012. **117**(D12).
7. Van Donkelaar, A., et al., *Satellite-based estimates of ground-level fine particulate matter during extreme events: A case study of the Moscow fires in 2010.* Atmospheric Environment, 2011. **45**(34): p. 6225-6232.
8. Zheng, Y., et al., *Estimating ground-level PM 2.5 concentrations over three megalopolises in China using satellite-derived aerosol optical depth measurements.* Atmospheric Environment, 2016. **124**: p. 232-242.
9. Ma, X., et al., *Can MODIS AOD be employed to derive PM2. 5 in Beijing-Tianjin-Hebei over China?* Atmospheric Research, 2016. **181**: p. 250-256.
10. Li, R., et al., *Estimating ground-level pm 2.5 using fine-resolution satellite data in the megacity of Beijing, China.* Aerosol Air Qual. Res, 2015. **15**: p. 1347-1356.
11. Li, S., et al., *Retrieval of aerosol optical depth over bright targets in the urban areas of North China during winter.* Science China Earth Sciences, 2012. **55**(9): p. 1545-1553.
12. Just, A.C., et al., *Using high-resolution satellite aerosol optical depth to estimate daily PM2. 5 geographical distribution in Mexico City.* Environmental science & technology, 2015. **49**(14): p. 8576-8584.
13. Myhre, G., et al., *Aerosol-cloud interaction inferred from MODIS satellite data and global aerosol models.* Atmospheric Chemistry and Physics, 2007. **7**(12): p. 3081-3101.
14. Alam, K., et al., *Variability of aerosol optical depth and their impact on cloud properties in Pakistan.* Journal of Atmospheric and Solar-Terrestrial Physics, 2014. **107**: p. 104-112.
15. Yu, C., et al., *Statistical evaluation of the feasibility of satellite-retrieved cloud parameters as indicators of PM2. 5 levels.* Journal of Exposure Science and Environmental Epidemiology, 2015. **25**(5): p. 457-466.
16. Bey, I., et al., *Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation.* Journal of Geophysical Research: Atmospheres, 2001. **106**(D19): p. 23073-23095.

17. Byun, D. and K.L. Schere, *Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system.* Applied Mechanics Reviews, 2006. **59**(2): p. 51-77.
18. Stern, R., et al., *A model inter-comparison study focussing on episodes with elevated PM10 concentrations.* Atmospheric Environment, 2008. **42**(19): p. 4567-4588.
19. Appel, K.W., et al., *Evaluation of the community multiscale air quality (CMAQ) model version 4.5: sensitivities impacting model performance; part II—particulate matter.* Atmospheric Environment, 2008. **42**(24): p. 6057-6066.
20. Appel, K.W., et al., *Examination of the Community Multiscale Air Quality (CMAQ) model performance over the North American and European domains.* Atmospheric environment, 2012. **53**: p. 142-155.
21. Butland, B.K., et al., *Measurement error in time-series analysis: a simulation study comparing modelled and monitored data.* BMC medical research methodology, 2013. **13**(1): p. 136.
22. Quennehen, B., *Multi-model evaluation of short-lived pollutant distributions over East Asia during summer 2008.* Atmos. Chem. Phys, 2015. **1680**: p. 7324.
23. Friberg, M.D., et al., *Method for fusing observational data and chemical transport model simulations to estimate spatiotemporally resolved ambient air pollution.* Environmental science & technology, 2016. **50**(7): p. 3695-3705.
24. Punger, E.M. and J.J. West, *The effect of grid resolution on estimates of the burden of ozone and fine particulate matter on premature mortality in the USA.* Air Quality, Atmosphere & Health, 2013. **6**(3): p. 563-573.
25. Lyapustin, A., et al., *Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm.* Journal of Geophysical Research: Atmospheres, 2011. **116**(D3).
26. Lyapustin, A., et al., *Multiangle implementation of atmospheric correction (MAIAC): 1. Radiative transfer basis and look‐up tables.* Journal of Geophysical Research: Atmospheres, 2011. **116**(D3).
27. Emili, E., et al., *High spatial resolution aerosol retrieval with MAIAC: Application to mountain regions.* Journal of Geophysical Research: Atmospheres, 2011. **116**(D23).
28. Hu, X., et al., *10-year spatial and temporal trends of PM 2.5 concentrations in the southeastern US estimated using high-resolution satellite data.* Atmospheric Chemistry and Physics, 2014. **14**(12): p. 6301-6314.
29. Di, Q., et al., *Assessing PM2. 5 exposures with high spatiotemporal resolution across the continental United States.* Environmental science & technology, 2016. **50**(9): p. 4712-4721.
30. Strickland, M.J., et al., *Pediatric Emergency Visits and Short-Term Changes in PM2. 5 Concentrations in the US State of Georgia.* Environmental health perspectives, 2016. **124**(5): p. 690.
31. Kloog, I., et al., *Estimating daily PM 2.5 and PM 10 across the complex geo-climate region of Israel using MAIAC satellite-based AOD data.* Atmospheric Environment, 2015. **122**: p. 409-416.
32. Jinnagara Puttaswamy, S., et al., *Statistical data fusion of multi-sensor AOD over the Continental United States.* Geocarto International, 2014. **29**(1): p. 48-64.
33. Holben, B.N., et al., *AERONET—A federated instrument network and data archive for aerosol characterization.* Remote sensing of environment, 1998. **66**(1): p. 1-16.

34. Rohde, R.A. and R.A. Muller, *Air pollution in China: Mapping of concentrations and sources.* PloS one, 2015. **10**(8): p. e0135749.
35. Dobson, J.E., et al., *LandScan: a global population database for estimating populations at risk.* Photogrammetric engineering and remote sensing, 2000. **66**(7): p. 849-857.
36. Zhang, Y., et al., *Probing into regional O3 and PM pollution in the US, Part I. A 1-year CMAQ simulation and evaluation using surface and satellite data.* Journal of Geophysical Research, 2009. **114**: p. D22304.
37. Chameides, W., et al., *Correlation between model‐calculated anthropogenic aerosols and satellite‐derived cloud optical depths: Indication of indirect effect?* Journal of Geophysical Research: Atmospheres, 2002. **107**(D10).
38. Junger, W. and A.P. de Leon, *Imputation of missing data in time series for air pollutants.* Atmospheric Environment, 2015. **102**: p. 96-104.
39. Acock, A.C., *Working with missing values.* Journal of Marriage and family, 2005. **67**(4): p. 1012-1028.
40. Harrell Jr, F.E., *Hmisc: Harrell Miscellaneous.* R package version 3.17-4. https://CRAN.R-project.org/package=Hmisc, 2016.
41. Kloog, I., et al., *Assessing temporally and spatially resolved PM 2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements.* Atmospheric Environment, 2011. **45**(35): p. 6267-6275.
42. Hu, X., et al., *Estimating ground-level PM 2.5 concentrations in the southeastern U.S. using geographically weighted regression.* Environmental Research, 2013. **121**: p. 1-10.
43. Lv, B., et al., *Improving the Accuracy of Daily PM2.5 Distributions Derived from the Fusion of Ground-Level Measurements with Aerosol Optical Depth Observations, a Case Study in North China.* Environmental Science & Technology, 2016. **50**(9): p. 4752-4759.
44. Kloog, I., et al., *A new hybrid spatio-temporal model for estimating daily multi-year PM2.5 concentrations across northeastern USA using high resolution aerosol optical depth data.* Atmospheric Environment, 2014. **95**: p. 581-590.
45. Smirnov, A., et al., *Cloud-Screening and Quality Control Algorithms for the AERONET Database.* Remote Sensing of Environment, 2000. **73**(3): p. 337-349.
46. Kang, N., et al., *Correlation analysis between AOD and cloud parameters to study their relationship over China using MODIS data (2003–2013): impact on cloud formation and climate change.* Aerosol Air Qual. Res, 2015. **15**: p. 958-973.
47. Ma, Z., et al., *Satellite-derived high resolution PM2.5 concentrations in Yangtze River Delta Region of China using improved linear mixed effects model.* Atmospheric Environment, 2016. **133**: p. 156-164.

Table 1. 1 Summary statistics and coefficients of fixed effects in the LME model of year 2013 and 2014.

	Mean (Std)		Coefficient	
	2013	2014	2013	2014
$PM_{2.5}$ (μg/m^3)	73 (57)	61 (35)		
AOD	0.98 (0.44)	1.11 (0.57)	33.97	48.5
PBLH[a] (m)	1.19×10^3 (413)	1.18×10^3 (401)	-1.85	-1.12
SH[b] (kg/kg)	9.93×10^{-3} (6.17×10^{-3})	1.06×10^{-2} (5.67×10^{-3})	-9.30	-5.85
Temp[c] (K)	294 (10)	294 (8)	3.62	3.50
SWGDN[d] (W/m^2)	517 (188)	506 (190)	1.10×10^{-2}	5.18×10^{-3}
Wind[e] (m/s)	5.81 (3.22)	5.44 (3.00)	-2.05	-1.11
NDVI[f]	0.26 (0.13)	0.28 (0.11)	-13.8	-9.74
PRECTOT[g] (kg/m^2s^2)	3.23 (10.08)	4.94 (12.69)	-6.67×10^{-2}	-2.93×10^{-2}
AOD^2			5.19	-9.56
AOD×PBLH			1.56	0.49
$SWGDN^2$			3.00×10-5	-1.47×10-5

a Planetary boundary layer height

b Specific humidity at 2 m

c Temperature at 2 m

d Surface incident shortwave flux

e Wind speed under planetary boundary layer

f Normalized difference vegetation index

g Total precipitation during the previous day

Table 1. 2 Performance of different gap-filling methods in a case study in 2014.

Method	MI	Lv et al.	Just et al.	Kloog et al.
Coverage (%)	100	94	94	94
R^2 of AOD evaluation	0.44	0.18		
R^2 of $PM_{2.5}$ evaluation	0.78		0.84	0.79

Figure 1. 1 Study region with a 50-km buffer, showing air quality monitoring stations and AERONET stations in the modeling region.

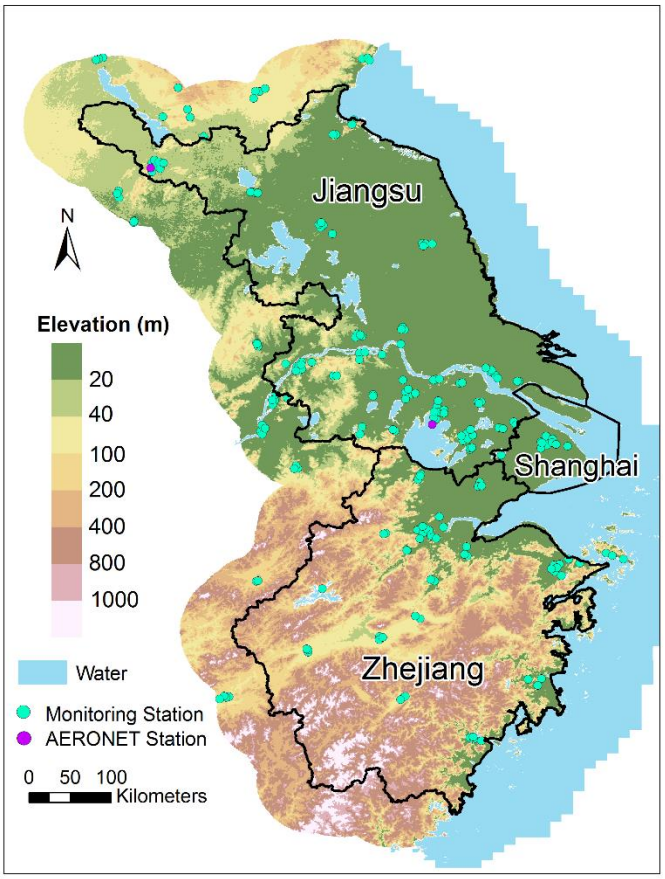


Figure 1. 2 The workflow of multiple imputation (light grey), first stage linear mixed-effects (LME) model (grey) and second stage generalized additive model (GAM) (dark grey).

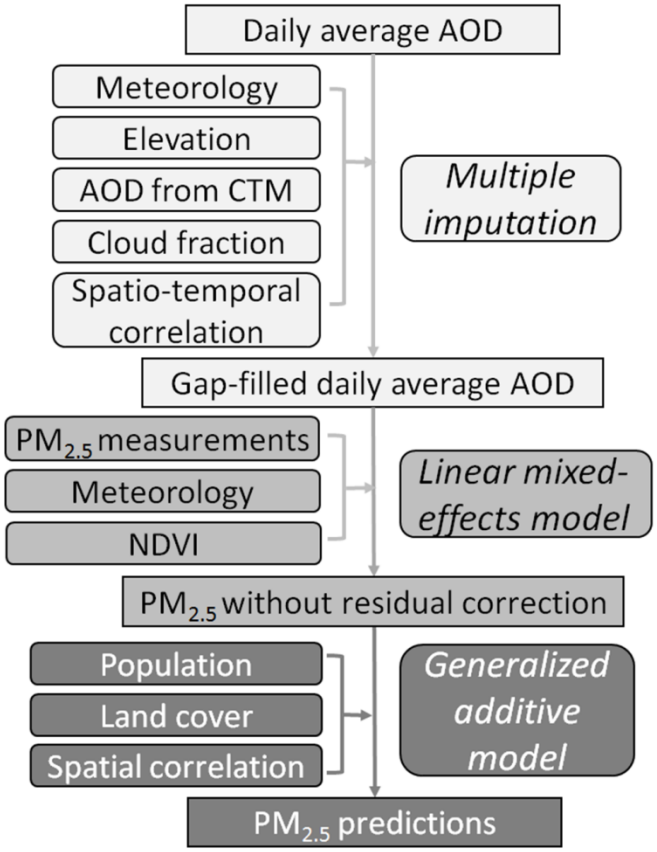


Figure 1. 3 Annual mean MAIAC AOD coverage (left) and summer (June to August) seasonal average MAIAC AOD coverage (right) over Yangtze River Delta during 2013-2014.

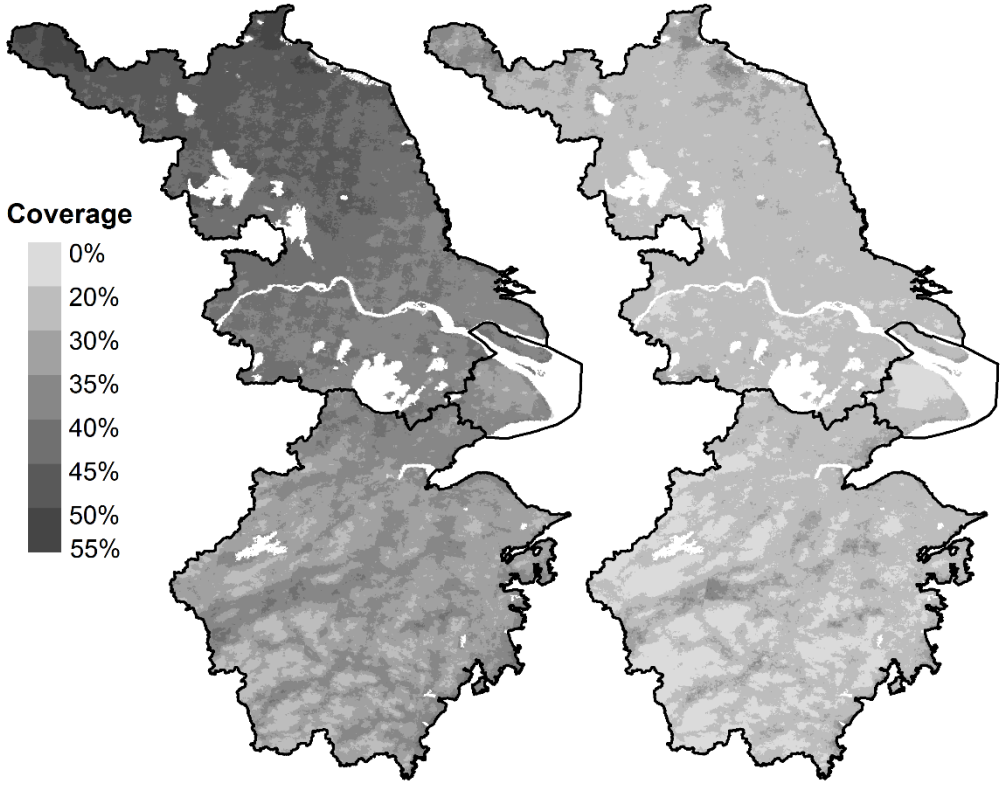


Figure 1. 4 Annual average AOD before (left) and after (right) imputation during 2013-2014.

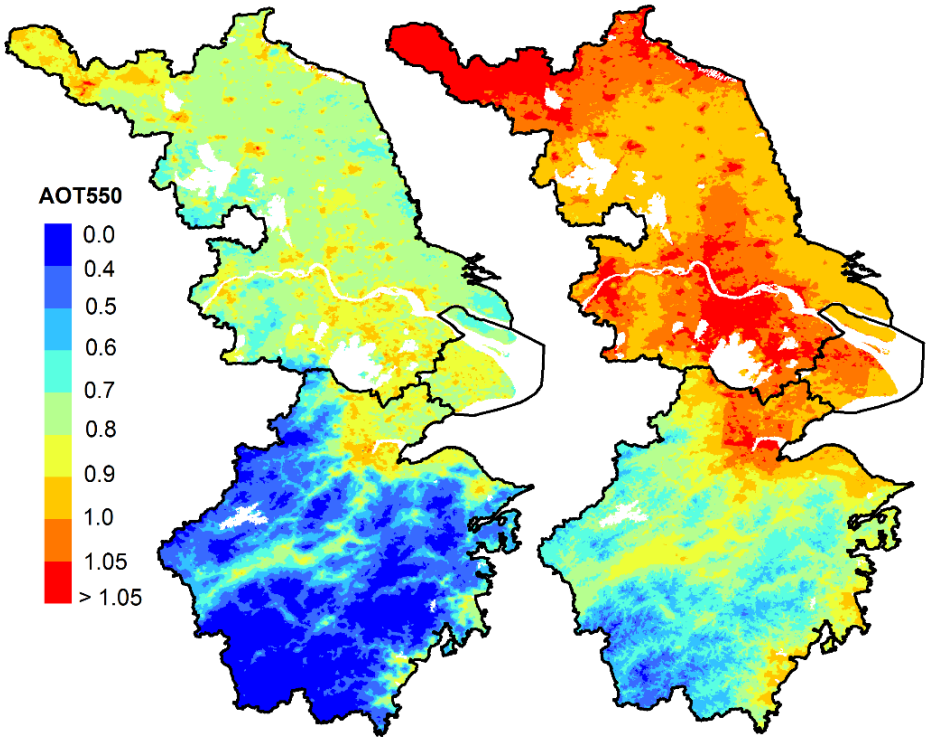


Figure 1. 5 Ten-fold cross-validation results of the two-stage prediction model.

The blue solid line shows the linear regression between $PM_{2.5}$ measurements and $PM_{2.5}$ predictions. The red dash line is the one-one line.

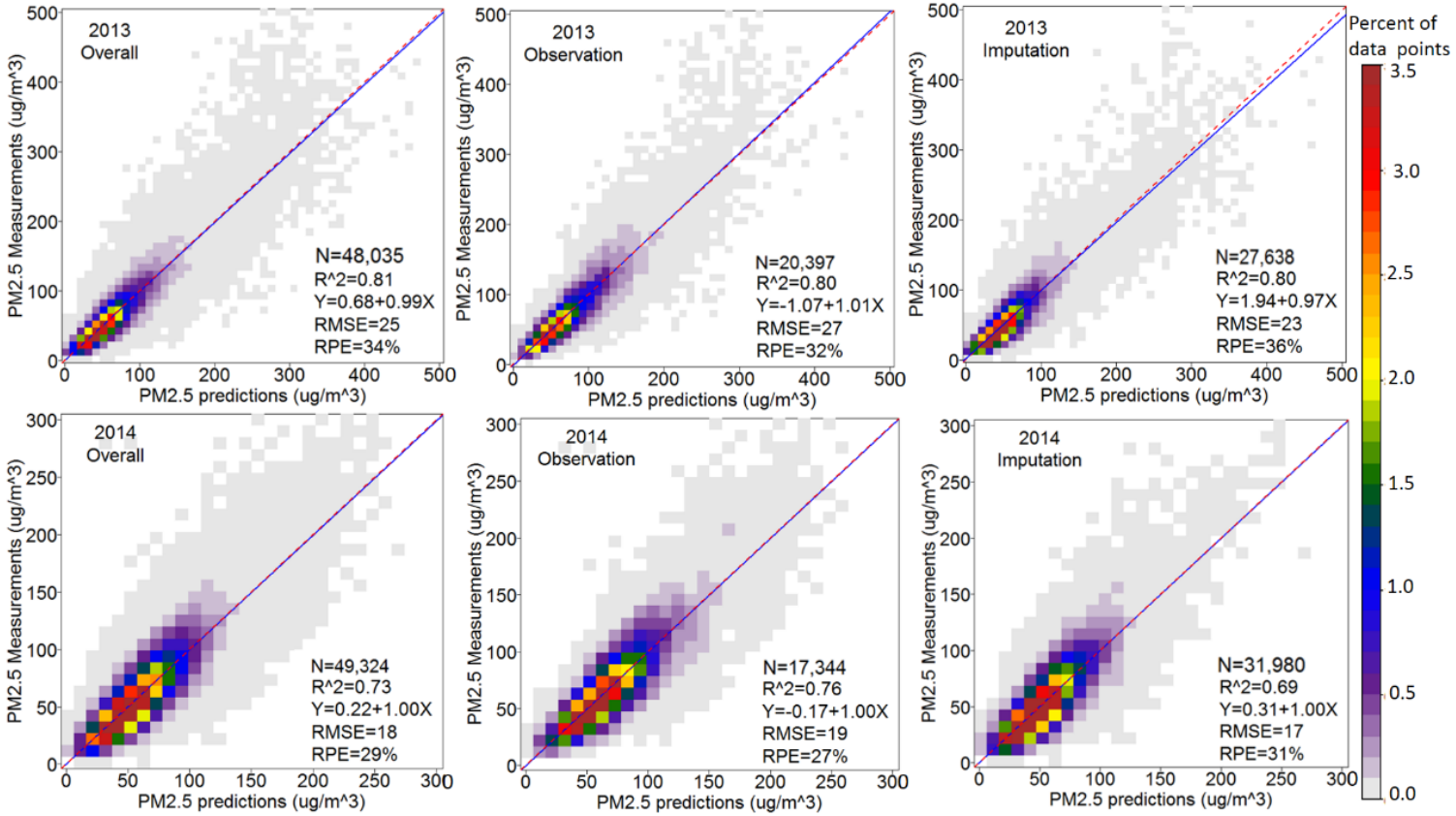


Figure 1. 6 Annual average $PM_{2.5}$ predictions.

