A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information

Gongbo Chen[a], Shanshan Li[a], Luke D. Knibbs[b], NAS Hamm[c], Wei Cao[d], Tiantian Li[e], Jianping Guo[f], Hongyan Ren[d], Michael J. Abramson[a], Yuming Guo[a,*]

[a] Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia;

[b] Department of Epidemiology and Biostatistics, School of Public Health, The University of Queensland, Brisbane, Australia;

[c] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands;

[d] Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China;

[e] National Institute of Environmental Health Sciences, Chinese Center for Disease Control and Prevention, Beijing, China;

[f] Sate Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China;

*** Corresponding author**

Y. Guo, Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Monash University. Level 2, 553 St Kilda Road, Melbourne, VIC 3004, Australia. Phone: +61 3 9905 6100. Fax: +61 3 9903 0556. Email: yuming.guo@monash.edu.

ABSTRACT

Background: Machine learning algorithms have very high predictive ability. However, no study has used machine learning to estimate historical concentrations of $PM_{2.5}$ (particulate matter with aerodynamic diameter ≤2.5 μm) at daily time scale in China at a national level.

Objectives: To estimate daily concentrations of $PM_{2.5}$ across China during 2005-2016.

Methods: Daily ground-level $PM_{2.5}$ data were obtained from 1,479 stations across China during 2014-2016. Data on aerosol optical depth (AOD), meteorological conditions and other predictors were downloaded. A random forests model (non-parametric machine learning algorithms) and two traditional regression models were developed to estimate ground-level $PM_{2.5}$ concentrations. The best-fit model was then utilized to estimate the daily concentrations of $PM_{2.5}$ across China with a resolution of 0.1 degree (≈10km) during 2005-2016.

Results: The daily random forests model showed much higher predictive accuracy than the other two traditional regression models, explaining the majority of spatial variability in daily $PM_{2.5}$ [10-fold cross-validation (CV) R^2 = 83%, root mean squared prediction error (RMSE) = 28.1 μg/m^3]. At the monthly and annual time-scale, the explained variability of average $PM_{2.5}$ increased up to 86% (RMSE=10.7 μg/m^3 and 6.9 μg/m^3, respectively).

Conclusions: Taking advantage of a novel application of modelling framework and the most recent ground-level $PM_{2.5}$ observations, the machine learning method showed higher predictive ability than previous studies.

Keywords: $PM_{2.5}$; Aerosol optical depth; Random forests; Machine learning; China

Capsule: Random forests approach could be used to estimate historical exposure to $PM_{2.5}$ in

China with high accuracy.

1 INTRODUCTION

Particulate matter (PM) is a complex mixture of solid and liquid particles suspended in the air of varying sizes, shapes, sources and composition (Jin et al., 2016; Pope and Dockery, 2006). Particle size is one characteristic of PM that is relevant to human health effects. Among different size fractions of PM, particles with aerodynamic diameter ≤ 2.5 μm ($PM_{2.5}$) attract the most scientific attention, as they are able to penetrate into the gas exchange area of the lung and potentially reach other parts of human body through the circulatory system (Feng et al., 2016).

As a consequence of rapid economic growth and urban expansion, China experiences some of the world's worst PM air pollution (Kan et al., 2009). $PM_{2.5}$ has been identified as the fourth-leading risk factor for mortality in China (Yang et al., 2013), and its associations with a range of diseases have also been reported, including respiratory and cardiovascular diseases, cancer, infectious disease and adverse birth outcomes (Chen et al., 2017b; Chen et al., 2017c; Guo et al., 2016; Lin et al., 2016; Liu et al., 2016; Liu et al., 2007). However, very few previous studies have examined the long-term health effects of $PM_{2.5}$ in China, as measurements of $PM_{2.5}$ at the national scale were not available prior to 2013. Moreover, no such study has been conducted in Western China (e.g., Tibet and Xinjiang), due to the scarcity of ground-monitoring data. To fill in the spatial gaps of ground measurements, satellite-retrieved aerosol optical depth (AOD), also known as aerosol optical thickness (AOT), has been applied to estimate ground-level $PM_{2.5}$ concentrations. This method has been increasingly employed in recent years (Chen et al., 2017a;

Hu et al., 2014c; Kloog et al., 2012; Lee et al., 2011; Ma et al., 2016; Van Donkelaar et al., 2015).