$PM_{2.5}$ predictions over the buffer region were not shown. A: annual average $PM_{2.5}$ predictions in 2013. B: annual average $PM_{2.5}$ predictions in 2014.C: zoom in map of annual average $PM_{2.5}$ predictions over Nanjing. D: satellite photo of Nanjing. Map data: Google, Landsat/Copernicus.

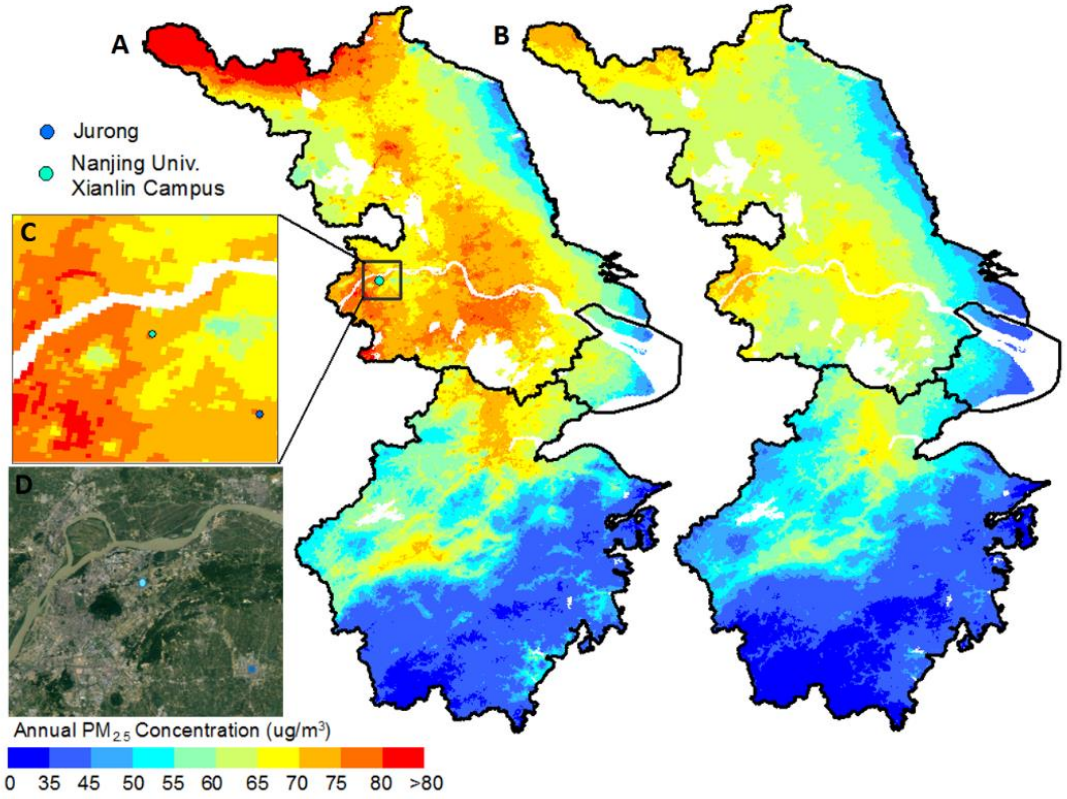
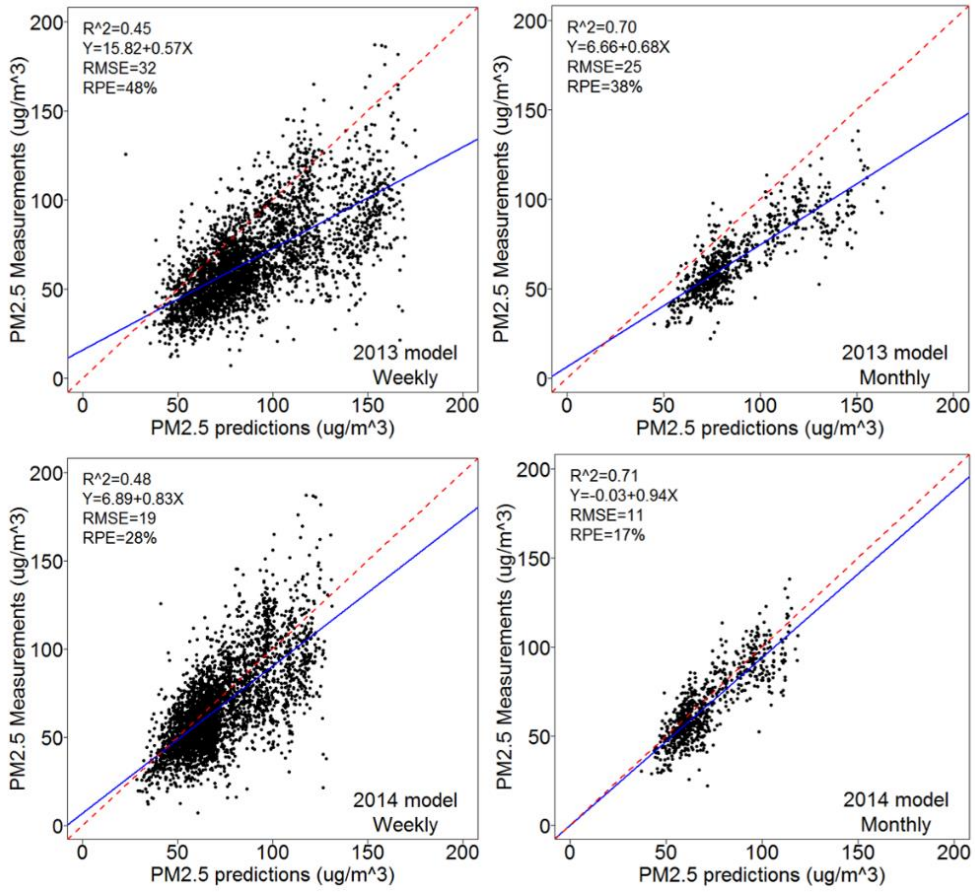


Figure 1. 7 Results of predicting 2015 weekly and monthly $PM_{2.5}$ levels with models fitted from data of year 2013 and 2014.



SUPPLEMENTARY MATERIALS

Figure 1.S 1 Map of Wuxi and Xuzhou with $PM_{2.5}$ monitoring stations.

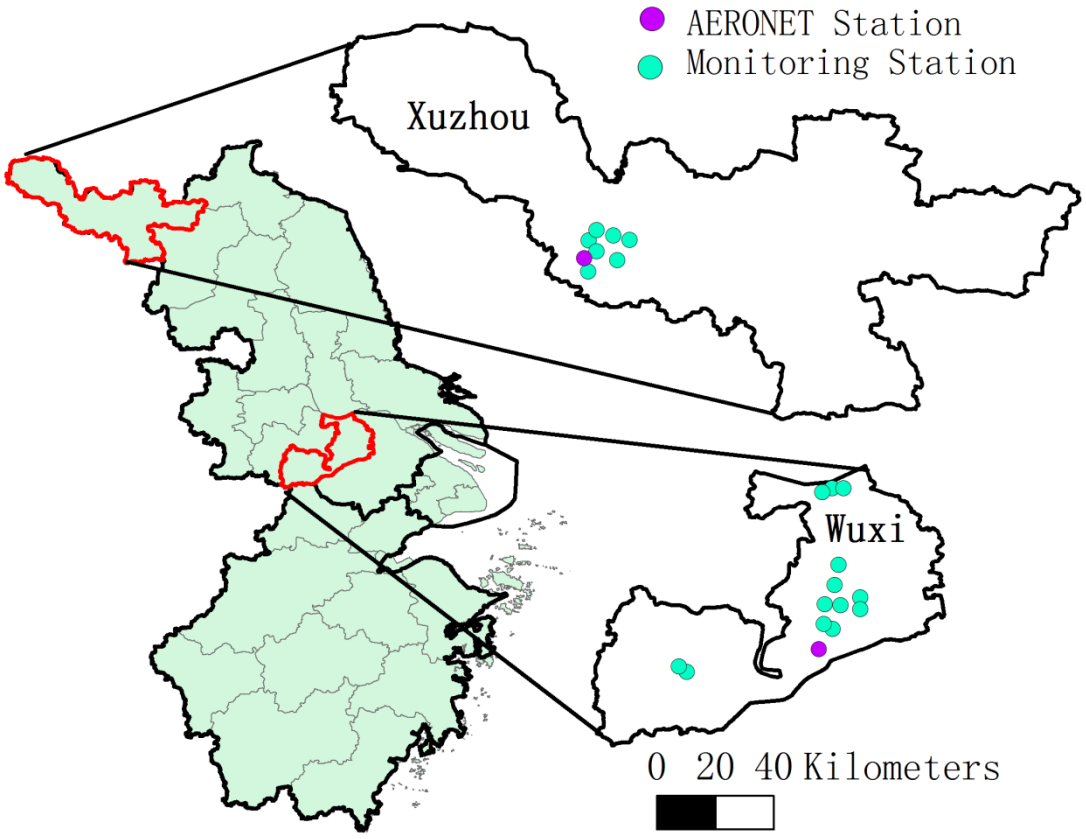


Figure 1.S 2 Comparing daily AERONET AOD (during 9:00-3:00 local time) with gap-filled MAIAC AOD, daily AERONET AOD with observational MAIAC AOD and daily AERONET AOD with imputed AOD.

The blue solid line shows the linear regression between MAIAC AOD and AERONET AOD. The red dash line is the one-one line.

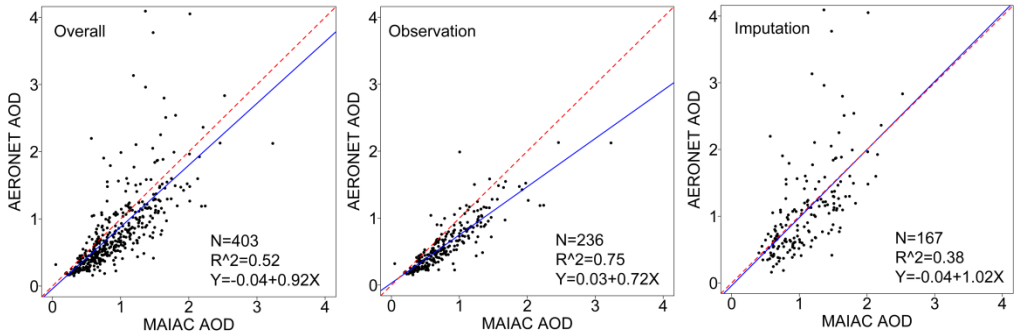
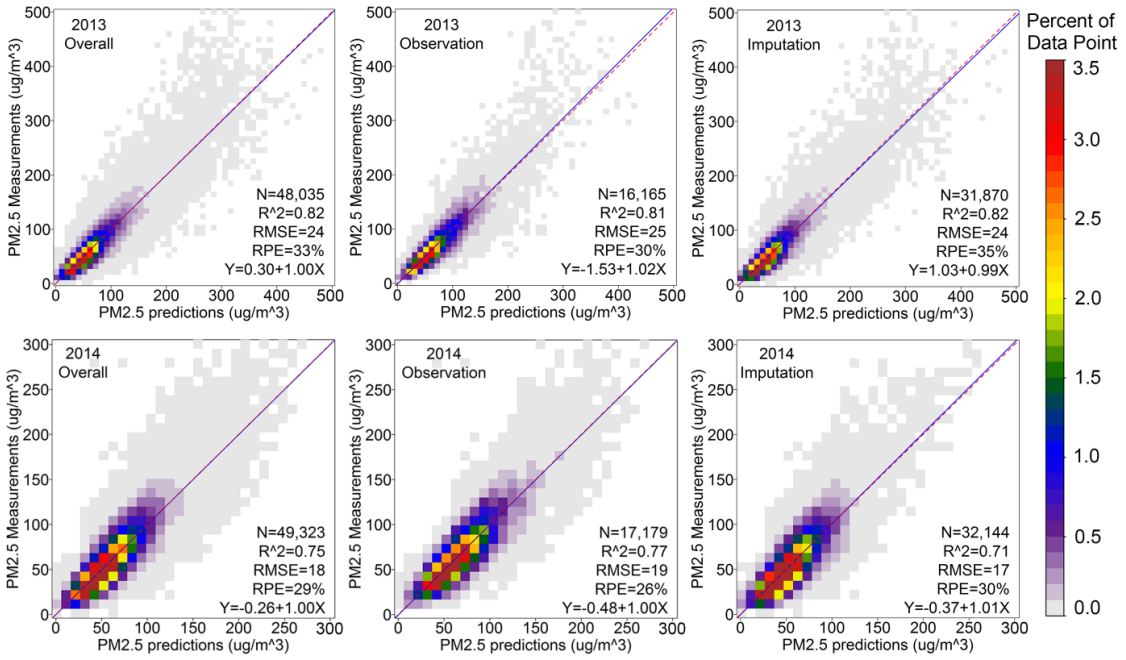


Figure 1.S 3 Model fitting results of the two-stage prediction model.



Chapter 2

Associations between birth outcomes and maternal $PM_{2.5}$ exposure in Shanghai: a comparison of three exposure assessment approaches

Qingyang Xiao, Hanyi Chen, Matthew J. Strickland, Haidong Kan, Howard H. Chang, Mitchel Klein, Chen Yang, Xia Meng, Yang Liu

ABSTRACT

Background: Few studies have estimated effects of maternal $PM_{2.5}$ exposure on birth outcomes in China due to the lack of historical air pollution data.

Objectives: We estimated the associations between maternal $PM_{2.5}$ exposure and birth outcomes including birth weight and preterm birth using gap-filled satellite estimates in Shanghai, China.

Methods: We obtained birth registration records of 132 783 singleton live births during 2011-2014 in Shanghai. $PM_{2.5}$ exposures were assessed from satellite-derived estimates or central-site measurements. Linear and logistic regressions were used to estimate associations with term birth weight and term LBW, respectively. Logistic and discrete-time survival models were used to estimate associations with preterm birth. Effect modifications by maternal age and parental education levels were investigated.

Results: A 10 $\mu g/m^3$ increase in gap-filled satellite-based whole-pregnancy $PM_{2.5}$ exposure was associated with a -12.85 g (95% CI: -18.44, -7.27) change in term birth weight, increased risk of preterm birth (OR 1.27, 95% CI: 1.20, 1.36), and increased risk of term LBW (OR 1.22, 95% CI: 1.06, 1.41). Sensitivity analyses during 2013-2014, when ground $PM_{2.5}$ measurements were available, showed that the estimated health effects using gap-filled satellite $PM_{2.5}$ were higher than using satellite $PM_{2.5}$ without accounting for missingness. The estimated health effects using gap-filled satellite $PM_{2.5}$ had similar magnitudes to those using central-site measurements, but with tighter confidence intervals.

Conclusions: The magnitude of associations between maternal $PM_{2.5}$ exposure and adverse birth outcomes in Shanghai was higher than previous findings. One reason could be reduced exposure error of the gap-filled high-resolution satellite $PM_{2.5}$ estimates.

INTRODUCTION

Preterm birth and low birth weight have been widely documented as significant predictors of infant mortality and have negative long-term effects in adulthood [1-3]. Liu, Oza [4] estimated that in 2013, preterm birth ranked as the first cause of death before age 5 and was responsible for 15.4% (0.965 million) of deaths before age 5 in the world. Adverse birth outcomes in association with maternal exposure to $PM_{2.5}$ (fine particulate matter with an aerodynamic diameter of 2.5 μm or less) have been studied in various populations [5, 6]. While results from recent meta-analyses support the link between maternal $PM_{2.5}$ exposure and adverse birth outcomes, substantial heterogeneity in health effect estimates exists among different studies [7, 8]. This heterogeneity is partly due to differences in exposure assessment methods and the authors reported that studies assessing individual-level exposures tended to report stronger associations relative to studies assessing regional-level exposures. In addition, similar to many other health endpoints reported in the literature, the overwhelming majority of the included studies in these meta-analyses were conducted in the U.S. where $PM_{2.5}$ levels are relatively low. Studies in highly polluted regions such as China can further elucidate the magnitude of $PM_{2.5}$-associated health effects and provide crucial information on the shape of concentration-response curve at high exposure levels. However, ground measurements of $PM_{2.5}$ levels are often very sparse or nonexistent in most part of the developing world. For countries where its air quality monitoring network was established recently, lack of long-term measurements remains an obstacle to studying the association between adverse birth outcomes and exposure to $PM_{2.5}$. Additionally, measurements from ground central monitors have limited spatial representativeness. Previous studies used specific buffers, ranging from 6.4 km to 50 km in radius, around monitoring stations to select study populations and assign exposure, with the intent of reducing exposure error (Chang et al. 2011; Darrow et al. 2011; Hyder et al. 2014). However, this method reduces sample size, and an optimal cutoff distance is difficult to determine.

To assess historical air pollution levels and characterize local-scale variability in air pollution, satellite-retrieved aerosol optical depth (AOD) has been used in health effects studies during the past decade [9]. Polar-orbiting satellites have global coverage, long data records, and high spatiotemporal resolution, but missingness in satellite data has raised concerns regarding its use in epidemiological studies. Annually, 30 to 70% satellite retrievals can be missing due to cloud cover and high surface reflectance [10]. Unfortunately, situations that lead to failed satellite retrievals often influence the production and deposition of $PM_{2.5}$, e.g. cloud cover leads to reduced photochemical reactions. Thus, using satellite predictions without accounting for the non-random missingness may result in exposure misclassification. Strickland, Hao [11] evaluated the influence of missing satellite-derived $PM_{2.5}$ predictions on the association between short-term $PM_{2.5}$ exposure and pediatric emergency department visits in Georgia, US. They reported that, in general, a large proportion of missing satellite predictions tended to overestimate regional average $PM_{2.5}$ exposure compared with ground measurements and led to lower health association estimates. To date, studies on the influence of missing satellite data on $PM_{2.5}$ longer-term exposure assessment are very limited.

We developed a gap-filling method that provided full-coverage daily $PM_{2.5}$ predictions at 1-km resolution using the Multi-Angle Implementation of Atmospheric Correction (MAIAC) aerosol product [12]. In this study, we analyzed the associations between birth outcomes (birth weight and preterm birth) and maternal $PM_{2.5}$ exposure in Shanghai, China, during 2011-2014, using three exposure metrics: satellite predictions with missingness, gap-filled satellite predictions with complete coverage, and measurements from ground central monitors. We reported that exposure errors can arise when satellite predictions without accounting for missing data were used in exposure assessment.

METHODS

Data and Outcome Assessments

Birth registration data for live births between January 1st, 2011 and December 31st, 2014 were obtained from Pudong New Area Centers for Disease Control and Prevention (CDC) (n=173 403). Data of all births born in Pudong New Area were collected by health facilities and reported to the Pudong New Area CDC. Although born in Pudong New Area, some of these births had maternal residential address in other distracts of Shanghai (Figure 2. A1). Shanghai is located on the east coast of China (Figure 2. A1) and is one of the largest cities in the world with more than 24 million residents. Benefited from the establishment of a special economic zone in 1993, Pudong New Area as well as Shanghai has become one the most economically developed regions in China as well as in East Asia.

The gestational age was calculated from the last menstrual period and the birth weight was measured at the time of birth using standard digital scales. The maternal residential address was geo-coded for exposure assignment. Maternal residential address outside Shanghai or with failed geo-coding, mainly due to incomplete address, were excluded (13%). Singleton births without congenital anomalies (96%) and with clinically estimated gestational age between 27 and 42 weeks were selected to ensure that we followed the complete first and second trimester of all births. We further excluded births with maternal age younger than 15 years or older than 44 years (0.04%). In birth cohort studies using birth records ascertained based on birth dates, the study population tends to include longer gestations at the start of the study period and shorter gestations at the end of the study period [13]. To avoid this issue, we included births with the estimated conception date after June 26, 2010 (27 weeks before January 1st, 2011) and before March 12, 2014 (42 weeks before December 31st, 2014) (n=133 120). Births with missing variables were excluded (0.2%). Thus, 132 783 births were analyzed in this study. Preterm birth was defined as a

birth with less than 37 weeks but at least 27 weeks of gestation. Term low birth weight (LBW) was defined as a full term birth (≥ 37 weeks of gestation) with a birth weight less than 2500 g.

Exposure Assessment

To analyze the influence of different exposure assessment approaches on estimated health effects, we considered three exposure metrics from satellite data and ground measurements: 1) daily $PM_{2.5}$ predictions from MAIAC AOD without accounting for missingness, 2) daily $PM_{2.5}$ predictions from gap-filled MAIAC AOD with complete coverage, and 3) daily average $PM_{2.5}$ measurements from ten monitoring stations. The satellite derived $PM_{2.5}$ predictions were estimated during 2010-2014. $PM_{2.5}$ measurements from central station were only available since 2013.

Details of the multiple imputation gap-filling method and the two-stage $PM_{2.5}$ concentration prediction model can be found elsewhere [12] and a brief description is provided here. First, we brought together the emerging MAIAC satellite aerosol optical depth (AOD) retrieval, the Moderate Resolution Imaging Spectroradiometer (MODIS) cloud fraction [14], the Community Multi-scale Air Quality (CMAQ) AOD simulations [15] and elevation data by multiple imputation to fill missing satellite AOD. After gap-filling, the coverage of satellite retrieval increased from below 40% to 100%. Then, for daily average $PM_{2.5}$ concentration prediction, we fitted a first stage linear mixed effects model driven by gap-filled AOD and meteorology variables, and a second stage generalized additive model (GAM) driven by land use information. This method provided $PM_{2.5}$ predictions at 1 km resolution with complete coverage in space and time. In Shanghai, the model cross validation R^2 (root mean square error) between daily model predictions and ground measurements was 0.74 (22 $\mu g/m^3$) [12]. When aggregating during a certain exposure window, the gap-filled predictions had smaller bias compared to ground measurements. For example, the gap-filled satellite predictions better estimated monthly average $PM_{2.5}$ concentrations (10-fold cross-validation R^2=0.92 and relative prediction error=9%) than

satellite predictions without accounting for missing data (R^2=0.84 and relative prediction error=14%) in 2014. Trimester-specific and whole-pregnancy $PM_{2.5}$ exposures were assigned by maternal residential address of each birth record and averaged from daily satellite derived $PM_{2.5}$ predictions, with and without gap-filling, across each exposure window based on the estimated gestation date.

Hourly $PM_{2.5}$ measurements from 10 air quality monitoring stations in Shanghai were published by the China National Environmental Monitoring Center (CNEMC, http://www.cnemc.cn/), and were downloaded from PM25.in (http://pm25.in/), a direct mirror of data from CNEMC. We removed repeated identical measurements for at least three continuous hours, assuming that such repetition was due to instrument malfunction (Rohde and Muller 2015). Daily average concentrations, calculated from hourly concentrations during 0:00-23:59 local time with at least 18 hourly measurements, were included for exposure assessment. Since more than 95% of the study population reside within 25 km of at least one monitoring station (Figure 2. 1) and these stations were clustered with high temporal correlations ranging between 0.88 and 0.99 (Figure 2. A.2), we used the daily regional average concentrations to assess exposure across each exposure window.

Meteorological variables (temperature, humidity, surface pressure) were obtained from the Goddard Earth Observing System Data Assimilation System GEOS-5 Forward Processing (GEOS 5-FP) [16].

Statistical Models

Health effects on birth weight and term LBW in association with maternal $PM_{2.5}$ exposure were estimated by linear regression and logistic regression among full term births, respectively. Logistic regression models were fitted to estimate associations between preterm birth and maternal $PM_{2.5}$ exposure during the first (from conception date through gestational week 13) and second (gestational week 14-26) trimester. Discrete-time survival models were fitted to estimate

associations during the third trimester (gestational week 27- date of birth) and the entire pregnancy [17], because the length of third trimester exposure and entire pregnancy exposure are affected by the birth date. Specifically, the discrete-time survival model assumes each birth was no longer at-risk (censored) of being preterm at week 37. The discrete-time survival model can be expressed as Equation 1 [18, 19].

$$LogitP(Y_{it}) = \beta_0 + \beta_1 E_{it} + \delta X_i + \gamma GestWeek_t \quad \text{Equation 1}$$

where Y_{it} indicates for pregnancy i, whether a birth occurred during gestational week t; E_{it} indicates the average pollution level for pregnancy i from gestational week 27 to gestational week t for associations with the third trimester exposure or from gestational week 1 to gestational week t for associations with the entire pregnancy exposure, respectively; X_i indicates covariates that were controlled in this study and $GestWeek_t$ indicates the gestational week.

We controlled the following covariates in all models: parental education levels (high school or lower, college, graduate school), maternal age (continuous), parity (1, 2, >2), birth location (hospital, maternal health service center, others), infant sex, average temperature and average surface pressure during the corresponding exposure period (continuous), season for conception (a categorical variable with four levels), and long-term temporal trend (a cubic spline with one knot per year) [18, 20]. Paternal age and average relative humidity were included in the initial regression, but they resulted in no meaningful changes in the point estimates and were excluded from the final model. Analyses were performed separately using each of the three exposure metrics described above. We used 10 $\mu g/m^3$ as the exposure increment to benefit the comparison of health association estimates across exposure metrics. Since $PM_{2.5}$ measurements are only available since 2013, we only included births with an estimated gestation date after January 1st 2013 when comparing across exposure metrics. We also reported health association estimates per interquartile range (IQR) change in $PM_{2.5}$ exposure to facilitate the comparison across exposure windows. IQRs were calculated from each trimester-specific/pregnancy exposures of all births.

To investigate the potential effect modification of maternal age, we stratified the study population by mothers younger than 35 years or not [21]. We also investigated the potential effect modification of parental social economic status by stratifying the study population by maternal and paternal education levels (college and higher or not), separately. To investigate the potential confounding due to unobserved spatial trends, we conducted a sensitivity analysis using generalized additive models with the spatial trends being controlled by a thin-plate spline of longitude and latitude of the centroid of each grid cell.

RESULTS

The annual average $PM_{2.5}$ concentrations from satellite predictions and central-site measurements in 2014 are shown in Figure 2. 1. $PM_{2.5}$ concentrations decreased from west to east. Characteristics of the study population are shown in Table 2. 1. The mean birth weight among term births was 3389 g, with a standard deviation of 403 g. The preterm birth rate in Shanghai during 2011-2014 was 4.41% and the term LBW rate was 0.95%. Previous studies reported higher preterm birth rate in China (4.8%) during 2004-2008 [22] and in one hospital in Shanghai (6.8%) during 2010-2012 [23]. However, Xue, Shen [23] included multiple pregnancies (17.7%) and births with fetal anomalies (0.8%) in their study. Multiple pregnancies have a significantly higher preterm birth rate than singletons [24]. Another study reported the preterm birth rate being 7.4% and term LBW rate being 2.06% during 2012-2014 [25] in China; however, this study also included multiple pregnancies (3.3%) in the analysis and reported that multiple pregnancies had a higher risk of term LBW than singletons (OR = 21.9, 95% CI: 20.9, 22.9). The LBW rate in developed regions of China, e.g. Shanghai, has been reported to be lower than the national average [26]. Additionally, our data included births from hospitals as well as births from maternal health service centers and other places. Consistent with previous findings, we noticed that births in hospital (preterm birth rate=4.81% and term LBW rate 1.09%) had a higher rate of adverse birth outcomes than births in other places (preterm birth rate=3.64% and term LBW rate 0.70%)

[25]. The male/female ratio was 1.13 and 96.8% of mothers are ethnic Han. Shanghai is one of the most developed regions in China and more than half of parents have a Bachelor's degree or higher. The births with failed geocoding of maternal address had similar preterm birth rate (4.6%), term LBW rate (0.95%), and birth weight distribution among term births (mean 3399, standard deviation 414). However, fewer mothers with failed geocoding had a Bachelor's degree or higher (42.2%) relative to mothers with successful geocoding (52.5%).

Table 2. 2 shows the characteristics of the three $PM_{2.5}$ exposure metrics. The $PM_{2.5}$ exposure derived from satellite data with missingness had higher average values (e.g., 72 $\mu g/m^3$ during the entire pregnancy) than exposure from gap-filled satellite predictions (60 $\mu g/m^3$) and central-site measurements (58 $\mu g/m^3$). One explanation of the lower average $PM_{2.5}$ predictions after gap-filling is that cloud cover leads to reduced photochemical reaction and precipitation removes $PM_{2.5}$ from the atmosphere, leading to decreased $PM_{2.5}$ dry mass concentration [27]. The $PM_{2.5}$ exposure assessed from gap-filled satellite predictions and from central-site measurements were highly correlated, with the Pearson's correlation coefficients of exposures during the first, second, third trimester and whole pregnancy being 0.96, 0.97, 0.97, and 0.83, respectively. The relatively low correlation coefficients of exposures during pregnancy is because when averaging during a longer time (pregnancy), the spatial variations became more important and the 1-km satellite predictions had larger footprints than point measurements from ground monitors. The correlation between exposure estimated from satellite data without accounting for missingness and from ground measurements was weaker (Table 2. A1).

Figure 2. 2 presents the adjusted health association estimates for term birth weight, preterm birth, and term LBW in relation to trimester-specific and pregnancy maternal $PM_{2.5}$ exposures for births during 2011-2014, using gap-filled satellite predictions. $PM_{2.5}$ exposures during all time windows were associated with decreased birth weight in term births. The associations per 10 $\mu g/m^3$ increase in $PM_{2.5}$ exposure during entire pregnancy were a -12.85 g (95% CI: -18.44, -7.27)

change in birth weight and increased risk of term LBW (OR = 1.22, 95% CI: 1.06, 1.41). Magnitudes of associations with term birth weight were higher for exposures during the first (-4.66 g, 95% CI: -8.15, -1.16 per 10 μg/m^3 increase in $PM_{2.5}$ exposure) and the third trimester (-4.47 g, 95% CI: -8.01, -0.93) compared with exposures during the second trimester (-2.55g, 95% CI: -5.99, 0.89). Regarding preterm birth, exposures during the first trimester were observed with a higher risk (OR= 1.15, 95% CI: 1.10, 1.20 per 10 μg/m^3 increase in $PM_{2.5}$ exposure) than exposures during the second (OR= 1.08, 95% CI: 1.03, 1.12) and the third trimester (OR= 1.06, 95% CI: 1.03, 1.08). Results using exposure window-specific IQRs as the exposure increment are shown in Figure 2. B1. Effects of exposure during the whole pregnancy had larger magnitudes with term birth weight (-13.60 g, 95% CI: -19.50, -7.69 per IQR increase in $PM_{2.5}$ exposure) and preterm birth (OR=1.40, 95% CI: 1.28, 1.52) relative to effects of trimester-specific exposures. The third trimester exposure showed the highest risk of term LBW (OR= 1.30, 95% CI: 1.03, 1.63 per IQR) relative to exposures during other time windows.

The comparison results across three exposure metrics for births with estimated gestation date in 2013-2014 are shown in Figure 2. 3. Non-random missingness in satellite predictions decreased the accuracy of exposure assessment. For both birth weight and preterm birth, estimated effects using exposure from satellite data without accounting for missingness were weaker than estimated effects using exposure from gap-filled satellite data and ground measurements. Estimated effects using ground-based exposure had similar magnitude, but wider confidence intervals, sometimes including the null, than those using gap-filled satellite-based exposure. For example, the estimated ORs of preterm birth in association with the second trimester exposures assessed with central-site measurements and gap-filled satellite predictions were 1.06 (95% CI 0.96, 1.18) and 1.10 (95% CI: 1.03, 1.18), respectively.

Figure 2. 4 shows adjusted health effect estimates respectively stratified by maternal age and maternal education level, using exposures based on gap-filled satellite predictions. In pregnant

women older than 35 years, we observed stronger associations between $PM_{2.5}$ exposure and adverse health outcomes. Due to the small proportion of births with mother older than 35 years, this group had a much wider 95% CI than the > 35 year-old group. Among women with college or higher education, the estimated risk of preterm birth (OR=1.19, 95% CI: 1.08, 1.32 per 10 $\mu g/m^3$ increase in entire pregnancy $PM_{2.5}$ exposure) were lower; however, the estimated effects of $PM_{2.5}$ exposure on birth weight (-45.57 g, 95% CI: -54.31, - 36.83) were higher. For mothers without college education, the estimated OR for preterm birth was 1.32 (95% CI: 1.22, 1.43) and the estimated decrease in birth weight was -26.01 g (95% CI: - 34.44, 17.59). We also stratified the population by paternal education level and had parallel findings (Figure 2. B2).

When controlling for spatial trends, the estimated trimester-specific associations with adverse birth outcomes became weaker while the association for whole pregnancy remain statistically significant (Table 2. 3), probably because the spatial variations in $PM_{2.5}$ were surrogated by the flexible spatial spline smoother. As seen in Figure 2. B3, the GAM-estimated thin-plate spline that related preterm births and $PM_{2.5}$ exposure showed a similar pattern to the spatial distribution of $PM_{2.5}$ concentrations in Shanghai, with higher values in the northwest. Due to this additional control of spatial trends, the confidence interval of most health association estimates expanded to contain the null value.

DISCUSSION

In this study, we observed associations between maternal $PM_{2.5}$ exposure during all exposure windows and adverse birth outcomes, including decreased birth weight, increased risk of term LBW, and increased risk of preterm birth in a highly polluted region. Exposure assessment approaches affected the estimated health effects and satellite based exposures without accounting for missing data led to underestimate of health effects. Maternal age and parental education levels appeared to modify the associations between maternal $PM_{2.5}$ exposure and birth outcomes.

The exposure levels in Shanghai are much higher than other areas that have been studied previously. Research in the U.S. reported mean $PM_{2.5}$ exposure during the entire pregnancy ranging between 9.9 $\mu g/m^3$ in Florida and 18.7 $\mu g/m^3$ in California [8]. In our study, the mean $PM_{2.5}$ exposure during the whole pregnancy was 60 $\mu g/m^3$, estimated from gap-filled satellite predictions. We reported associations with higher magnitude between $PM_{2.5}$ exposure and birth outcomes than previous studies. Previous meta-analyses reported combined estimate of OR for preterm birth per 10 $\mu g/m^3$ increase in $PM_{2.5}$ exposure during the whole pregnancy being 1.13 (95% CI: 1.03, 1.24) [7] and 1.02 (95% CI: 0.93, 1.12) [6]. Dadvand, Parker [28] collected data from 14 centers (nine countries) and reported the combined OR for term LBW being 1.10 (95% CI: 1.03, 1.18) per 10 $\mu g/m^3$ increase in $PM_{2.5}$ exposure during entire pregnancy. We reported the OR for preterm birth and term LBW being 1.27 (95% CI: 1.20, 1.36) and 1.22 (95% CI: 1.06, 1.41) in association with 10 $\mu g/m^3$ increase in $PM_{2.5}$ exposure during entire pregnancy. This difference could be partly due to different exposure assessment methods, i.e., most previous studies used central-site measurements for exposure assessment while we used high-resolution satellite predictions in this study. Since most monitoring sites are located in urban centers, their measurements may overestimate individual exposure and attenuate health effect estimates. We found two studies that estimated health effects of $PM_{2.5}$ exposure on birth outcomes in China. Qian, Liang [29] reported the OR of preterm birth was 1.03 (95% CI: 1.02, 1.05) per 5 $\mu g/m^3$ increase in $PM_{2.5}$ exposure during the entire pregnancy in Wuhan, China. Fleischer, Merialdi [22] analyzed data of the World Health Organization Global Survey (WHOGS) and reported that the OR of preterm birth and LBW were 1.11 (95% CI: 1.04, 1.17) and 1.07 (95% CI: 1.01, 1.14) per 10 $\mu g/m^3$ increase in $PM_{2.5}$ exposure during the entire pregnancy in China, separately. These two studies assessed exposure from two ground monitoring stations [29] or from seasonal adjusted long-term (2001-2006) average $PM_{2.5}$ predictions from satellite [22]. By employing high-resolution satellite predictions for exposure assessment, we reduced potential exposure misclassification in this study.

We observed that satellite-based $PM_{2.5}$ predictions without accounting for non-random missingness overestimated gestational $PM_{2.5}$ exposure compared to ground measurements [12], thus led to attenuation in health effect estimates. This finding is consistent with previous studies [11, 30]. As discussed in previous research, in studies using daily exposures, the missing daily air pollution estimates result to a smaller study population, and gap-filling intends to increase the precision of estimated effects. In studies using long-term average exposures, such as this study, the missing daily air pollution estimates did not decrease the population size, rather increases the exposure error since the observed daily pollution levels are systematically different from the missing daily pollution levels. Thus, gap-filling is needed to increase the accuracy of exposure assessments. In our study, health associations estimated using satellite-based exposures had similar magnitudes but tighter confidence intervals compared to health associations estimated using ground-based exposures. Although central-site measurements have high accuracy, they ignored the local-scale spatial variability in $PM_{2.5}$ exposure, leading to more Berkson error and wider confidence intervals. Employing high-resolution complete-coverage satellite data may be able to improve accuracy of exposure assessments and benefit the health effect estimates.

In this study, we examined potential effect modification and found higher estimated risk of $PM_{2.5}$ exposure on adverse birth outcomes among pregnant women older than 35 years. We also found that pregnant women with higher education level had a lower estimated risk of preterm birth, but a larger decrease in birth weight in association with $PM_{2.5}$ exposure. Findings from previous studies on the effect modification of maternal education level were not consistent [18, 31-34]. Pregnant women with higher education level may be more vulnerable to air pollution-associated decrease in birth weight due to less exposure to competing risks, e.g. smoking and alcohol consumption.

In Shanghai, we found high correlation coefficients between daily regional average $PM_{2.5}$ and other air pollutants: CO (0.87), NO_2 (0.75), and SO_2 (0.76). Thus, a multipollutant model can

result in unstable estimates due to the high collinearity and we did not fit such a model. Our estimated $PM_{2.5}$ associations may in fact represent a broader pollutant mixture in our study.

There are 13% of births with failed geocoding mainly due to incomplete address which were removed from the study population. However, these excluded births due to failed geocoding did not have a higher risk of adverse birth outcomes than those with successful geocoding. Another limitation of this study is the potential spatial confounding. There is a benefit of allowing both spatial and temporal contrast in this study; however, spatial confounding, especial social economic status, may bias the health association estimates. To account for potential confounding due to social economic factors, we controlled parental education levels in statistical models. We also conducted sensitivity analyses by adding a spatial spline smoother to control potential spatial confounding. When controlling for spatial trends, almost all the health association estimates decreased, partly due to the control of spatial variation in $PM_{2.5}$ exposure. The ability of flexible spatial smoothers to attenuate effects of spatial covariates due to collinearity has been well documented by previous studies [35, 36].

Another limitation of this study is the lack of behavioral information. Indoor smoking and alcohol consumption were treated as potential confounders in some previous studies on birth outcomes [18, 20], but this information is not recorded in birth registration dataset. However, Ritz, Wilhelm [37] reported that adjusting for personal behavioral variables, including active and passive smoking, marital status, and alcohol consumption, did not change the health effect estimates of preterm birth in association with air pollution. Darrow, Woodruff [38] also reported that though maternal smoking was a strong predictor of infant respiratory mortality, it did not confound the associations between ambient air pollution and mortality. Residential mobility may be an additional limitation, leading to exposure misclassification when assigning maternal exposure with $PM_{2.5}$ concentrations at the residence of birth. Pennington, Strickland [39] estimated that without accounting for residential mobility led to -2% to -10% bias towards the null in cohort that

18.6% of children were born to mothers changed resident during pregnancy. A previous study reported that 8.4% of pregnant women changed residence in Wuhan, China [29]. Unfortunately, there is no study on residential mobility in Shanghai, but we expect residential mobility to be nondifferential with respect to exposure and birth outcomes, thus the potential bias would be toward to null.

CONCLUSIONS

We reported decreased birth weight as well as increased risk of preterm birth and term LBW in association with maternal $PM_{2.5}$ exposure in Shanghai, China, from 2011-2014. The magnitude of associations between maternal $PM_{2.5}$ exposure and birth outcomes was slightly higher than previously reported findings. Health association estimates were influenced by exposure assessment approaches, and when using satellite predictions for exposure assessment, researchers should account for missing data. We observed higher magnitudes of associations between first and third trimester exposure and birth weight, as well as between first trimester exposure and preterm birth. Mothers older than 35 years and without college education tended to have higher risk of preterm birth.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health [grant number R01ES027892]; the Public Welfare Research Program of National Health and Family Planning Commission of China [grant number 201502003]; and the General Program of Health Bureau of Shanghai Pudong New Area [grant number PW2016A-6].

REFERENCES

1. Rogers, L.K. and M. Velten, *Maternal inflammation, growth retardation, and preterm birth: insights into adult cardiovascular disease.* Life sciences, 2011. **89**(13-14): p. 417-421.
2. Swamy, G.K., *Preterm birth, a known risk factor for infant and childhood death, is an independent risk factor for mortality in early childhood and young adulthood.* BMJ evidence-based medicine, 2011: p. ebmed-2011-100361.
3. CDC, *Infant mortality and low birth weight among black and white infants--United States, 1980-2000.* MMWR. Morbidity and mortality weekly report, 2002. **51**(27): p. 589.
4. Liu, L., et al., *Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis.* The Lancet, 2015. **385**(9966): p. 430-440.
5. Polichetti, G., et al., *Effects of ambient air pollution on birth outcomes: an overview.* Crit Rev Environ Sci Technol, 2013. **43**(12): p. 1223-1245.
6. Li, X., et al., *Association between ambient fine particulate matter and preterm birth or term low birth weight: An updated systematic review and meta-analysis.* Environ Pollut, 2017.
7. Sun, X., et al., *The association between fine particulate matter exposure during pregnancy and preterm birth: a meta-analysis.* BMC Pregnancy Childbirth, 2015. **15**(1): p. 300.
8. Sun, X., et al., *The associations between birth weight and exposure to fine particulate matter (PM 2.5) and its chemical constituents during pregnancy: A meta-analysis.* Environ Pollut, 2016. **211**: p. 38-47.
9. Seltenrich, N., *Remote-sensing applications for environmental health research.* Environ Health Perspect, 2014. **122**(10): p. A268.
10. Xiao, Q., et al., *Evaluation of VIIRS, GOCI, and MODIS Collection 6 AOD retrievals against ground sunphotometer observations over East Asia.* Atmos Chem Phys, 2016. **16**(3): p. 1255-1269.
11. Strickland, M.J., et al., *Pediatric emergency visits and short-term changes in PM2. 5 concentrations in the US State of Georgia.* Environ Health Perspect, 2016. **124**(5): p. 690.
12. Xiao, Q., et al., *Full-coverage high-resolution daily PM 2.5 estimation using MAIAC AOD in the Yangtze River Delta of China.* Remote Sens Environ, 2017. **199**: p. 437-446.
13. Strand, L.B., A.G. Barnett, and S. Tong, *Methodological challenges when estimating the effects of season and seasonal exposures on birth outcomes.* BMC Med Res Methodol, 2011. **11**(1): p. 49.
14. Platnick, S., et al., *The MODIS cloud products: Algorithms and examples from Terra.* IEEE Transactions on Geoscience and Remote Sensing, 2003. **41**(2): p. 459-473.
15. Appel, K.W., et al., *Examination of the Community Multiscale Air Quality (CMAQ) model performance over the North American and European domains.* Atmospheric environment, 2012. **53**: p. 142-155.
16. Lucchesi, R., *File Specification for GEOS-5 FP (Forward Processing).* GMAO Office Note No. 4 (Version 1.1), 2013.
17. Chang, H.H., B.J. Reich, and M.L. Miranda, *Time-to-event analysis of fine particle air pollution and preterm birth: results from North Carolina, 2001–2005.* Am J Epidemiol, 2011. **175**(2): p. 91-98.
18. Hao, H., et al., *Air pollution and preterm birth in the US State of Georgia (2002–2006): associations with concentrations of 11 ambient air pollutants estimated by combining Community Multiscale Air Quality Model (CMAQ) simulations with stationary monitor measurements.* Environ Health Perspect, 2016. **124**(6): p. 875.