Many statistical models have been used to estimate ground-level $PM_{2.5}$ from AOD and other predictors, including multiple linear regression, generalized additive model (GAM), and mixed effects models (Gupta and Christopher, 2009; Lee et al., 2011; Liu et al., 2009). However, these regression models may not fully capture the complex relationships between $PM_{2.5}$ and a wide range of spatial and temporal predictors. Moreover, traditional regression models are restricted by some assumptions, e.g., the independence of observations and distribution of monitored $PM_{2.5}$ (Hu et al., 2017).

One approach to overcoming these limitations is machine learning, a newly developed method of data analysis that can automate statistical model development. Random forests models are non-parametric machine learning algorithms that could be used for prediction with high accuracy (Liu et al., 2018). Random forests consist of a collection of classifiers with tree structure. These classifiers are randomly and independently selected vectors with the same distribution that vote for the most popular class (Breiman, 2001). Random forests model have been successfully used for the prediction of $PM_{2.5}$ in the U.S. (Hu et al., 2017), but no study has been done at a national scale in China. In this study, we first compare the performance of the random forests approach with two traditional regression models and then estimate the spatiotemporal trends of $PM_{2.5}$ concentrations in China during 2005-2016 with satellite-retrieved AOD data, meteorological and land use information using a random forests approach.

2 METHOD AND MATERIALS

2.1 Ground-based $PM_{2.5}$ measurements

Daily ground-level measurements of $PM_{2.5}$ from May 13, 2014 through to December 31, 2016 were obtained from the China National Environmental Monitoring Center (CNEMC) (http://www.cnemc.cn/). The recently expanded network of CNEMC consists of 1,479 monitoring sites covering more than 300 cities in 31 provinces and municipalities of China. The locations of the monitoring sites are shown in Figure 1. Concentrations of $PM_{2.5}$ were measured at all sites using a Tapered Element Oscillating Microbalance (TEOM). The accuracy of daily mean concentration of $PM_{2.5}$ for this network was ±1.5 $\mu g/m^3$ (You et al., 2016). Strict quality controls were applied and abnormal values, accounting for nearly 5%, were removed (Fang et al., 2016). After data cleaning, daily mean concentrations of $PM_{2.5}$ were calculated for all stations within the network.

2.2 Satellite-retrieved AOD data

Moderate Resolution Imaging Spectroradiometer (MODIS) AOD data (Collection 6) from January 1, 2005 through to December 31, 2016 were downloaded from Level 1 and Atmosphere Archive & Distribution System of NASA (https://ladsweb.modaps.eosdis.nasa.gov/). “Deep Blue” (DB) and “Dark Target” (DT) AOD are two types daily Level-2 aerosol data from MODIS Aqua, produced at a spatial resolution of 10 km (Levy and Hsu, 2015). DB AOD shows better performance over bright areas (e.g., desert), while DT AOD works over dense and dark areas (e.g., vegetation). As neither algorithm outperforms the other consistently, a merged product of them two is recommended (Sayer et al., 2014). To improve the spatial coverage of AOD data, DB and DT AOD were combined after filling the gaps between them; where missing DB AOD, with corresponding valid DT AOD, was estimated with the linear regression model below and vice- versa (Chen et al., 2017a; Jinnagara Puttaswamy et al., 2014). Linear regressions of DB and DT AOD were fitted as follows:

$$AOD_{DB} = \beta * AOD_{DT} + \alpha$$

$$\text{or } AOD_{DT} = \beta * AOD_{DB} + \alpha$$

where AOD_{DB} and AOD_{DT} are DB and DT AOD values, respectively; β is the coefficient and α is the intercept of linear regression. In total, 25.4% and 0.1% of DT and DB AOD values were filled with the linear regressions shown above, respectively.

Ground-level observations of AOD were obtained from Aerosol Robotic Network (AERONET) of ground-based sun photometers (https://aeronet.gsfc.nasa.gov/new_web/index.html). The details of AERONET data downloading and processing are shown in the "Interpolation of AOD at 550 nm" section of the Supplementary Material. DB and DT AOD values were compared with corresponding AERONET AOD values at all AERONET monitoring sites in China. Then, combined AOD data were generated by merging DB and DT AOD using the Inverse Variance Weighting method reported previously (Ma et al., 2015). Compared to merged dark target-deep blue MODIS Collection 6 AOD product, the combined AOD data with this method showed substantial increase in spatial coverage and similar accuracy (Ma et al., 2015).

2.3 Meteorological data

Meteorological data during the study period (12 years) were obtained from 824 weather stations of China Meteorological Data Sharing Service System (http://data.cma.cn/). The distribution of all weather stations in mainland China is shown in Figure S2 in the Supplementary Material. Four meteorological variables were collected: daily mean temperature (℃), relative humidity (%), barometric pressure (kPa) and wind speed (km/h). For areas not covered by the weather stations, daily values of meteorological variables were interpolated using kriging (Diggle and Ribeiro, 2007; Furrer et al., 2009). Details of the interpolation of the meteorological variables are shown in the "Interpolation of meteorological variable" section of the Supplementary Material.

2.4 Land cover data and other predictors

Collection 5.1 annual urban cover data from 2004 to 2012 at a spatial resolution of 500 meter were downloaded from Global Mosaics of the standard MODIS land cover type data of the Global Land Cover Facility (http://glcf.umd.edu/) (Friedl et al., 2010). As 2012 urban cover is the most recent data, they were used for the estimation from 2012 through to 2016. MODIS Level 3 monthly average Normalized Difference Vegetation Index (NDVI) data at a spatial resolution of 0.1 degree (≈10 km) were downloaded from the NASA Earth Observatory (http://neo.sci.gsfc.nasa.gov/). Daily MODIS fire counts (Collection 6) during 2005-2016 were downloaded from NASA Fire Information for Resource Management System (FIRMS) (https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms/active-fire-data) (Hu et al., 2014b). The global Shuttle Radar Topography Mission (SRTM) Version 4 elevation data for China at a spatial resolution of 3 arc-seconds (approximately 90 m) were downloaded from The CGIAR Consortium for Spatial Information (http://srtm.csi.cgiar.org/).

2.5 Model development

The random forests approach generated a large number of decision trees using independent bootstrap samples of the data set. Each node of decision tree was split depending on the best among a subset of all variables which were randomly selected at that node, and then, a simple majority vote was used for prediction (Liaw and Wiener, 2002). A wide range of spatial and temporal predictors (Table S2 in the Supplementary Material) associated with $PM_{2.5}$ reported by previous studies were considered in our model development (Fang et al., 2016; Ma et al., 2015; Ma et al., 2014). All predictors were firstly included in the random forests model, and then, those included in the final model were selected according to the change in mean square error and the increase in node purities which were two variable importance measures of random forests approach. In this study, we set the thresholds of these two measures as 100 and 50000, respectively. Predictors with an increase in mean square error of less than 100 and an increase in node purities of less than 50000 were not included in the final model, as they did not improve predictive ability. The final random forests model with the best performance is shown as following;

$$PM_{2.5ij} = AOD_{ij} + TEMP_{ij} + RH_{ij} + BP_{ij} + WS_{ij} + NDVI_j + Urban\_cover_j + doy_j + \log(elev_j) \quad (1)$$

where $PM_{2.5ij}$ is the $PM_{2.5}$ on day *i* at station *j*; AOD_{ij} is the combined AOD; *TEMP*, *RH*, *BP* and *WS* are mean temperature, relative humidity, barometric pressure and wind speed on day i, respectively; *NDVI* is the monthly average NDVI value; *Urban_cover* is the percentage of urban cover with a buffer radius of 10 km; *doy* is day of the year; *log(elev)* is the log transferred elevation.

As random forests are non-parametric machine learning algorithms, we only set two parameters, the number of predictors in the random subset of each node (m_{try}) as the default value and the number of trees in the forest (n_{tree}) as 100, in the model. The selections of optimal buffer radius for percentage of urban cover and NDVI values based on median R^2 and mean square errors (*mse*). Details of these selections are shown in Tables S3 in the Supplementary Material.

In this study, we compared the performance of random forests model with traditional generalized additive model (GAM) and a non-linear exposure-lag-response model as following;

$$PM_{2.5ij} = AOD_{mij} + ns(TEMP_{ij}, 3) + ns(RH_{ij}, 3) + ns(BP_{ij}, 3) + ns(WS_{ij}, 3) + NDVI + ns(Urban\_cover, 3) + ns(doy, 8) + log(elev) \quad (2)$$

$$PM_{2.5ij} = AOD_{mij} + \mathrm{cb}\_TEMP_{ij} + \mathrm{cb}\_RH_{ij} + \mathrm{cb}\_BP_{ij} + \mathrm{cb}\_WS_{ij} + NDVI + ns(Urban\_cover, 3) + ns(doy, 8) + log(elev) \quad (3)$$