19. Johnson, S., et al., *Ambient fine particulate matter, nitrogen dioxide, and preterm birth in New York City.* Environmental health perspectives, 2016. **124**(8): p. 1283.
20. Woodruff, T.J., et al., *Methodological issues in studies of air pollution and reproductive health.* Environ Res, 2009. **109**(3): p. 311-320.
21. Ferré, C., *Effects of maternal age and age-specific preterm birth rates on overall preterm birth rates—United States, 2007 and 2014.* MMWR Morb Mortal Wkly Rep, 2016. **65**.
22. Fleischer, N.L., et al., *Outdoor air pollution, preterm birth, and low birth weight: analysis of the world health organization global survey on maternal and perinatal health.* Environ Health Perspect, 2014. **122**(4): p. 425.
23. Xue, Q., et al., *An analysis of the medical indications for preterm birth in an obstetrics and gynaecology teaching hospital in Shanghai, China.* Midwifery, 2016. **35**: p. 17-21.
24. Kurdi, A.M., et al., *Multiple pregnancy and preterm labor.* Saudi medical journal, 2004. **25**(5): p. 632-637.
25. Tang, W., et al., *Low birthweight in China: evidence from 441 health facilities between 2012 and 2014.* The Journal of Maternal-Fetal & Neonatal Medicine, 2017. **30**(16): p. 1997-2002.
26. Chen, Y., et al., *An epidemiological survey on low birth weight infants in China and analysis of outcomes of full-term low birth weight infants.* BMC pregnancy and childbirth, 2013. **13**(1): p. 242.
27. Yu, C., et al., *Statistical evaluation of the feasibility of satellite-retrieved cloud parameters as indicators of PM2. 5 levels.* Journal of Exposure Science and Environmental Epidemiology, 2015. **25**(5): p. 457-466.
28. Dadvand, P., et al., *Maternal exposure to particulate air pollution and term birth weight: a multi-country evaluation of effect and heterogeneity.* Environ Health Perspect, 2013. **121**(3): p. 267.
29. Qian, Z., et al., *Ambient air pollution and preterm birth: a prospective birth cohort study in Wuhan, China.* Int J Hyg Environ health, 2016. **219**(2): p. 195-203.
30. Jerrett, M., et al., *Comparing the health effects of ambient particulate matter estimated using ground-based versus remote sensing exposure estimates.* Environ Health Perspect, 2017. **125**(4): p. 552.
31. Pedersen, M., et al., *Ambient air pollution and low birthweight: a European cohort study (ESCAPE).* Lancet Respir Med, 2013. **1**(9): p. 695-704.
32. Laurent, O., et al., *Sources and contents of air pollution affecting term low birth weight in Los Angeles County, California, 2001–2008.* Environ Res, 2014. **134**: p. 488-495.
33. Vinikoor-Imler, L.C., et al., *Associations between prenatal exposure to air pollution, small for gestational age, and term low birthweight in a state-wide birth cohort.* Environ Res, 2014. **132**: p. 132-139.
34. Genereux, M., et al., *Neighbourhood socioeconomic status, maternal education and adverse birth outcomes among mothers living near highways.* J Epidemiol Community Health, 2008. **62**(8): p. 695-700.
35. Paciorek, C.J., *The importance of scale for spatial-confounding bias and precision of spatial regression estimators.* Stat Sci, 2010. **25**(1): p. 107.
36. Hodges, J.S. and B.J. Reich, *Adding spatially-correlated errors can mess up the fixed effect you love.* The American Statistician, 2010. **64**(4): p. 325-334.
37. Ritz, B., et al., *Ambient air pollution and preterm birth in the environment and pregnancy outcomes study at the University of California, Los Angeles.* Am J Epidemiol, 2007. **166**(9): p. 1045-1052.

38. Darrow, L.A., T.J. Woodruff, and J.D. Parker, *Maternal smoking as a confounder in studies of air pollution and infant mortality.* Epidemiology, 2006. **17**(5): p. 592-593.
39. Pennington, A.F., et al., *Measurement error in mobile source air pollution exposure estimates due to residential mobility during pregnancy.* J of Expo Sci Environ Epidemiol, 2016.

Table 2. 1 Descriptive statistics of the birth cohort in Shanghai during 2011-2014 (n=132 783).

Variable	Level	
Birth weight among term births (g)	Mean (std)	3352 (447)
Preterm birth (%)		4.41
Gestational age (week)	Mean (std)	39.0 (1.3)
Term low birth weight (%)		0.95
Gender (%)	Female	47.0
Parity (%)	1	69.9
	2	27.5
	>2	2.6
Birth location (%)	Hospital	65.0
	Maternal health service center	35.0
	others	0.02
Mother's age (year)	Mean (std)	28 (5)
Father's education level (%)	Graduate	9.4
	College	43.6
	High school or lower	47.0
Mother's education level (%)	Graduate	7.1
	College	45.4
	High school or lower	47.5

Table 2. 2 Descriptive statistics of the three $PM_{2.5}$ exposure metrics.

	Pregnancy		1st trimester		2nd trimester		3rd trimester	
	Mean (std)	IQR	Mean (std)	IQR	Mean (std)	IQR	Mean (std)	IQR
Gap-filled $PM_{2.5}$ prediction	60 (9)	11	63 (17)	24	61 (16)	23	57 (16)	25
$PM_{2.5}$ prediction with missingness	72 (10)	13	71 (19)	29	69 (18)	26	65 (19)	19
$PM_{2.5}$ measurements from stations	58 (7)	14	59 (18)	18	58 (17)	16	55 (18)	19

Table 2. 3 Health effect estimates per 10 μg/m^3 increase in gap-filled satellite based $PM_{2.5}$ exposure for births during 2011-2014, controlling for spatial and temporal trends by generalized additive model.

	Change in birth weight	OR for preterm birth
First trimester	1.08 (-2.81, 4.96)	1.06 (1.01, 1.11)
Second trimester	0.66 (-3.61, 4.93)	0.96 (0.91, 1.01)
Third trimester	-1.40 (-5.78, 2.97)	1.00 (0.97, 1.03)
Entire pregnancy	-15.32 (-25.76, -4.88)	1.06 (0.95, 1.17)

Figure 2. 1 Annual average $PM_{2.5}$ concentrations from gap-filled satellite predictions and central-site measurements (circle) in 2014.
The boundary of Shanghai (black line) and the main roads (thin grey lines) are overlaid on the map.

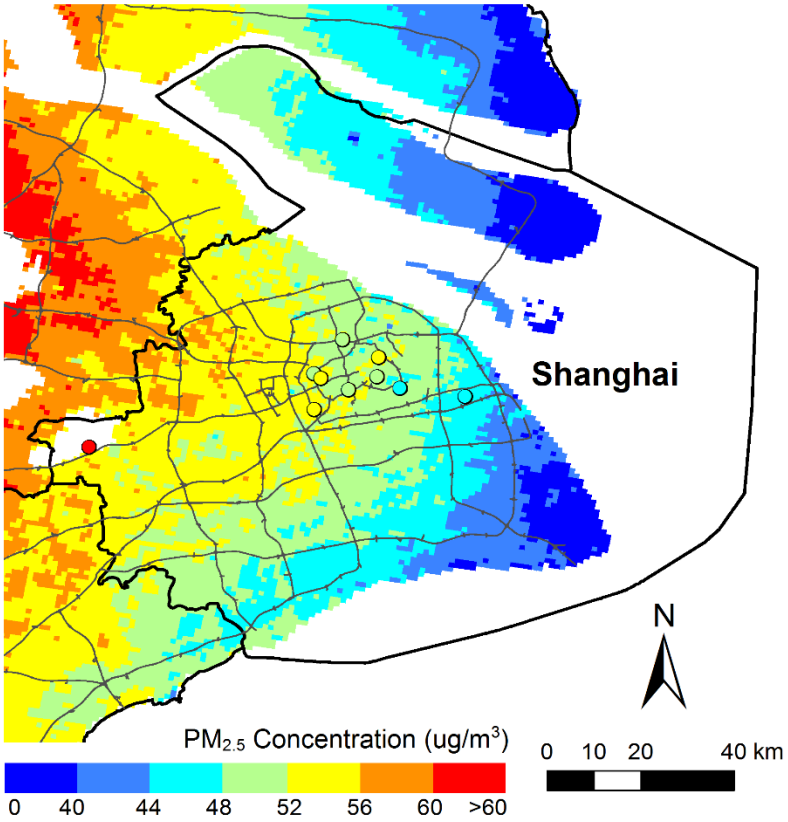


Figure 2. 2 Adjusted health association estimates per 10 ug/m^3 increase in $PM_{2.5}$ exposure during each trimester and entire pregnancy for births between 2011 and 2014, using exposure assessed from gap-filled satellite predictions.

Left: adjusted birth weight change among term births and 95% confidence interval (CI). Right: adjusted odds ratio (OR) and 95% CI for preterm birth and term low birth weight.

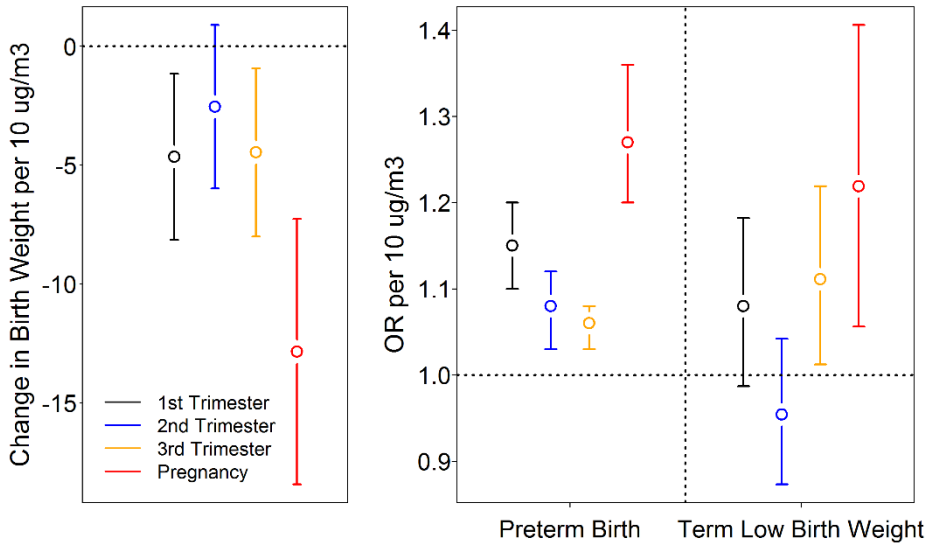


Figure 2. 3 Adjusted OR for preterm birth (left), adjusted birth weight change in term births (middle) and adjusted OR for term LBW (right) per 10 µg/m^3 increase in $PM_{2.5}$ exposure during each trimester and entire pregnancy for births between 2013 and 2014.

Exposures were based on gap-filled satellite predictions, satellite predictions without accounting for missingness, and ground measurements.

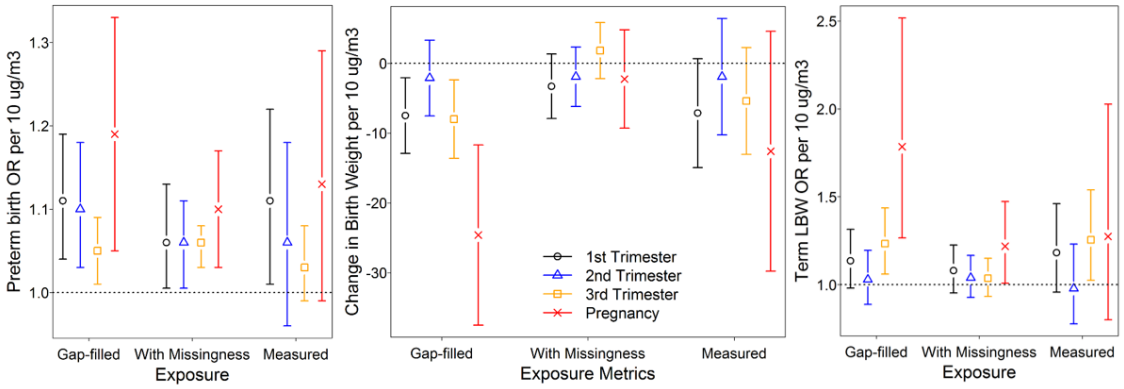
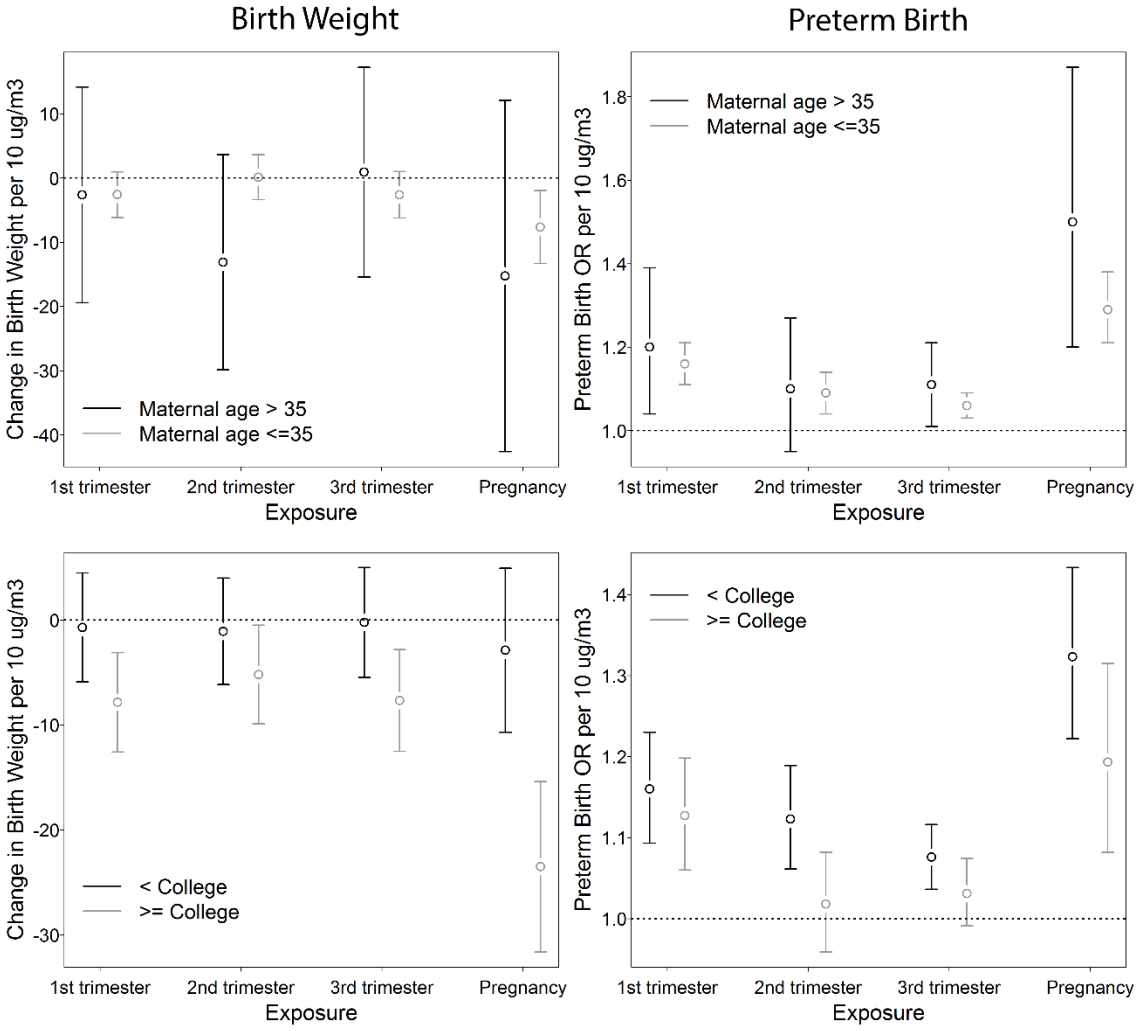


Figure 2. 4 Adjusted health effect estimates per 10 µg/m^3 increase in trimester-specific and entire pregnancy $PM_{2.5}$ exposures, stratified by maternal age and maternal education level for term births during 2011-2014. Exposures were based on gap-filled satellite predictions.



APPENDIX A

Figure 2.A 1 Study population distribution in Shanghai, China, during 2011-2014.

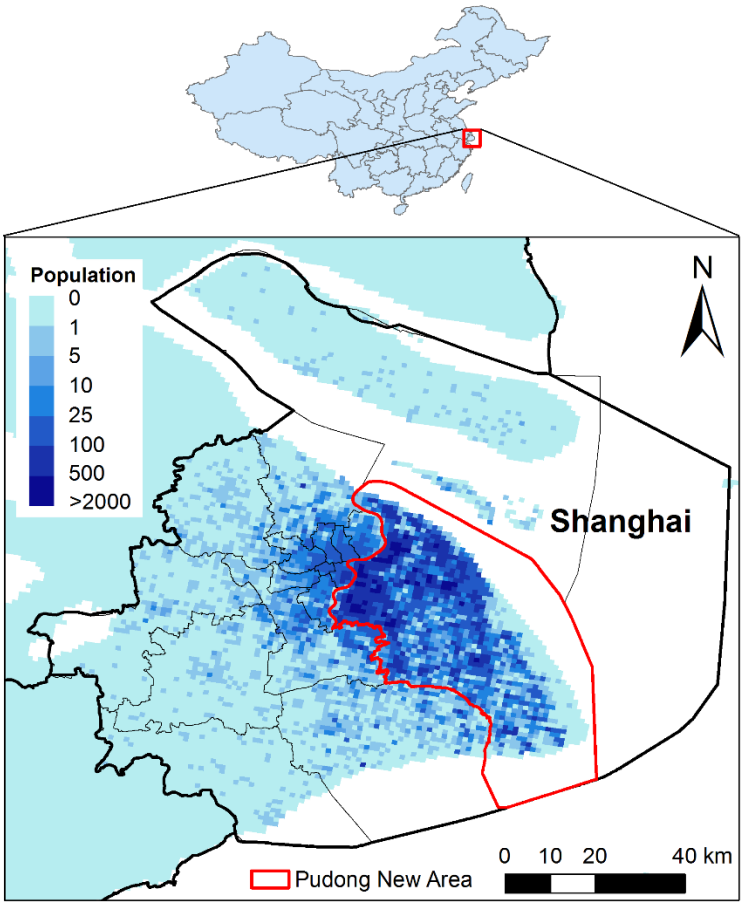


Figure 2.A 2 Temporal trends of $PM_{2.5}$ concentrations measured at ten monitoring stations (1-10) in Shanghai during 2013-2014.

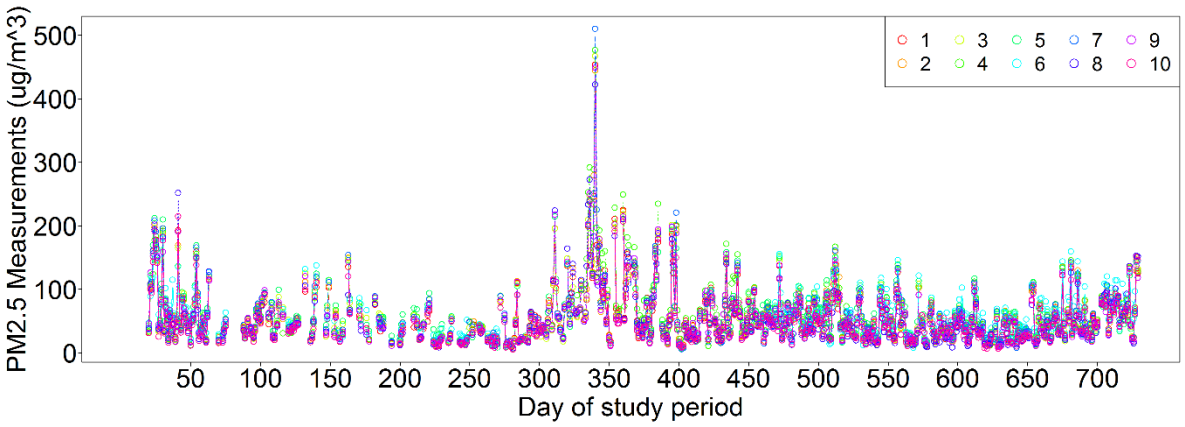


Table 2.A 1 Correlation coefficients between the three $PM_{2.5}$ exposure metrics.

	Gap-filled predictions vs. Measurements	With-missing predictions vs. Measurements	Gap-filled predictions vs. with-missing predictions
1st trimester	0.96	0.88	0.93
2nd trimester	0.97	0.87	0.91
3rd trimester	0.97	0.89	0.92
Pregnancy	0.83	0.80	0.96

APPENDIX B

Figure 2.B 1 Adjusted health effect estimates per IQR increase in $PM_{2.5}$ exposures during each trimester and entire pregnancy for births between 2011 and 2014, using exposure assessed from gap-filled satellite predictions. Left: adjusted birth weight change among term births and 95% confidence intervals. Right: adjusted odds ratio (OR) and 95% confidence intervals for preterm birth and term low birth weight.

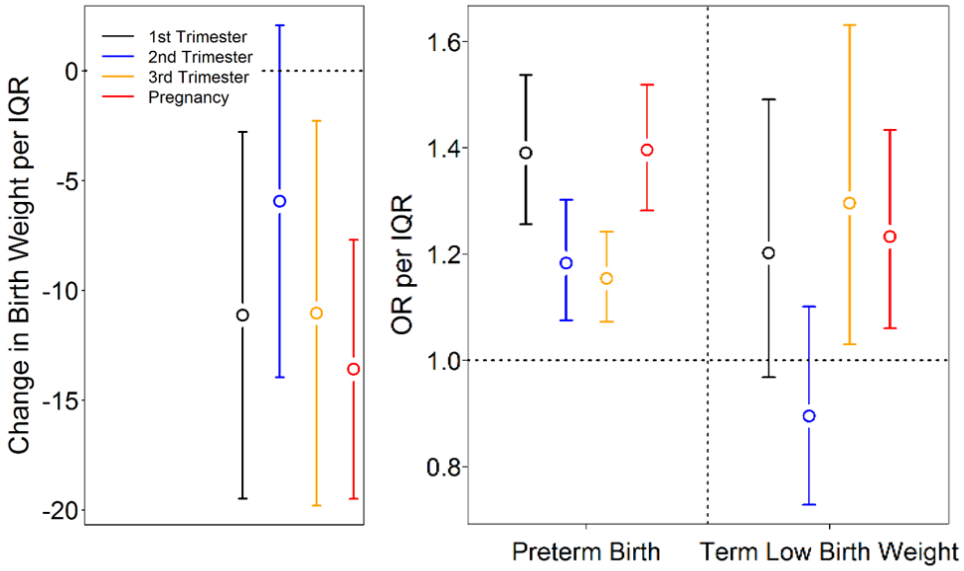


Figure 2.B 2 Adjusted health effect estimates per 10 µg/m^3 increase in trimester-specific and entire pregnancy $PM_{2.5}$ exposure, stratified by paternal education level for term births during 2011-2014. Exposures were based on gap-filled satellite predictions.

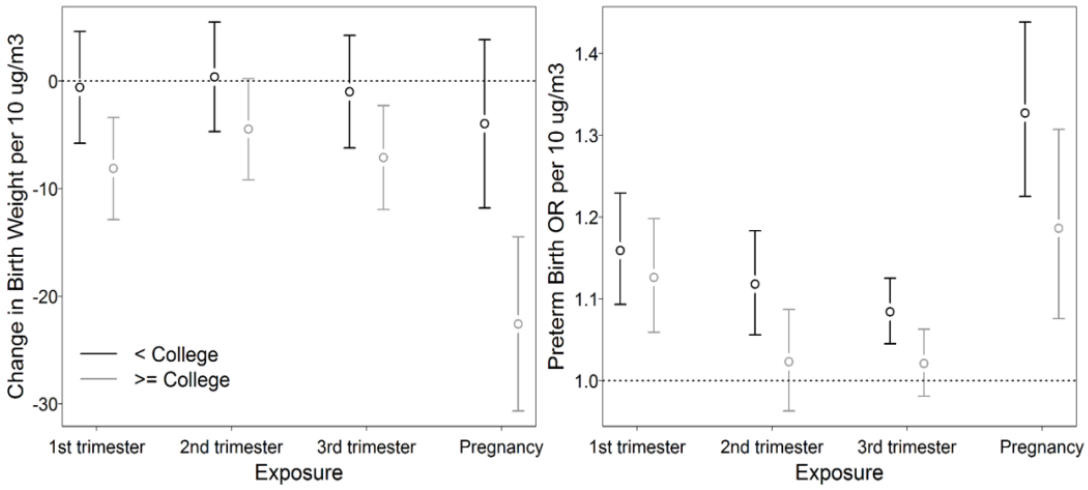
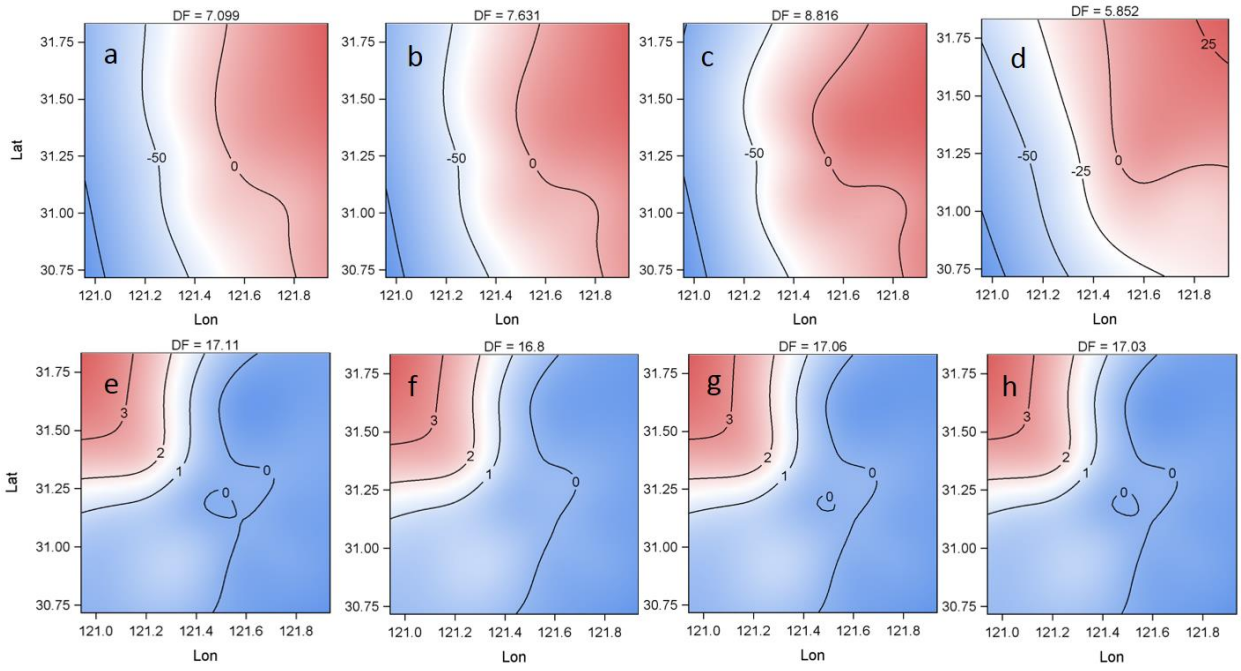


Figure 2.B 3 The fitted spatial patterns from GAM in the sensitivity analysis.

a-d: spatial patterns in models relating term birth weight to exposure during the first (a), second (b), third (c) trimester and entire pregnancy (d). e-f: spatial patterns in models relating preterm birth to exposure during the first (e), second (f), third (g) trimester and entire pregnancy (h).



Chapter 3

A machine-learning model to predict historical $PM_{2.5}$ concentrations in China from satellite data

Qingyang Xiao, Howard H. Chang, Guannan Geng, Yang Liu

ABSTRACT

Background: The long satellite aerosol data record enables assessments of historical $PM_{2.5}$ level in developing countries such as China where routine $PM_{2.5}$ monitoring began only recently. However, most previous models reported decreased prediction accuracy when predicting $PM_{2.5}$ levels outside the model-training period. This limitation greatly hinders the application of satellite-driven exposure assessments in the research on health effects of long-term $PM_{2.5}$ exposure.

Objectives: We proposed an ensemble machine learning approach that provided reliable $PM_{2.5}$ hindcast capabilities in China.

Methods: Non-random missing satellite data due to cloud cover were first filled by multiple imputation to ensure unbiased long-term exposure estimates. Then the modeling domain, China, was divided into seven regions using a spatial clustering method to control for unobserved spatial heterogeneity. A set of machine learning models including random forest, generalized additive model, and extreme gradient boosting were trained in each region separately. Finally, a generalized additive ensemble model was developed to combine predictions from different algorithms.

Results: The ensemble prediction characterized the spatiotemporal distribution of daily $PM_{2.5}$ well with the cross-validation (CV) R^2 (RMSE) of 0.79 (21 $\mu g/m^3$). The cluster-based sub-region models outperformed national models and improved the CV R^2 by ~0.05. Compared with previous studies, our model provided more accurate hindcasts at the daily level (R^2 = 0.53, RMSE = 28 $\mu g/m^3$) and monthly level (R^2 = 0.81, RMSE = 13 $\mu g/m^3$).

Conclusions: Our hindcast modeling system allows for the construction of long-term, unbiased historical $PM_{2.5}$ levels that can support epidemiologic studies on the chronic health effects of $PM_{2.5}$ in China.

INTRODUCTION

Numerous studies have documented associations between $PM_{2.5}$ (fine particulate matter with an aerodynamic diameter of 2.5 μm or less) and adverse health outcomes, including respiratory diseases, cardiovascular diseases, and lung cancer [1, 2]. The 2015 Global Burden of Diseases study identified ambient $PM_{2.5}$ as the fifth largest overall risk factor for global mortality and estimated that $PM_{2.5}$ exposure is responsible for 4.2 million deaths in 2015 [3]. However, studies on the health effects, especially chronic health effects, of $PM_{2.5}$ exposure are limited in highly polluted regions [4], due to the lack of $PM_{2.5}$ measurements. For instance, in China, the national air quality monitoring network was established in 2013 such that $PM_{2.5}$ measurements before 2013 were unavailable, making it difficult to assess long-term $PM_{2.5}$ exposure levels. To extend ground air quality monitoring networks, satellite-retrieved aerosol optical depth (AOD) has been increasingly used for air pollution monitoring and population exposure assessment in the past decade. Satellite data with broad spatial coverage, a long data record and high spatial resolutions could support the assessment of historical air pollution levels in environmental epidemiological studies.

Previous studies revealed that the relationship between satellite AOD and ground $PM_{2.5}$ concentration is complex and non-linear. Various statistical models have been presented to describe this relationship, addressing the effects of meteorological parameters, emission sources, and land use information [5-9]. Benefited from the long $PM_{2.5}$ ground monitoring record, previous $PM_{2.5}$ prediction models in the North America and Europe have aimed to extend the spatial coverage of $PM_{2.5}$ monitoring networks rather than to generate historical $PM_{2.5}$ levels. These models often included daily random effects or day-stratification to improve performance. Although day-specific intercepts and slopes can capture the unobserved fine-scale temporal trends in the associations between $PM_{2.5}$ concentration and explanatory variables, applying the

daily effects outside the model fitting period imposes a strong and often unrealistic assumption that the estimated daily effects during the model fitting period will remain constant during the hindcast period. When applying these models in regions lacking historical measurements, e.g. China, the model performance degraded significantly outside the model fitting period. For example, Ma et al. (2016) reported that when using a model fitted with data of 2013 to predict daily $PM_{2.5}$ concentrations in 2014, the R^2 was 0.41 compared with the 10-fold CV R^2 of 0.79. He and Huang [10] also reported that using model fitted with data of 2015 to predict daily $PM_{2.5}$ concentrations in 2014 had R^2 of 0.47, wheare the model CV R^2 was 0.80.

Another $PM_{2.5}$ modeling approach, driven by atmospheric chemical transport model (CTM) simulations, has also been reported [11, 12]. This approach estimated the scaling factor between AOD and $PM_{2.5}$ from model simulations, and applied the scaling factor to satellite retrieved AOD to get $PM_{2.5}$ estimations. Because CTMs simulate historical AOD and $PM_{2.5}$, this approach can estimate historical $PM_{2.5}$ levels from satellite AOD at global scale. However, the relatively low accuracy of CTM simulations limited the performance of this approach and the prediction accuracy was not comparable to statistical models. For example, Geng, Zhang [13] reported that the R^2 of the linear relationship between five-month-mean $PM_{2.5}$ predictions and ground measurements was 0.72 in China in 2013.

Most recently, machine learning algorithms have been applied to $PM_{2.5}$ prediction. Machine learning algorithms can deal with complex non-linear relationships with interactions, making them promising in air pollution prediction. Di, Kloog [9] fitted a neural network to predict $PM_{2.5}$ concentrations from the Multi-Angle Implementation of Atmospheric Correction (MAIAC) AOD at 1 km resolution over the continental U.S. The model 10-fold CV R^2 is 0.84. Hu, Belle [14] fitted a random forest model with the Moderate Resolution Imaging Spectroradiometer (MODIS) AOD and reported the 10-fold CV R^2 of 0.80. These two studies included convolutional layers of $PM_{2.5}$ estimated from ground measurements to characterize the spatial correlation of $PM_{2.5}$. As a

result, these US-based models cannot estimate historical $PM_{2.5}$ levels when ground measurements were unavailable. Li, Shen [15] trained a deep belief network in China with the 10-fold CV R^2 of 0.88, but this model relied on both spatial and temporal correlations of $PM_{2.5}$ estimated from ground observations, making it unable to hindcast in space and time. Gradient boosting [16] and Generalized regression neural network [17] have also been employed to predict daily $PM_{2.5}$ concentrations in China, with the 10-fold CV R^2 of 0.76 and 0.67, respectively. Although these models did not rely on $PM_{2.5}$ measurements to construct input variables, neither of these two studies examined their models' hindcast ability.

Still at its infancy, machine learning $PM_{2.5}$ models could be improved in several aspects. For example, previous studies revealed significant spatial heterogeneity in relationships between $PM_{2.5}$, satellite AOD, and meteorological parameters [9, 18]. Thus, dividing a large modeling domain and training regional models could help control for unobserved spatial features and improve model performance [14]. Additionally, parallelizing the cluster-based model fitting process can significantly increase computational efficiency, especially for machine learning algorithms that normally require a longer time to converge. Previous studies in the U.S. divided study domains according to climate regions defined by the National Oceanic and Atmospheric Administration (NOAA) [14, 19], but it is not clear how to divide China into reasonable sub-regions. Ma et al. (2016) fitted their multi-stage model for each province in China but provincial areas vary dramatically, ranging from 0.03 million km^2 (Hainan) to 1.6 million km^2 (Xinjiang). In addition, provincial boundaries do not necessarily reflect any geographic or emission patterns and observations from one province is generally insufficient to support a complex model. Thus, researchers had to select different buffer radii manually to ensure sufficient model fitting data in each province-based region and the buffer radii changed when the number of ground monitoring stations changed.

In this study, we proposed a machine-learning approach specifically designed to provide high-quality historical $PM_{2.5}$ concentration estimates in developing countries such as China. We developed a clustering method to divide China into seven temporally stable regions. Then we trained a set of machine learning models in each region that did not rely on daily effects during 2013-2016. We finally combined predictions from various models by an additive ensemble model. We evaluated model hindcast predictions in 2017 and during the 2008 Beijing Olympic Games.

METHODS

Data

The study domain covers mainland China, Hong Kong special administrative region and Taiwan (Figure 1). We constructed a 0.1 degree modeling grid covering this study domain for data integration. We used data during 2013-2016 for model training, and data during the 2008 Beijing Olympic Games as well as the first seven months of 2017 for hindcast evaluation.

$PM_{2.5}$ measurements

Hourly $PM_{2.5}$ concentrations in 2013-2017 were measured at ~1,593 air quality monitoring stations across mainland China (Figure 1). Since the national air quality monitoring network was under development during the study period, the number of monitoring stations increased over the years. Measurements are published by the China National Environmental Monitoring Center (CNEMC, http://www.cnemc.cn/), and were downloaded from PM25.in (http://pm25.in/), a direct mirror of data from CNEMC. Additionally, we collected $PM_{2.5}$ measurements in Hong Kong from the Hong Kong environmental protection department (http://epic.epd.gov.hk/) and $PM_{2.5}$ measurements of Taiwan from the Taiwan environmental protection agency (http://taqm.epa.gov.tw/). We removed repeated identical measurements for at least three continuous hours, assuming that such repetition was due to instrument malfunction. Daily average concentrations were calculated from hourly measurements during 0:00-23:59 local time. Days with less than 18 hourly measurements were excluded. Daily average $PM_{2.5}$ measurements from

stations located within the same grid cell were averaged. Finally, we analyzed $PM_{2.5}$ concentrations during the 2008 Summer Olympic Games measured at three locations in Beijing, China: Tsinghua University, Daxing District, and Miyun District during June to October, 2008 [20] as a test of model hindcast capabilities. These three temporary sampling sites were established during a field experiment and their locations do not coincide with any regulatory monitors later.

Satellite data

The MODIS Collection 6 level 2 aerosol products at 10 km resolution from Aqua and Terra satellites were downloaded from the Atmospheric Archive and Distribution System (http://ladsweb.nascom.nasa.gov/). Since MODIS retrievals were affected by the bow-tie effect (pixels were stretched at the border of each scan), to correctly assign AOD retrievals to the 0.1 degree-grid cell, we created Thiessen polygons from centroid of AOD pixels. Two retrieval algorithms, Deep Blue (DB) algorithm and Dart Target (DT) algorithm, have been developed to retrieve AOD at 10 km resolution [21, 22]. These two algorithms use different methods to characterize and remove surface reflectance. Thus, they are suitable for retrievals over different land surfaces. The Dark Target algorithm provides high quality retrievals over vegetation covered land, while the Beep Blue algorithm is able to retrieve AOD over bright land, e.g. urban regions. The MODIS Collection 6 products also provided a parameter, "combined AOD" that combines high quality retrievals from Deep Blue and Dark Target algorithms, accounting for surface situations. Since the combination only includes high quality retrievals, its coverage is very limited. In this study, we included all three AOD parameters as separate inputs in our machine learning models.

Due to cloud cover or high surface reflectance, about 40-70% of satellite retrievals are missing on average in East Asia[23]. To improve the coverage of satellite retrievals without decreasing retrieval quality, we filled data gaps in DB AOD, DT AOD and combined AOD separately using

multiple imputation. The details of this method are provided elsewhere and here is a brief summary [7]. We first fitted daily linear regressions between AOD retrievals from Aqua satellite (overpass time at 1:30 pm local time) and Terra satellite (over pass time at 10:30 am local time), and used the regression coefficients to estimate the missing Aqua/Terra AOD when only one of them is present. Then the observed and predicted AOD values were averaged to reflect daily aerosol loadings [24]. We then filled the missing daily average AOD by multiple imputation with an additive model driven by chemical transport model AOD simulations, temperature and humidity in the boundary layer, elevation, and MODIS cloud fraction [25]. Each missing daily AOD was imputed five times to account for the additional uncertainty due to imputation and the average of the five imputed AODs served as a predictor in machine learning models.

The MODIS active fire data were obtained from the Fire Information for Resource Management System (FIRMS, https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms). We developed buffers with various radii, including 20 km, 30 km, 50km, and 75 km, to assign fire information to the corresponding 0.1° modeling grid cell. Specifically, we searched and summed the number of fire spots within each buffer centered on the centroid of each grid cell. We extracted *cloud_fraction_day* from Aqua and Terra Collection 6 level 2 cloud products (MYD06_L2 and MOD06_L2), at 5 km spatial resolution. Daily cloud fraction was calculated as the average of Aqua and Terra cloud fraction that were interpolated to 0.1 degree grid cell by the nearest neighbor approach. Normalized Difference Vegetation Index (NDVI) data were obtained from Terra MODIS monthly global NDVI dataset at 1 km resolution (MOD13A3). The NDVI value of each grid cell was assigned as the average of NDVI pixels falling within the corresponding grid cell. Missing data in NDVI were interpolated by inverse distance weighting.

The tropospheric vertical column NO_2 density and absorbing aerosol index (AAI) data in visible light and UV light from Ozone Monitoring Instrument (OMI) was downloaded from the Goddard Earth Sciences Data and Information Services Center (https://mirador.gsfc.nasa.gov/). We

extracted and processed the parameters *ColumnAmountNO2Trop* from the OMI NO2 level 2 data (OMNO2), *AerosolIndexUV* and *AerosolIndexVIS* from the OMI Aerosol Extinction Optical Depth and Aerosol Types level 2 data (OMAERO), and *UVAerosolIndex* from the OMI Near-UV Aerosol Absorption and Extinction Optical Depth and Single Scattering Albedo level 2 data (OMAERUV). These retrievals are at 13×24 km^2 resolution at nadir. Due to row anomaly started from 2007, retrievals with the cross track anomaly flag as nonzero were removed and oversampling was conducted to smooth the systematic noise. Regarding the NO_2 column density, the value of each 0.1° grid cell was assigned as the average of samples from a 20 km-radius buffer centered on this grid cell during each season. Regarding the AAI parameters, retrievals with lower than 0.5% percentile were removed and the value of each 0.1° grid cell were assigned as the average of samples from a 30 km-radius buffer centered on this grid cell during each season. This oversampling approach led to ~100 NO_2 column density measurements and ~250 AAI measurements being averaged per season in each grid cell.

Meteorological and land use data

Meteorological parameters in 2013-2017 were extracted from the Goddard Earth Observing System Data Assimilation System GEOS-5 Forward Processing (GEOS 5-FP) at 0.25° latitude × 0.3125° longitude resolution. Meteorological parameters in 2008 were extracted from the Goddard Earth Observing System Model, Version 5 (GEOS 5) at 0.5° × 0.5° resolution. The meteorological data were downscaled to 0.1 degree grid cell through a daily smooth surface estimated by inverse distance weighting. The elevation data were obtained from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM) version 2 at 30 m resolution. Population density data were obtained from the LandScan Global Population Database at 1 km resolution [26].

Since we extracted various wind parameters at different heights of the atmosphere (wind direction, u and v component of wind speed at 10 m, averaged in the boundary layer, and at 500 mb), to

reduce feature space and avoid the curse of dimensionality [27], we applied dimension reduction by Linear Discriminant Analysis (LDA) on these wind parameters [28]. The categorical output required by LDA was defined by separating the continuous $PM_{2.5}$ concentrations into 164 levels. We extracted the first and second components from LDA that cumulatively explained over 95% of variabilities in all wind parameters.

MERRA-2 $PM_{2.5}$ reanalysis data

We obtained daily $PM_{2.5}$ simulations from the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) [29]. The MERRA-2 $PM_{2.5}$ simulations have complete coverage and relatively high accuracy at 0.5° latitude × 0.625° longitude resolution. Evaluation studies in the U.S. showed that MERRA-2 $PM_{2.5}$ simulations agreed well with ground measurements [30]. MERRA-2 data provided additional information on $PM_{2.5}$ distribution at broad scale. The total concentration of $PM_{2.5}$ was calculated using the following equation [31, 32]:

$$\mathrm{PM}_{2.5} = 1.375 \times \mathrm{SO}_4 + 2.1 \times \mathrm{OC} + \mathrm{BC} + \mathrm{Dust}_{2.5} + \mathrm{Sea\ salt}_{2.5}$$

where SO_4, OC, BC represent the MERRA-2 concentration of sulfate ion, organic carbon, and black carbon, respectively. $Dust_{2.5}$ and Sea $salt_{2.5}$ are the concentration of dust and sea salt with a radius less than 2.5 μm. Since MERRA-2 simulates dust and sea salt by five size bins, we summed dust concentrations of Bin 1 (radius 0.1~1.0 μm), Bin 2 (radius 1~1.5 μm), and Bin 3 (radius 1.5~3.0 μm), and sea salt concentrations of Bin 1 (radius 0.03~0.1 μm), Bin 2 (radius 0.1~0.5 μm), and Bin 3 (radius 0.5~1.5 μm). We multiplied SO_4 by 1.375 to get the concentration of sulfate aerosol, assuming that sulfate is primarily presented as ammonium sulfate. The ratio between organic carbon and organic matter, 2.1, was estimated from $PM_{2.5}$ observations and MERRA-2 organic carbon simulations in China during 2013-2016. The MERRA-2 $PM_{2.5}$ simulations at 50 km resolution was interpolated by inverse distance weighting to the 0.1° modeling grid.