Model 2 is the GAM linking $PM_{2.5}$ and predictors. In contrast to Model 1, we fitted four meteorological variables and percentage of urban cover with natural cubic splines giving 3 degrees of freedom (df), considering their potential non-linear effects (Chen et al., 2017a) . We also fitted day of the year with a natural cubic spline giving 8 df. Model 3 is the non-linear exposure-lag-response model developed by incorporating distributed lag non-liner model (DLNM) into GAM, considering the potential lag effects of meteorological variables on $PM_{2.5}$-AOD association (Chen et al., 2018), where *cb_TEMP*, *cb_RH*, *cb_BP* and *cb_WS* are mean temperature, relative humidity, barometric pressure and wind speed on the current day and previous two days (lag 0-2 days) fitted using *crossbasis()* function of DLNM with 3 df (Gasparrini, 2011; Gasparrini, 2014), respectively. The selections of optimal df for non-linear variables, buffer radius for urban cover and maximum lag day for meteorological variables in Model 2 and Model 3 were based on adjusted R^2 and Generalized Cross Validation (GCV) value of the model. Details of these selections are shown in Tables S3-S4 in the Supplementary Material.

2.6 Validation and estimation

To evaluate the predictive ability of the models, a ten-fold cross-validation (CV) was performed with ground measurements of $PM_{2.5}$ during 2014-2016 by randomly selecting 148 (10% of total) stations as the validation set and the rest of the stations as the training set. This process was repeated 200 times. The overall adjusted R^2, Root Mean Square Error (RMSE), regression slope and coefficients were calculated.

A grid with a resolution of 0.1 degree (≈10 km) covering the entirety of China was created. In total, 96103 grid cells were included. Data on predictors included in the final model were integrated into the grid and they were linked by location and calendar date for each grid cell. Mean values of AOD and land cover variables were calculated where multiple values fell within one grid cell. The final random forests model, based on ground measured $PM_{2.5}$ during 2014-2016, was then used to estimate the daily concentrations of $PM_{2.5}$ for all grid cells during 2005-2016. Because no historical measurement data were available to validate these predictions, we thus assumed the relationship between $PM_{2.5}$ and its predictors observed for 2014-16 held true back to 2005. As no ground measured data were available in Taiwan, we did the estimation in Taiwan using the model built for Fujian province, which is the nearest province to Taiwan in mainland China. Daily results of estimation were aggregated into monthly and seasonal averages. Considering the regional variations of $PM_{2.5}$-AOD associations (Zhang et al., 2009), models were developed and the predictions were performed by each province separately.

To investigate the trends of estimated $PM_{2.5}$ over time, linear regressions of annual mean $PM_{2.5}$ and calendar year were fitted for each grid cell. Coefficients of calendar year were extracted to indicate the change of $PM_{2.5}$ over time. Positive coefficients indicated increase in $PM_{2.5}$ over time and negative coefficients indicated decrease in $PM_{2.5}$

3 RESULTS

Means of daily concentrations of $PM_{2.5}$ at 1,479 ground monitoring stations during 2014-2016 are shown in Figure 1. Overall, the mean concentration of $PM_{2.5}$ in China was 50.1 μg/m^3. The mean value of combined AOD was 0.6. The largest concentrations of ground-level measured $PM_{2.5}$ (≥ 85 μg/m^3) were observed in the south of Hebei, the north of Henan and western remote areas of Xinjiang, while the lowest levels (< 25 μg/m^3) were present in the southwestern areas of China, such as Hainan, Yunnan and Tibet. A summary of ground measurements of $PM_{2.5}$ in each province is shown in Table S5 in the Supplementary Material.

The variable importance measures of all predictors are shown in Table S2 in the Supplementary Material. In total, 12 predictors were considered in the model development stage and 9 of them were included in the final random forests model. Day of the year, AOD and daily temperature were the top three important predictors. The results of 10-fold cross-validation at the national scale in China are shown in Figure 2. These showed that daily model explained most of the variability in ground measured $PM_{2.5}$ (CV R^2=83%, RMSE=18.0 μg/m^3). Aggregated into monthly and seasonal average, the model explained 86% (RMSE=10.7 μg/m^3 and 6.9 μg/m^3, respectively) of variability in $PM_{2.5}$, respectively. Daily GAM and non-linear exposure-lag-response model showed similar predictive abilities. They explained 55% (RMSE=29.1 μg/m^3) and 51% (RMSE=30.3 μg/m^3) of $PM_{2.5}$ variability, respectively. Daily random forests model had much higher CV R^2 and lower RMSE than GAM and non-linear exposure-lag-response model.