Visibility data

The visibility data were extracted from the Integrated Surface Dataset (IDS) from the U.S. National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Information (NCEI). Visibility was measured at 407 stations in China (Figure 1). The daily average visibility was interpolated by inverse distance weighting and assigned to the 0.1° modeling grid.

METHODS

A diagram of our modeling method is shown in Figure 2. First, we divided our study domain according to the coefficient surface estimated from geographically weighted regression (GWR) by the K-Means algorithm. Then we trained three machine learning models, including random forest, extreme gradient boosting (XGBoost), and generalized additive model (GAM) in each region, separately. The decision tree based algorithms, random forest and XGboost, provided the estimated importance of predictors that guided parameter selection and construction of future models. These two algorithms performed well when predicting $PM_{2.5}$ concentrations in the U.S. [14, 33]. The GAM model has been widely used to characterize the spatial distribution of $PM_{2.5}$ [8, 34]. Finally, to improve the hindcast accuracy and robustness, we combined predictions from the three individual machine learning models by a GAM ensemble model. We trained and evaluated prediction models at the daily level, while predictions outside the model fitting period were aggregated to monthly level for hindcast performance evaluation, because studies on the chronic health effects of $PM_{2.5}$ normally assess exposure levels over a relatively long exposure window. Training prediction models at daily level allows flexible start and end dates of exposure windows in epidemiological studies.

The R package “mlr” was used to optimize hyperparameters of each algorithm through 5-fold CV and fixed holdout. Since this study aims to train a model with accurate hindcasts, we favor low variance than low bias in the bias-variance trade off. We evaluated the model performance by 10-

fold CV at daily level that we randomly selected 90% of data to train individual models and the ensemble model, and then we used the remaining 10% of data to examine the model performance. This process was repeated 10 times so that each data record was left for testing once. Because in such a standard cross validation, the randomly selected training dataset usually contains enough observations to estimate local spatial and temporal trends that may not hold constant outside the model fitting domain and period, we also conducted 10-fold CV spatially and temporally to detect potential spatial and temporal overfitting. For the spatial CV, we used data from randomly selected 90% of monitors to train the models and used data from the remaining 10% of monitors to test the model. Similarly, for the temporal CV, we used data from randomly selected 90% of days during the modeling period to train the model and used data from the remaining 10% of days to test the model. Since the CV results may underestimate the hindcast prediction error, we used data outside the training period (i.e. 2017) and outside the existing monitoring network (three temporary sites in 2008), to further characterize the prediction error.

Cluster analysis

We applied Geographically Weighted Regression (GWR), K-Means algorithm and GIS methods to identify appropriate clusters and divide our study domain to sub-regions. First, we fitted a GWR model with the annual average $PM_{2.5}$ concentrations together with annual mean DB AOD, meteorological variables, population density, and elevation (Equation 1). DB AOD was included because it had the highest coverage before gap-filling. GWR has been widely used to analyze spatially varying relationships [35]. It generates a continuous surface of regression coefficients through a spatial weighting mechanism from observations within a certain distance from each location. Since we aimed to control the spatial trend by clustering, we used annual average values for GWR fitting and ignored the temporal variations to avoid short-term fluctuations in cluster patterns.

$$PM2.5_{t,i} \sim Elev_i + DB\_AOD_{t,i} + Pop_{t,i} + Tem_{t,i} + Humidity_{t,i} + Prec_{t,i} + PBLH_{t,i} + AAI\_UV_{t,i} + Column\_NO2_{t,i} + e_{t,i}$$

Equation 1

where $PM2.5_{t,i}$ represents the annual average $PM_{2.5}$ concentrations of year t at grid cell i; $Elev_i$ represents the elevation of grid cell i; $DB\_AOD_{t,i}$ represents the annual average Deep Blue AOD of year t at grid cell i; $Pop_{t,i}$ represent the population of year t at grid cell i; $Tem_{t,i}$, $Humidity_{t,i}$, $Prec_{t,i}$, $PBLH_{t,i}$, $AAI\_UV_{t,i}$, and $Column_NO2_{t,i}$represent the annual average temperature, humidity, precipitation, planetary boundary layer height, AAI in UV light, and tropospheric vertical column NO_2 density of year t at grid cell, respectively.

After fitting the GWR, we clustered $PM_{2.5}$ monitors according to the vector of estimated coefficients by the K-Means algorithm and assigned $PM_{2.5}$ monitoring stations to different clusters. The number of clusters (k) was decided after comparing the clustering results using various values of k, ranging between 4 and 20. The estimated coefficients from GWR were normalized before clustering and we gave longitude and latitude higher weights to favor spatially continuous clusters. To examine the effects of randomization on the clustering results, we examined 20 different random seeds when selecting initial centroids and compared the K-Means clustering results. We used the most common clustering pattern for the following analysis. Thiessen polygons were generated from monitors and we assigned grid cells within each Thiessen polygon to the same cluster of the corresponding monitor in the center (Figure 3). We added a one degree buffer to each region and averaged the $PM_{2.5}$ predictions from different regional models in the buffer to ensure that the daily $PM_{2.5}$ predictions are spatially continuous. To examine the long-term stability of the clusters, we also estimated the clustering pattern by year as a sensitivity analysis.

Generalized additive model

GAM is a non-parametric model where the dependent variable depends linearly on smooth functions of predictors. We log transformed $PM_{2.5}$ concentrations to improve the prediction accuracy of high $PM_{2.5}$ values. The GAM model is shown as:

$$\lg\_PM2.5_{i,j} = s((Lon, Lat)_i) + s(DB\_AOD_{i,j}) + s(DT\_AOD_{i,j}) + s(AAI\_UV_{i,j}) + s(Prec_{i,j}) + s(Prec\_lag1_{i,j}) + s(Column\_NO2_{i,j}) + s(Humidity_{i,j}) + s(Tem_{i,j}) + s(Visibility_{i,j}) + s(MERRA2\_PM2.5_{i,j}) + s(Pop_{t,i}) + PBLH_{i,j} + e_{i,j}$$

Equation 2

where $lg\_PM2.5_{i,j}$ represents the log of $PM_{2.5}$ concentrations on day j at grid cell i; $s((Lon, Lat)_i)$ represents a thin plate surface of longitude and latitude of grid cell i; s() represents a smooth function of the corresponding parameter.

Random forest model

Initially proposed by Breiman [36], the random forest algorithm is a bagged classifier based on decision tree. The random forest algorithm offers several advantages over other machine learning algorithms: it can handle a large number of features without overfitting; it allows both continuous and categorical input variables; it is robust to outliers; and it provides variable importance as well as out of bag error for model evaluation. The random forest algorithm has been widely used for classification and regression. One limitation of the random forest algorithm is that with the increase of number of trees and complexity of each tree, the model training and prediction time can increase significantly. The hyperparameters of the random forest model were optimized by grid search and the training model performance was evaluated by out-of-bag statistics (Supplementary Text 1). Since the contribution of each predictor varied across regions, we selected predictors separately in each region.

Extreme gradient boosting model

The XGBoost algorithm is developed from gradient boosting [37]. Gradient boosting model has been shown to outperform various statistical and machine learning models in predicting $PM_{2.5}$ levels during a wildfire event [33]. XGBoost requires less training and predicting time than

random forest and has been widely used in data mining competitions [38, 39]. The R package, xgboost, was used to train the XGBoost model [40]. The hyperparameters of XGBoost model were selected by grid search (Supplementary Text 1). To avoid overfitting, only parameters with the evaluation statistic Gain, which describes the improvement in accuracy after splitting on the corresponding feature, larger than 0.01 were included in the model.

Ensemble model

To ensure a spatially continuous prediction surface, we fitted a national GAM model including predictions from the three individual models during the 4-year modeling period, 2013-2016. Predictions from the GAM model were transformed to normal scale before training the ensemble model. The ensemble model is shown as:

$$PM2.5_{i,j} = s(Pred\_RandomForest_{i,j}) + s\left(Pred\_XGBoost_{i,j}\right) + s\left(Pred\_GAM_{i,j}\right) + e_{i,j}$$

Equation 3

where $Pred\_RandomForest_{i,j}$, $Pred\_XGBoost_{i,j}$ and $Pred\_GAM_{i,j}$ are the predictions of $PM_{2.5}$ concentrations on day j at grid cell i from random forest, XGBoost and GAM, respectively.

RESULTS

Cluster analysis results

The estimated cluster map is shown in Figure 3. As expected, the separation of clusters did not follow provincial boundaries. Three northeastern provinces, i.e., Heilongjiang, Jilin, and Liaoning, as well as the northern Inner Mongolia constituted the Northeast cluster, characterized by its long winter/heating season and large presence of heavy industry. The North China Plain constituted the North cluster, characterized by its coal consumption [41, 42] and stagnant atmospheric conditions in winter, contributing to frequent regional haze events [43, 44]. The Yangtze River Delta was separated into two clusters: the relatively cold north (YRD) with central heating in winter and the relatively warm south without central heating (Southeast). The Pearl River Delta (PRD) was another cluster, located on the coast and characterized by its warm

weather. The PRD and Southeast clusters also produce more hydroelectricity than other regions [45]. The Qinghai-Tibetan Plateau, Sichuan, Yunnan, and Gansu province constituted the largest cluster (West) with a high altitude and low population density. Xinjiang Uygur Autonomous Region dominated the Northwest cluster, characterized by substantial dust emissions from the Taklamakan Desert. Changing the initial randomly selected centroid only led to slightly different cluster patterns (Figure S1). This cluster pattern was also stable across years (Figure S2), slightly affected by the increase in number of monitors during the modeling period. Increasing the number of clusters resulted in some scattered small clusters that did not have enough samples for model training, and decreasing the number of clusters led to merging of clusters into larger clusters. Thus, we used this seven-cluster map for modeling.

Individual machine learning model performance

Table 1 shows the model fitting and CV performance of individual cluster-based models and national models without clustering at the daily level. The density plots of model fitting performance of cluster-based models are shown in Figure S3. The density plots of the CV performance of cluster-based models and the reference national models are shown in Figure S4. The ensemble prediction outperformed all individual models in cross-validation, with a CV R^2 of 0.79, RMSE of 21 μg/m^3, slope of 1.00 and intercept of 0.0 at the daily level. The XGBoost model had the lowest CV RMSE (21 μg/m^3) and the highest CV R^2 (0.78) among individual models, followed by random forest (CV R^2 0.77, RMSE 22 μg/m^3). The random forest and GAM performed equally well in model training and standard 10-fold CV, while the XGBoost model's CV R^2 was 0.06 lower than model fitting R^2.

The cluster-based approach performed better than the national approach in all three algorithms (Table 1). The CV R^2 values of cluster-based XGBoost, random forest, and GAM model were 0.06, 0.06, and 0.05 higher than their national counterparts. As expected, prediction error increased in temporal and spatial CV relative to the standard CV, indicating that unobserved

spatial and temporal trends contributed to the prediction of $PM_{2.5}$. The random forest algorithm and the XGBoost algorithm relied more on the temporal trend: the R^2 in spatial CV was approximately 0.03 higher than the R^2 in temporal CV. On the contrary, GAM relied more on the spatial trend and showed no temporal overfitting with the spatial CV R^2 (0.58) lower than the temporal CV R^2 (0.65).

Model hindcast performance

We observed large variations in model hindcast performance across clusters and across algorithms. In general, the Northeast cluster had the lowest prediction accuracy (Table 2). The Northeast cluster had a long winter up to five months, leading to significant missingness in AOD retrievals due to snow/ice cover. These missing satellite data can hardly be accurately imputed by the current imputation model that aims to fill missing data due to cloud cover. The Northwest cluster showed the best hindcast performance that the monthly hindcast R^2 was 0.87, 0.85, 0.83 from random forest, GAM, and XGBoost respectively. This high prediction accuracy may be due to the single major particulate source from the Taklimakan Desert in this region thus the aerosol type had less variation across years. The relative model performance using different algorithms remained stable across clusters, i.e., the random forest model outperformed the XGBoost model in all clusters, and the XGBoost model outperformed GAM in all clusters expect the northwest cluster. However, the magnitude of model performance statistics varied in space. For example, all the algorithms described $PM_{2.5}$ levels in the Southeast cluster well, but random forest model provided significantly better hindcast in the Northeast cluster.

Nationwide, the cluster-based random forest model had the highest R^2 (0.83) and the lowest RMSE (12 $\mu g/m^3$), followed by the XGBoost model (R^2 0.82, RMSE 13$\mu g/m^3$) and GAM model (R^2 0.78, RMSE 15 $\mu g/m^3$) (Figure S5). Unfortunately, all these models underestimated the high $PM_{2.5}$ values. Consistent with a previous study [10], we noticed that the model temporal 10-fold CV error still underestimated the daily hindcast prediction error. For example, the cluster-based

random forest model had the temporal 10-fold CV R^2 (RMSE) of 0.72 (25 μg/m^3), while the daily hindcast R^2 was 0.57 (26 μg/m^3). This result suggested that long-term changes in $PM_{2.5}$ emission sources due to economic development and policy changes might affect the relationships between $PM_{2.5}$ and its predictors, but such changes in emission profiles were not well characterized in our current model.

Figure 4 shows hindcast performance of the ensemble model. The ensemble hindcast prediction had a slightly lower R^2 than the random forest model and the XGBoost model, but the linear regression between the ensemble hindcast predictions and ground measurements produced a slope closest to 1 (1.05) and a intercept closest to 0 (-3.18 μg/m^3), indicating a smaller prediction bias. To better evaluate our model's hindcast performance, we predicted $PM_{2.5}$ levels in 2008, five years before the model training period (Figure 4). During the Beijing Olympic and Paralympic Games, our daily hindcast prediction matched well with ground measurements from three stations in Beijing, with an R^2 value of 0.57 and RMSE of 27 μg/m^3. Compared with the daily hindcast performance in 2017 over the North region (R^2 = 0.53), the model performance did not appear to deteriorate in time.

The annual $PM_{2.5}$ distribution map in 2008 (Figure 5) indicated some hot spots of $PM_{2.5}$ in Beijing, Tianjin, Hebei province and Henan province. As a demonstration of our ensemble hindcast model, we estimated the annual $PM_{2.5}$ change rate during 2008-2016 with linear regression and noticed that the air quality at these hot spots was significantly improved during this eight-year period. During this eight-year period, $PM_{2.5}$ levels decreased or remained constant in most part of China [46]. The largest improvement in annual average $PM_{2.5}$ concentration occurred in Sichuan basin, followed by Henan province, Hebei province, Tianjin City, Taiyuan City, the Yangtze River Delta and Pearl River Delta, at more than 3 μg/m^3 per year. However, $PM_{2.5}$ levels in Northeast and Western China increased. For example, $PM_{2.5}$ levels in Heilongjiang, Jilin, Gansu, Qinghai, Shandong province have increased at approximately 1-2 μg/m^3 per year. The Taklimakan Desert

also experienced an increasing trend of $PM_{2.5}$ levels that was possibly due to the increased frequency of blowing dust events [47]. A previous study also reported increase in PM_{10} levels measured in this region after 2008 [48].

DISCUSSION

The ensemble machine-learning model developed in this study has several unique advantages when used to hindcast $PM_{2.5}$ levels. First, our results indicated that spatial clustering improved model performance for all three algorithms in this study. This is expected because the relations between $PM_{2.5}$ and its predictors would vary across our large spatial domain. By controlling unobserved spatial heterogeneity, the cluster-based models are able to capture the spatiotemporal variation in $PM_{2.5}$ more accurately than the national model. To our knowledge, this is the first data-driven method that divided China into stable regions for $PM_{2.5}$ modeling purpose. This clustering pattern is different from the so-called Heihe-Tengchong Line that divides China into two roughly equal parts with contrasting population density and economic development status [49] as many environmental factors can influence air pollution patterns. Compared with previous clustering methods, our method generated temporally stable regions that reflected geospatial heterogeneity. This clustering approach could aid modeling efforts in the future by other researchers.

Second, although we removed day-specific effects that were often used to improve model performance within the model training period, our machine learning model had similar CV performance compared with previous statistical models [10, 50]. The CV R^2 of our model was lower than some previous machine learning models that included spatial or/and temporal smooth surfaces of $PM_{2.5}$ estimated from ground measurements [9, 15]. Our model excluded measurement-based predictors because ground measurements of $PM_{2.5}$ are unavailable during the

hindcast period. Our modeling system was better suited for estimating historical $PM_{2.5}$ levels outside the model training period.

Third, to increase the robustness of our model, we included four years' worth of data for model training, whereas previous studies only used one- or two- year data, or trained a separate model for each year [9, 15, 50]. Similar to the spatial clustering, training model during a short time period, or temporal clustering, can better characterize short-term relationships. However, these annual models may estimate temporally unstable relationships that cannot be applied outside the modeling year. For example, when using only data of year 2013 to fit the XGBoost model, our model fitting R^2 increased to 0.94, but the hindcast R^2 decreased to 0.44 (Table S1). It is worth noting that the hindcast performance of the annually fitted model improved when the model-training year getting closer to 2017, the hindcast year (Table S1). This result suggested that the hindcast ability of annual model deteriorated when predicting $PM_{2.5}$ levels long before the model-training year. On the contrary, our ensemble hindcast prediction agreed well with ground measurements in 2008, five years before the model-training period. By recruiting four year' worth of data, we increased sample size and estimated temporally stable relationships. Increasing the training sample size normally leads to increased bias but decreased variance in the bias-variance trade-off [28]. This is another reason why our model had lower CV R^2 than some previous machine learning models. Similarly, to ensure a robust modeling system in space and time, we preferred low-complexity models, e.g. trees with smaller height and smaller number of leaves. We noticed that although increasing model complexity to a certain degree improved model standard CV performance, it also increased the risk of spatial and temporal overfitting (i.e., lower spatial and temporal CV R^2 values).

Finally, we combined predictions from different models that characterized different aspects of the complex relationships between $PM_{2.5}$ and various predictors. For example, the random forest and XGBoost were less prone to spatial overfitting and GAM was less prone to temporal overfitting

in cross-validation (Table 1). Additionally, while generating accurate $PM_{2.5}$ estimates, decision tree based machine learning algorithms, e.g. random forest and XGBoost, have difficulties handling spatial predictors and including time-fixed spatial parameters led to unsmooth prediction maps. By combining predictions from various algorithms, we were able to predict the spatial and temporal variations in $PM_{2.5}$ better.

We observed spatial heterogeneity in parameter importance. Visibility and MERRA-2 $PM_{2.5}$ simulations are two of the most important parameters in all clusters. Different from satellite AOD that describes vertical column aerosol loading, visibility is an indicator of horizontal aerosol loading and are associated with ground $PM_{2.5}$ concentrations [51]. MERRA-2 $PM_{2.5}$ simulations integrate data from various sources and have been shown to accurately describe large-scale $PM_{2.5}$ distributions in the U.S. and Europe [31, 32]. However, both parameters are at relatively low spatial resolutions: the visibility data was measured at ~400 stations in China and MERRA-2 simulations are at 0.625° × 0.5° resolution. The tropospheric vertical column NO_2 density and AAI from OMI also contributed significantly in $PM_{2.5}$ predictions. However, resampling of the OMI data is necessary due to a row anomaly, leading to reduced temporal resolution. Although satellite AOD retrievals were not the most important variables in random forest and XGBoost models, they provided valuable information describing the fine-resolution spatial distribution of $PM_{2.5}$ at the daily level.

One limitation of our ensemble prediction model is the underestimation of some high $PM_{2.5}$ values (e.g., monthly mean $PM_{2.5}$ levels above 200 $\mu g/m^3$, daily $PM_{2.5}$ levels above 300 $\mu g/m^3$, Figure 4), which could be attributed to the retrieval error in AOD and the relatively coarse resolution of our national model. Previous studies indicated that MODIS collection 6 AOD retrievals tend to overestimate AOD values [52, 53]. Calibrating satellite AOD against ground measurements from NASA's Aerosol Robotic Network (AERONET) may further improve the accuracy of AOD retrievals. However, there were only 10 operational AERONET stations in

China during 2013-2016 that cannot support a reliable nationwide calibration. Assuming that the quality of satellite retrievals remain constant in time, extending the study period to include more AERONET stations could support a reliable calibration of satellite AOD, therefore improve the performance of our $PM_{2.5}$ prediction models. Regarding model resolution, we constructed a national 0.1° x 0.1° grid for data integration and model fitting since the highest resolution predictors, MODIS level 2 AOD retrievals, are at a 10-km nominal resolution. Additionally, for a national model, the 0.1° grid cells revealed enough spatial variations. However, some abnormally high $PM_{2.5}$ concentrations due to local emission sources can hardly be captured at this spatial scale. As shown in Figure S7, although the residual distribution did not show any spatial patterns, suggesting that the model had no systematic bias, we observed considerable spatial variations in $PM_{2.5}$ residual within the 0.1° grid cell. As a result, the misalignment between grid level $PM_{2.5}$ predictions and point measurements may lead to underestimate of very high $PM_{2.5}$ measurements. Employing AOD products with a higher spatial resolution, e.g. MAIAC aerosol products [54, 55], and constructing a finer modeling grid could result in better model performance at high $PM_{2.5}$ levels.

CONCLUSIONS

In this study, we presented a data-driven clustering method that divided China into seven stable regions and improved model performance. We then developed a hindcast model that improved model hindcast performance to provide reliable estimations of historical $PM_{2.5}$ level. We observed that during 2008-2016, $PM_{2.5}$ levels decreased or remained constant in most part of China. In the Taklimakan Desert and Northeast China, the annual $PM_{2.5}$ levels increased. Our hindcast model could support epidemiological studies on the chronic health effects of $PM_{2.5}$ in regions without historical $PM_{2.5}$ monitoring.

ACKNOWLEDGMENTS

This work was supported by the NASA Applied Sciences Program (Grant # NNX16AQ28G, PI: Liu) and the National Institutes of Health (Grant # R01ES027892, PI: Chang).

REFERENCES

1. Brook, R.D., et al., *Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association.* Circulation, 2010. **121**(21): p. 2331-2378.
2. Hamra, G.B., et al., *Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis.* Environmental health perspectives, 2014. **122**(9): p. 906.
3. Forouzanfar, M.H., et al., *Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015.* The Lancet, 2016. **388**(10053): p. 1659-1724.
4. Tonne, C., *A call for epidemiology where the air pollution is.* The Lancet Planetary Health, 2017. **1**(9): p. e355-e356.
5. Hu, X., et al., *10-year spatial and temporal trends of PM 2.5 concentrations in the southeastern US estimated using high-resolution satellite data.* Atmospheric Chemistry and Physics, 2014. **14**(12): p. 6301-6314.
6. Ma, Z., et al., *Estimating ground-level PM2. 5 in China using satellite remote sensing.* Environmental science & technology, 2014. **48**(13): p. 7436-7444.
7. Xiao, Q., et al., *Full-coverage high-resolution daily PM2. 5 estimation using MAIAC AOD in the Yangtze River Delta of China.* Remote Sensing of Environment, 2017. **199**: p. 437-446.
8. Kloog, I., et al., *A new hybrid spatio-temporal model for estimating daily multi-year PM 2.5 concentrations across northeastern USA using high resolution aerosol optical depth data.* Atmospheric Environment, 2014. **95**: p. 581-590.
9. Di, Q., et al., *Assessing PM2. 5 exposures with high spatiotemporal resolution across the continental United States.* Environmental science & technology, 2016. **50**(9): p. 4712-4721.
10. He, Q. and B. Huang, *Satellite-based mapping of daily high-resolution ground PM 2.5 in China via space-time regression modeling.* Remote Sensing of Environment, 2018. **206**: p. 72-83.
11. Van Donkelaar, A., et al., *Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application.* Environmental health perspectives, 2010. **118**(6): p. 847.
12. Van Donkelaar, A., R.V. Martin, and R.J. Park, *Estimating ground‐level PM2. 5 using aerosol optical depth determined from satellite remote sensing.* Journal of Geophysical Research: Atmospheres, 2006. **111**(D21).
13. Geng, G., et al., *Estimating long-term PM 2.5 concentrations in China using satellite-based aerosol optical depth and a chemical transport model.* Remote Sensing of Environment, 2015. **166**: p. 262-270.

14. Hu, X., et al., *Estimating PM2. 5 Concentrations in the Conterminous United States Using the Random Forest Approach.* Environmental Science & Technology, 2017.
15. Li, T., et al., *Estimating ground‐level PM2. 5 by fusing satellite and station observations: A geo‐intelligent deep learning approach.* Geophysical Research Letters, 2017.
16. Zhan, Y., et al., *Spatiotemporal prediction of continuous daily PM 2.5 concentrations across China using a spatially explicit machine learning algorithm.* Atmospheric Environment, 2017. **155**: p. 129-139.
17. Li, T., et al., *Point-surface fusion of station measurements and satellite observations for mapping PM 2.5 distribution in China: Methods and assessment.* Atmospheric Environment, 2017. **152**: p. 477-489.
18. Kloog, I., et al., *Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal PM2. 5 exposures in the Mid-Atlantic states.* Environmental science & technology, 2012. **46**(21): p. 11913-11921.
19. Kloog, I., et al., *Estimating daily PM 2.5 and PM 10 across the complex geo-climate region of Israel using MAIAC satellite-based AOD data.* Atmospheric Environment, 2015. **122**: p. 409-416.
20. Liu, Y., et al., *A statistical model to evaluate the effectiveness of PM 2.5 emissions control during the Beijing 2008 Olympic Games.* Environment international, 2012. **44**: p. 100-105.
21. Levy, R., et al., *The Collection 6 MODIS aerosol products over land and ocean.* Atmospheric Measurement Techniques, 2013. **6**: p. 2989-3034.
22. Hsu, N., et al., *Enhanced Deep Blue aerosol retrieval algorithm: The second generation.* Journal of Geophysical Research: Atmospheres, 2013. **118**(16): p. 9296-9315.
23. Xiao, Q., et al., *Evaluation of VIIRS, GOCI, and MODIS Collection 6 AOD retrievals against ground sunphotometer observations over East Asia.* Atmos Chem Phys, 2016. **16**(3): p. 1255-1269.
24. Jinnagara Puttaswamy, S., et al., *Statistical data fusion of multi-sensor AOD over the Continental United States.* Geocarto International, 2014. **29**(1): p. 48-64.
25. Platnick, S., et al., *The MODIS cloud products: Algorithms and examples from Terra.* IEEE Transactions on Geoscience and Remote Sensing, 2003. **41**(2): p. 459-473.
26. Dobson, J.E., et al., *LandScan: a global population database for estimating populations at risk.* Photogrammetric engineering and remote sensing, 2000. **66**(7): p. 849-857.
27. Verleysen, M. and D. François. *The curse of dimensionality in data mining and time series prediction*. in *International Work-Conference on Artificial Neural Networks*. 2005. Springer.
28. Friedman, J., T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Vol. 1. 2001: Springer series in statistics New York.
29. Randles, C., et al., *The MERRA-2 aerosol reanalysis, 1980 onward. Part I: System description and data assimilation evaluation.* Journal of Climate, 2017. **30**(17): p. 6823-6850.
30. Buchard, V., et al., *The MERRA-2 aerosol reanalysis, 1980 onward. Part II: Evaluation and case studies.* Journal of Climate, 2017. **30**(17): p. 6851-6872.
31. Buchard, V., et al., *Evaluation of the surface PM 2.5 in Version 1 of the NASA MERRA Aerosol Reanalysis over the United States.* Atmospheric Environment, 2016. **125**: p. 100-111.

32. Provençal, S., et al., *Evaluation of PM surface concentrations simulated by Version 1 of NASA's MERRA Aerosol Reanalysis over Europe.* Atmospheric Pollution Research, 2017. **8**(2): p. 374-382.
33. Reid, C.E., et al., *Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning.* Environmental science & technology, 2015. **49**(6): p. 3887-3896.
34. Yanosky, J.D., et al., *Spatio-temporal modeling of particulate air pollution in the conterminous United States using geographic and meteorological predictors.* Environmental Health, 2014. **13**(1): p. 63.
35. Brunsdon, C., A.S. Fotheringham, and M.E. Charlton, *Geographically weighted regression: a method for exploring spatial nonstationarity.* Geographical analysis, 1996. **28**(4): p. 281-298.
36. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
37. Chen, T. and C. Guestrin. *Xgboost: A scalable tree boosting system*. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016. ACM.
38. Anand, T.R. and O. Renov. *Machine learning approach to identify users across their digital devices*. in *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. 2015. IEEE.
39. Mangal, A. and N. Kumar. *Using big data to enhance the bosch production line performance: A kaggle challenge*. in *Big Data (Big Data), 2016 IEEE International Conference on*. 2016. IEEE.
40. Chen, T. and T. He, *Xgboost: extreme gradient boosting.* R package version 0.4-2, 2015.
41. Li, H., et al., *Wintertime aerosol chemistry and haze evolution in an extremely polluted city of the North China Plain: significant contribution from coal and biomass combustion.* Atmospheric Chemistry and Physics, 2017. **17**(7): p. 4751-4768.
42. Huang, L., et al., *Impacts of power generation on air quality in China—part I: an overview.* Resources, Conservation and Recycling, 2017. **121**: p. 103-114.
43. Xu, W., et al., *Characteristics of pollutants and their correlation to meteorological conditions at a suburban site in the North China Plain.* Atmospheric Chemistry and Physics, 2011. **11**(9): p. 4353-4369.
44. Zhao, X., et al., *Analysis of a winter regional haze event and its formation mechanism in the North China Plain.* Atmospheric Chemistry and Physics, 2013. **13**(11): p. 5685-5696.
45. Liu, J., et al., *China's rising hydropower demand challenges water sector.* Scientific reports, 2015. **5**.
46. Guan, D., et al., *The socioeconomic drivers of China's primary PM2. 5 emissions.* Environmental Research Letters, 2014. **9**(2): p. 024010.
47. Yang, X., et al., *Spatial and temporal variations of blowing dust events in the Taklimakan Desert.* Theoretical and applied climatology, 2016. **125**(3-4): p. 669-677.
48. Zhang, X.-X., et al., *Dust deposition and ambient PM 10 concentration in northwest China: spatial and temporal variability.* Atmospheric Chemistry and Physics, 2017. **17**(3): p. 1699-1711.
49. Guo, H., et al., *Scientific big data and digital earth.* Chinese science bulletin, 2014. **59**(35): p. 5066-5073.
50. Ma, Z., et al., *Satellite-based spatiotemporal trends in PM2. 5 concentrations: China, 2004-2013.* Environmental Health Perspectives (Online), 2016. **124**(2): p. 184.
51. Liu, M., J. Bi, and Z. Ma, *Visibility-Based PM2. 5 Concentrations in China: 1957–1964 and 1973–2014.* Environmental Science & Technology, 2017. **51**(22): p. 13161-13169.

52. Fan, A., et al., *Evaluation and Comparison of Long-Term MODIS C5. 1 and C6 Products against AERONET Observations over China.* Remote Sensing, 2017. **9**(12): p. 1269.
53. de Leeuw, G., et al., *Two decades of satellite observations of AOD over mainland China using ATSR-2, AATSR and MODIS/Terra: data set evaluation and large-scale patterns.* Atmos. Chem. Phys., 2018. **18**(3): p. 1573-1592.
54. Lyapustin, A., et al., *Multiangle implementation of atmospheric correction (MAIAC): 1. Radiative transfer basis and look‐up tables.* Journal of Geophysical Research: Atmospheres, 2011. **116**(D3).
55. Lyapustin, A., et al., *Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm.* Journal of Geophysical Research: Atmospheres, 2011. **116**(D3).

Table 3. 1 The model fitting and 10-fold CV results at the daily level for individual cluster-based models, individual national models, and the ensemble model.

R^2 (RMSE ($\mu g/m^3$))	Individual Model			Ensemble Model
	XGBoost	Random Forest	GAM	
Model Fitting (cluster)	0.84 (18)	0.77 (22)	0.65 (28)	0.85 (18)
Standard CV (cluster)	0.78 (21)	0.77 (22)	0.65 (28)	0.79 (21)
Standard CV (national)	0.72 (24)	0.71 (25)	0.60 (29)	
Temporal CV (cluster)	0.71 (25)	0.72 (25)	0.65 (28)	0.73 (24)
Spatial CV (cluster)	0.74 (22)	0.75 (23)	0.58 (30)	0.76 (22)

Table 3. 2 Performance of 2017 monthly hindcast from individual models over each cluster.

	Northeast	North	YRD	Southeast	PRD	West	Northwest
Random Forest	0.74	0.83	0.82	0.82	0.83	0.84	0.87
XGBoost	0.71	0.83	0.81	0.81	0.80	0.84	0.83
GAM	0.68	0.79	0.75	0.80	0.79	0.80	0.85

Figure 3. 1 Map of the study domain with elevation.

Air quality monitors are shown as red dots and the weather stations included in the National Centers for Environmental Information (NCEI) dataset are shown as green triangles.

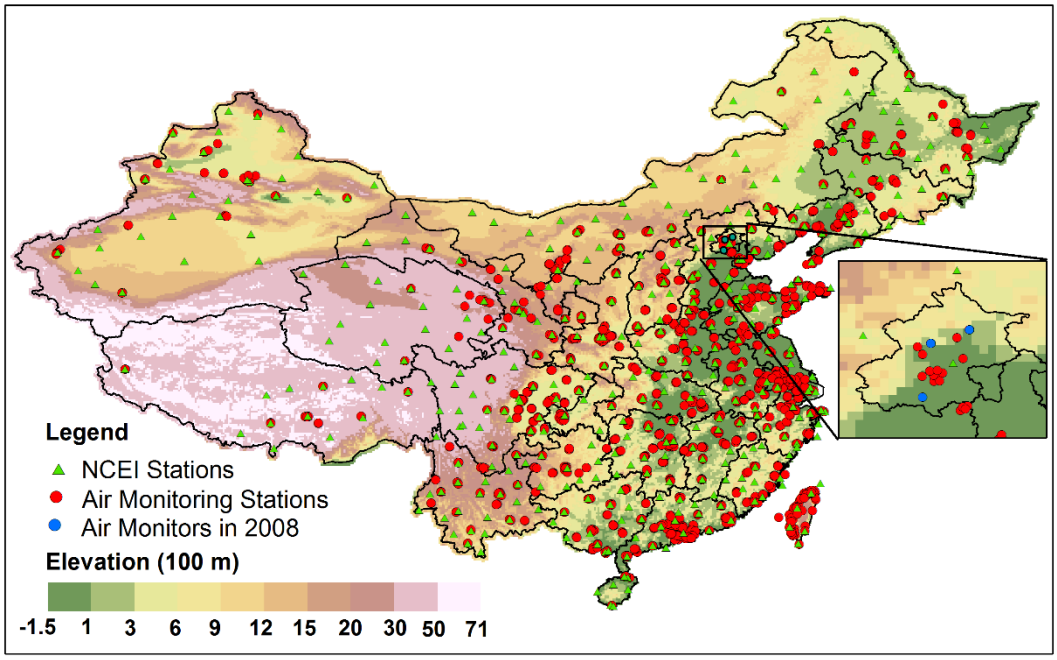


Figure 3. 2 Model structure.

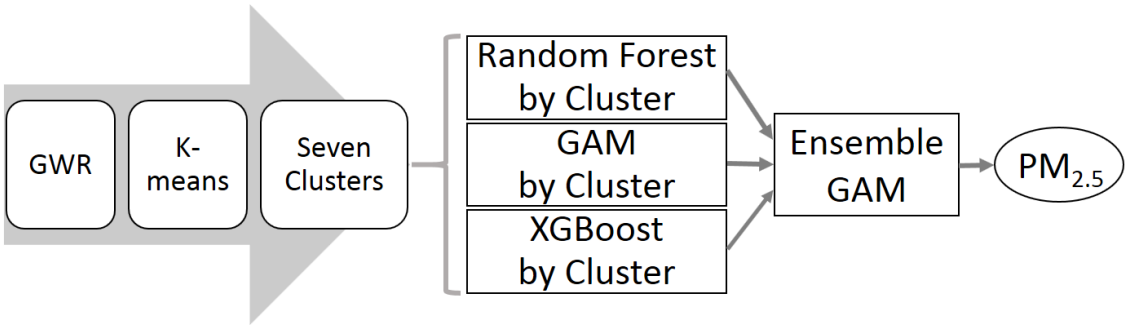


Figure 3. 3 The seven clusters covering the study domain.

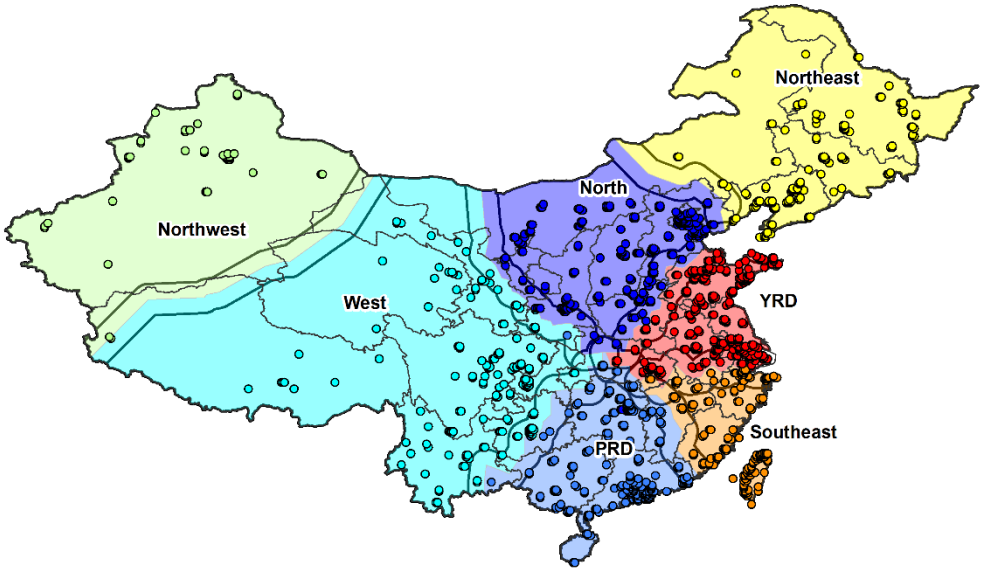


Figure 3. 4 The hindcast performance of the ensemble model in 2017 and 2008.

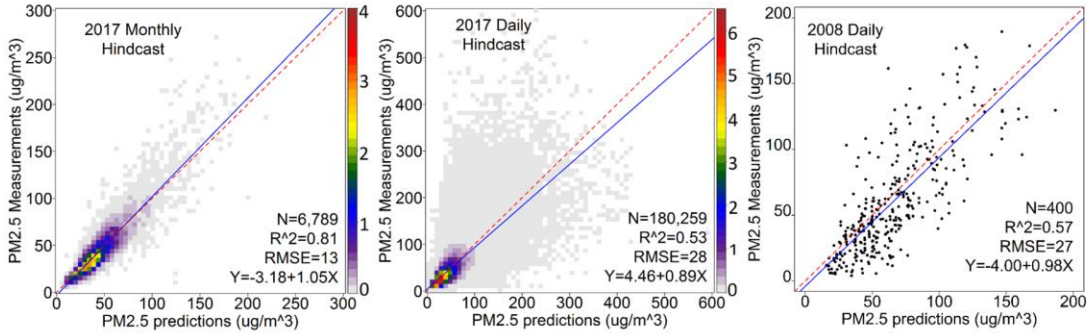
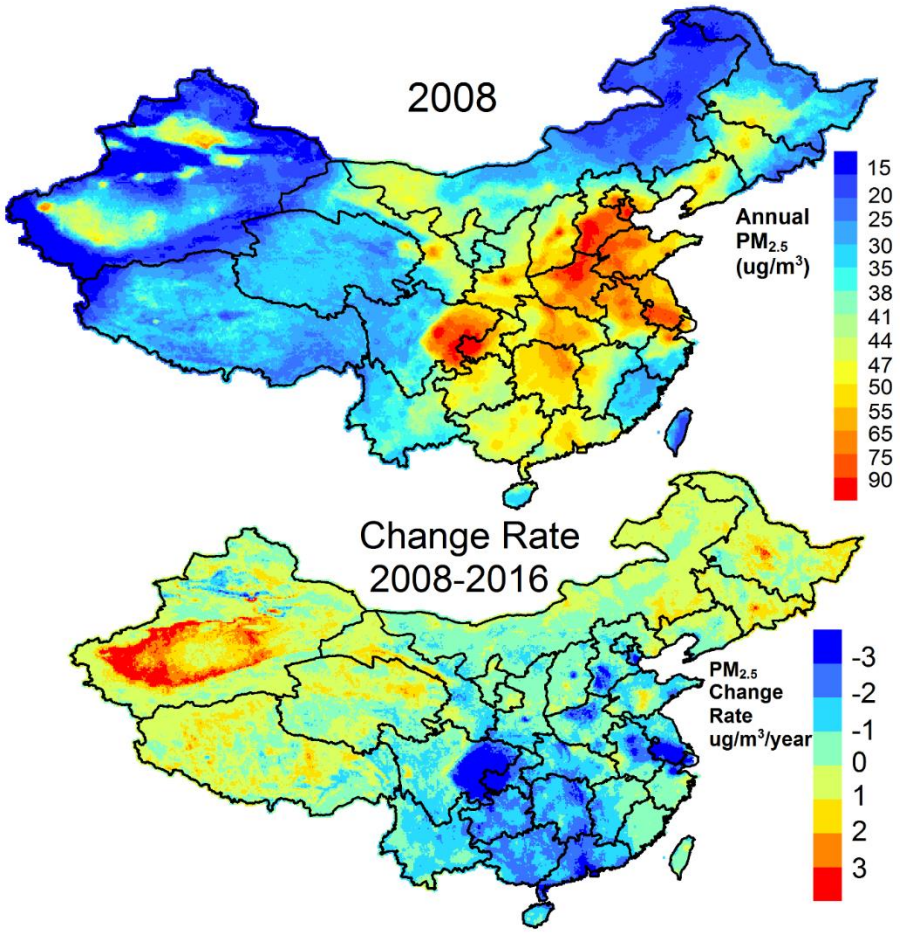


Figure 3. 5 Annual PM2.5 distribution in 2008 (above) and the estimated PM2.5 change rate during 2008-2016 (below).



SUPPLEMENTARY MATERIALS

Supplementary Text 1

The hyperparameters of random forest model were optimized by grid search. The search space was: maximum number of nodesfrom $2^8$to 2^{13}; minimum leaf node size among 10, 20, 30, 40 and 50; number of parameters for split among 4, 6, 8, 10, and 12; and number of trees among 100, 200, 300, 400, and 500. The model performance was evaluated by out-of-bag statistics and the input parameters were selected by removing the least important variables that kept the decrease in out of bag explained variance less than 1%.

The hyperparameters of XGBoost model were selected by grid search. The search space was: maximum tree depth from 6to 12; minimum child weight from 3 to 8; subsample ratio from 0.5 to 0.9; and the subsample ratio of columns from 0.5 to 0.9.

Table 3.S 1 Model fitting and hindcast performance of XGBoost model fitted separately by each year.

Data of year	Model fitting R^2	Hindcast $PM_{2.5}$ in 2017 R^2
2013	0.94	0.44
2014	0.93	0.46
2015	0.90	0.50
2016	0.90	0.54

Figure 3.S 1 Clustering results with different random seeds. Different colors represent different clusters.

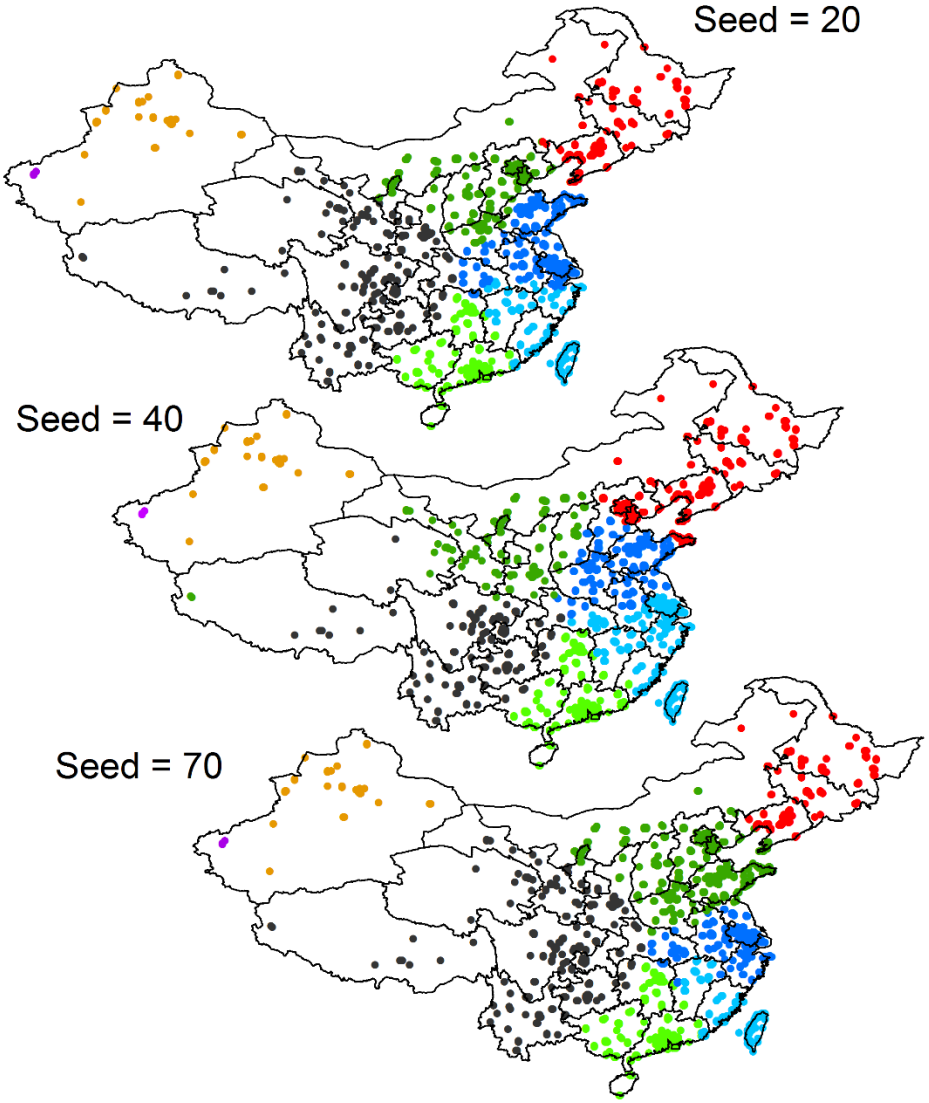


Figure 3.S 2 Clustering results by year.

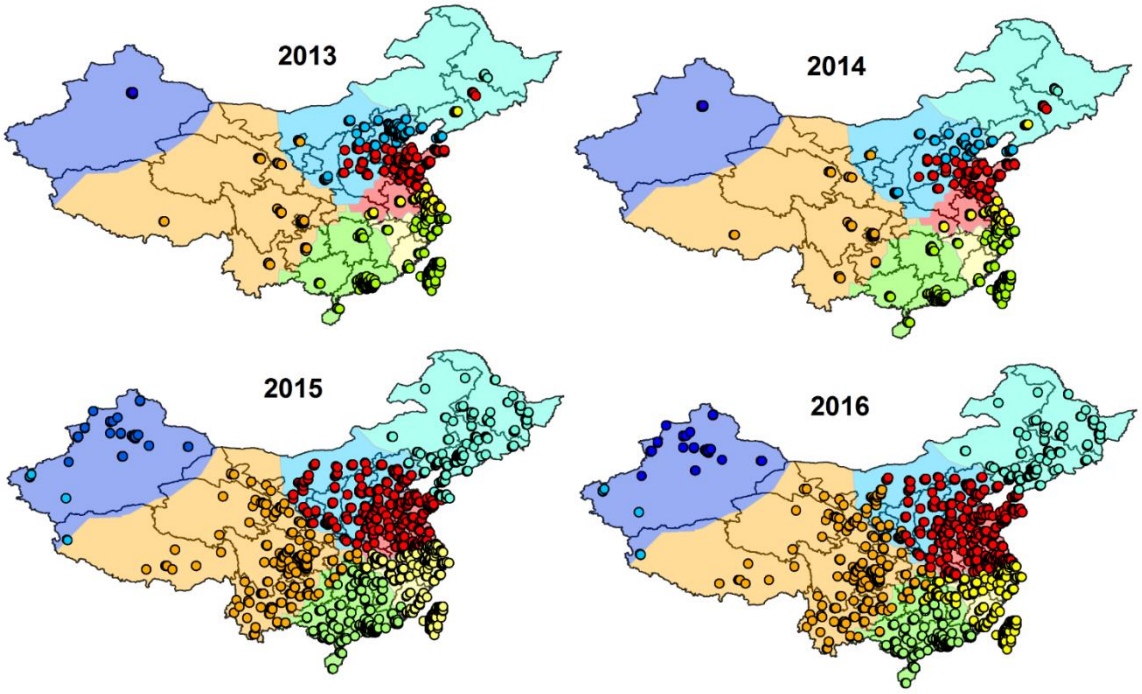


Figure 3.S 3 Scatter density plots showing the model fitting results of individual models.

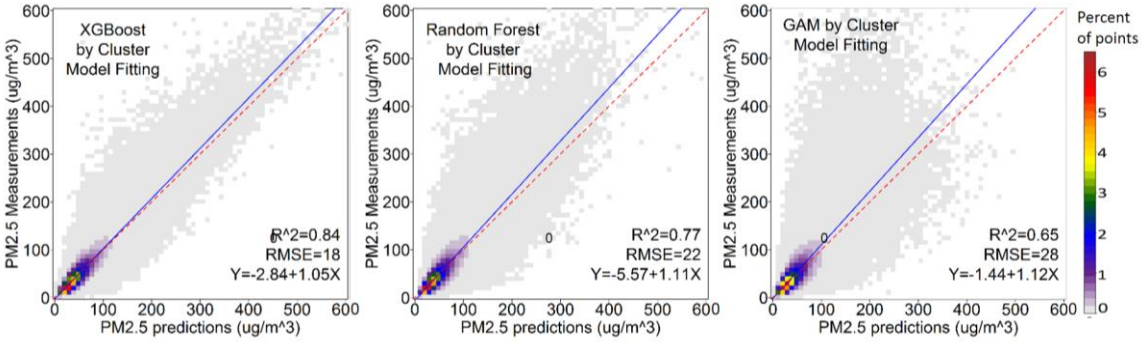


Figure 3.S 4 Scatter density plots showing the standard 10-fold CV (the first row), temporal 10-fold CV (the second row), spatial 10-fold CV (the third row) results of the individual cluster based model, as well as the standard 10-ford CV results of national models (the fourth row).

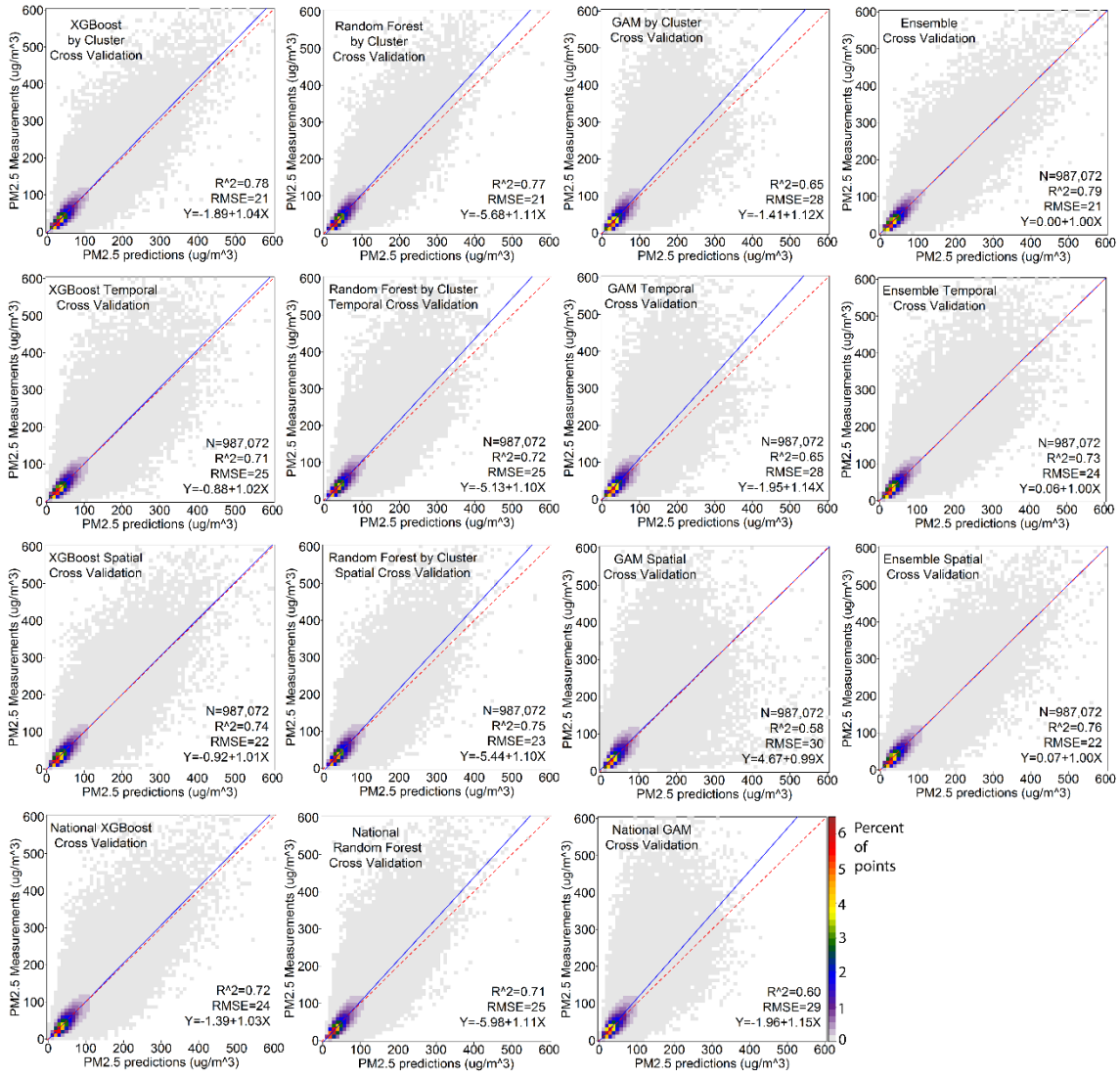


Figure 3.S 5 Hindcast performance at a daily level (above) and monthly level (below). The color scale shows the percent of points within the grid cell.

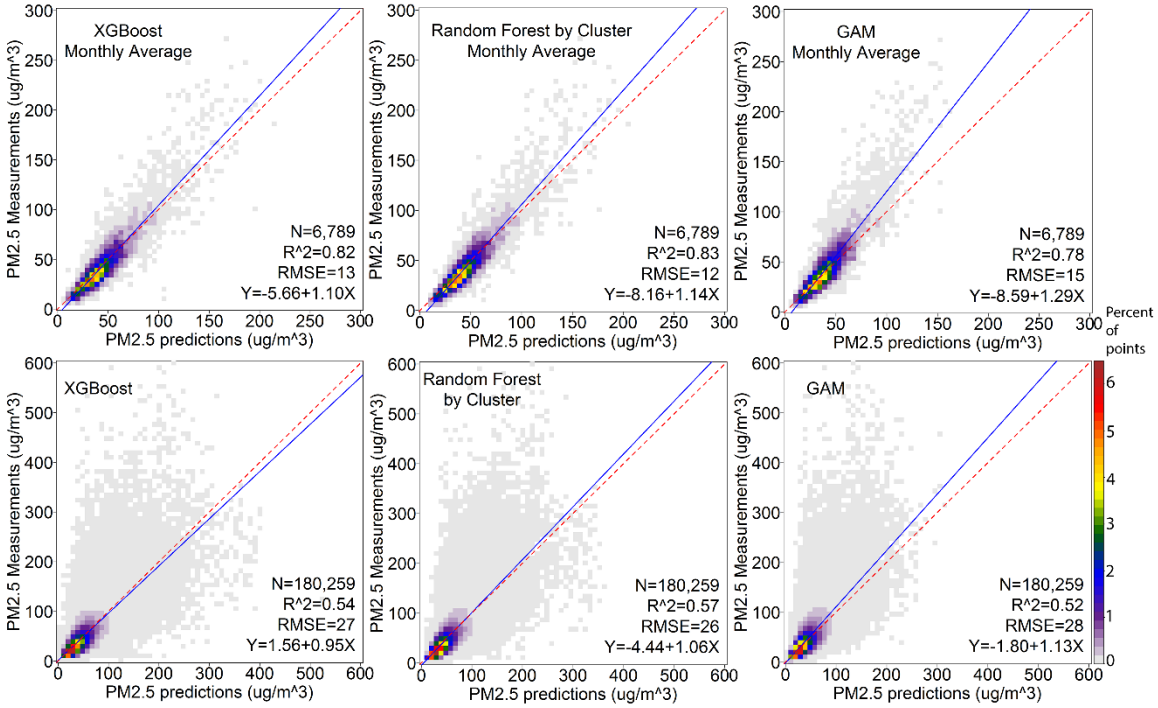


Figure 3.S 6 Annual PM2.5 distribution estimated from individual models.

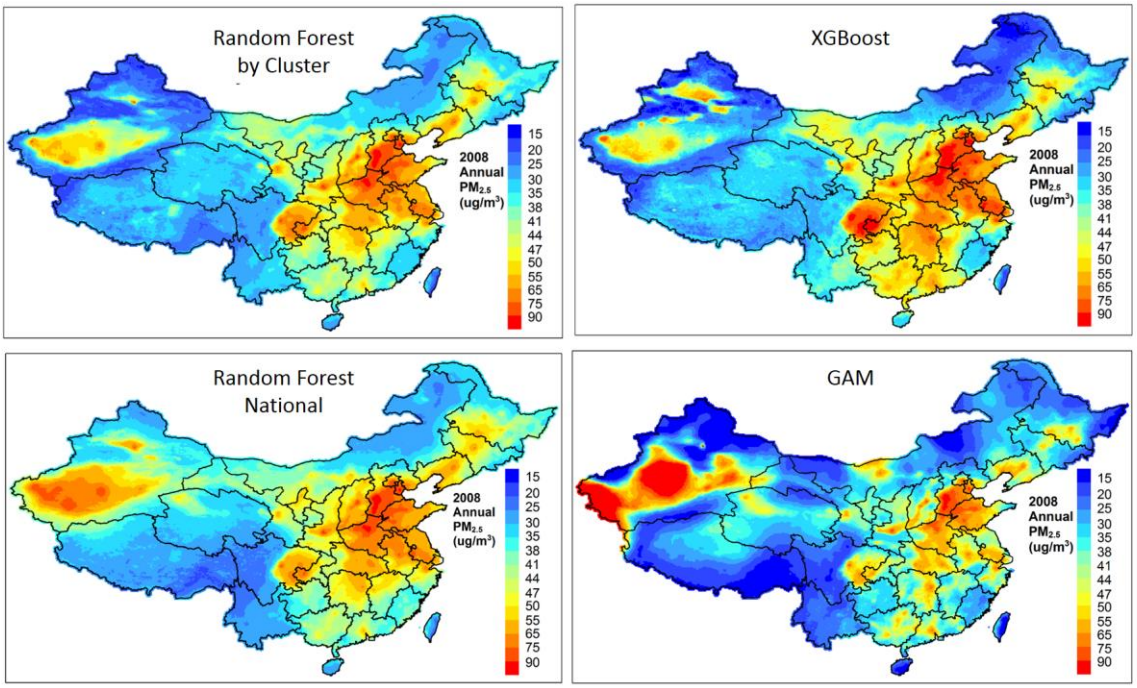
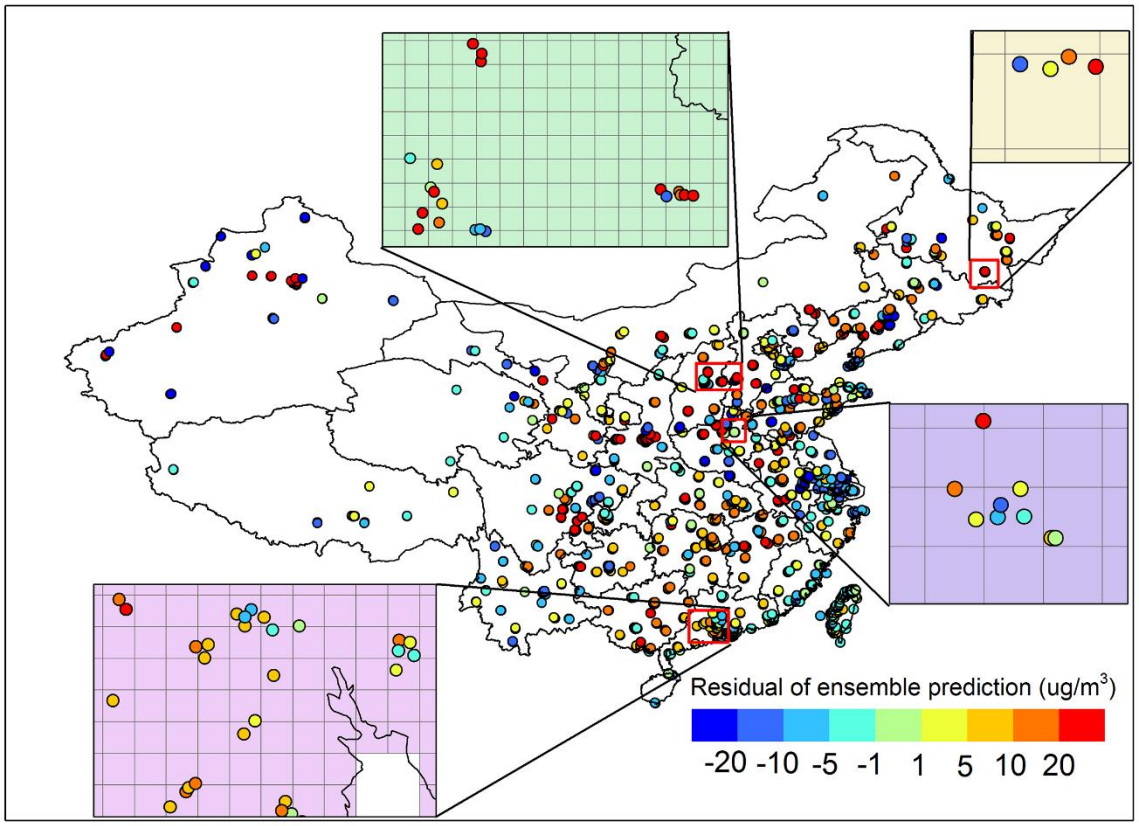


Figure 3.S 7 Average residual of ensemble predictions during 2017.



CONCLUSIONS

In most developing regions with serious air pollution, e.g. China, the routine monitoring of $PM_{2.5}$ does not existent or started recently. This lack of $PM_{2.5}$ monitoring hindered exposure assessment as well as epidemiological studies on health effects of $PM_{2.5}$ in highly polluted regions. Thus, the dose-response curve of $PM_{2.5}$ at high exposure levels is not well studied and disease burden attributable to ambient $PM_{2.5}$ in these regions are poorly estimated.

Satellite data with a long record can contribute to prediction of historical $PM_{2.5}$ concentrations, but most previous satellite data based $PM_{2.5}$ prediction models were developed in the U.S. with limited hindcast abilities. In order to predict historical $PM_{2.5}$ concentrations in developing regions, my work focus on two major challenges of $PM_{2.5}$ hindcast model: to fill non-random missing satellite data and to improve model hindcast accuracy. In Aim 1 we developed a gap-filling method that improved coverage of satellite data to 100 percent and without relying on ground measurements. Then we predicted $PM_{2.5}$ concentrations at 1-km with the gap-filled satellite data. In Aim 2, we applied satellite predictions for exposure assessment in an epidemiological study assessing associations between maternal $PM_{2.5}$ exposures and adverse birth outcomes. For Aim 3, we developed a national ensemble machine learning model in China that outperformed previous models in hindcast accuracy. The historical $PM_{2.5}$ predictions from our ensemble model could be used to assess chronic health effects of $PM_{2.5}$ in China.

Accurately estimating $PM_{2.5}$ levels in developing regions is challenging. Previous well-developed prediction models in the U.S. applied ground measurements to fill missing

data and improve model performance, but hindcast models in developing regions cannot employ the same strategy due to the lack of ground measurements. Additionally, previous models in the U.S. aimed to extend the spatial coverage of $PM_{2.5}$ monitors and ignored potential temporal overfitting that limits model's hindcast abilities when predicting historical $PM_{2.5}$ levels. Our work in Aim 1 and Aim 3 developed methods that contributed to prediction of historical $PM_{2.5}$ levels. From these work, we learned that cloud-aerosol interactions provide information on aerosol loadings and $PM_{2.5}$ concentrations. We observed spatial heterogeneity in model performance and showed that dividing a large modeling domain to appropriate smaller domains can improve model performance. These methods may motive future models in $PM_{2.5}$ prediction, $PM_{2.5}$ hindcast, and $PM_{2.5}$ forecast.

Another question that has not been well studied is the contribution of satellite-based exposure assessment in epidemiological studies. In Aim 2, we assessed maternal exposure using three datasets: $PM_{2.5}$ predictions from gap-filled satellite, $PM_{2.5}$ predictions from satellite data without accounting for missingness, and central site measurements. We reported that when aggregate exposure in time, missing satellite data led to overestimate of long-term exposure levels and attunation of health effects. We also noticed that fine-resolution satellite predictions revealed local-scale variations in exposure thus improved precision of estimated health effects. This exploratory study in one city observed higher magnitude of estimated health effects than previous findings. One reason could be the improved exposure assessment quality using fine-resolution satellite predictions.

Overall, our models are beneficial for environmental health studies in regions without historical $PM_{2.5}$ monitoring and our work met the urgent need of estimating health effects of ambient air pollution in highly polluted regions.