Table 1 shows the results of 10-fold cross-validation in each province of China. The random forests model had highest CV R^2 in provinces in Northern China (e.g., Hebei, Beijing and Tianjin), while the lowest CV R^2 in Western China (e.g., Tibet, Qinghai and Yunnan). On average, the CV R^2 of daily random forests model was 30% higher than that of GAM and non-linear exposure-lag-response model.

Thus, daily concentrations of $PM_{2.5}$ across China were estimated with random forests model rather than GAM or non-linear exposure-lag-response model. Figure 3 shows the estimated mean concentrations of $PM_{2.5}$ across China during 2005-2016. The highest levels of $PM_{2.5}$ (>85 μg/m^3) were observed in North China Plain (central and southern areas of Hebei). Apart from Hebei, severe $PM_{2.5}$ pollution were also present in Shandong, Henan, Yangtze River Delta, Sichuan Basin and Taklimakan Desert of Xinjiang. The lowest levels of $PM_{2.5}$ (<25 μg/m^3) were observed in south-western and northern remote areas of China, including Yunnan, Tibet and Inner Mongolia.

Figure 4 shows the seasonal patterns of estimated $PM_{2.5}$ across China. Levels of $PM_{2.5}$ in the entire China were the highest in winter (mean $PM_{2.5}$ = 40.6 μg/m^3) while lowest in summer (mean $PM_{2.5}$ = 21.6 μg/m^3). In spring and autumn, levels of $PM_{2.5}$ were similar (Mean $PM_{2.5}$ = 31.0 μg/m^3 and 29.1 μg/m^3, respectively).

Figure 5 illustrates the time trends of estimated $PM_{2.5}$ during the study period. Overall, modest changes of $PM_{2.5}$ were observed in China during 2005-2016. Increasing trends of $PM_{2.5}$ were present in Beijing-Tianjin-Hebei region and Yangtze River Delta, while decreasing trends were present in the Pearl River Delta. When divided the whole study period into three 4-year periods, substantial increases in $PM_{2.5}$ were observed in most parts of China during 2005-2008, while the concentrations decreased during the following 8 years (2009-2016).

4 DISCUSSION

In this study, a random forests model was developed to estimate $PM_{2.5}$ in China with MODIS AOD data, meteorological and land use information. The model showed much higher predictive ability than two traditional regression models. It was then used to estimate concentrations of $PM_{2.5}$ across China during 2005-2016. According to our estimates, the highest levels of $PM_{2.5}$ were observed in Southern Hebei, while the lowest levels were present in South-Western and Northern China in remote areas. Overall, levels of $PM_{2.5}$ in China peaked in 2008 and decreased from that year on.

Several previous studies have attempted to estimate $PM_{2.5}$ in China. Ma et al. (2015) analyzed the spatial and temporal trends of $PM_{2.5}$ in China during 2004-2013 with satellite-retrieved estimation (Ma et al., 2015). The CV R^2 for daily model, monthly average and seasonal average were 41%, 73% and 79%, respectively. Fang at al. (2016) estimated the annual concentrations of $PM_{2.5}$ across China from June 2013 through to May 2014 (Fang et al., 2016). The CV R^2 was 80%. Wei et al. (2016) estimated levels of $PM_{2.5}$ in China in 2013 and compared satellite-based models with different AOD products (You et al., 2016). The CV R^2s for annual estimation were 76% for MODIS AOD and 81% for MISR AOD. Our prediction with the random forests approach showed higher accuracy than those studies.

In contrast to previous studies, we employed non-parametric machine learning algorithms to estimate daily concentrations of $PM_{2.5}$ across China. Our study is consistent with previous studies showing advantages in prediction compared traditional regression models (Brokamp et al., 2017; Were et al., 2015). The injection of randomness (bagging and random features) contributes to substantial increase in accuracy of classification and regression, which makes this method robust to noise (Breiman, 2001). This method is user-friendly, as there is no need to define the complex relationships between predictors (e.g., linear or nonlinear relationships and interactions) and the variable importance measures provided by random forests help user to identify important variables and noise variables (Liaw and Wiener, 2002). Finally, this method makes full use of the strength of each predictor and their correlations and it is robust to overfitting (Breiman, 2001). The random forest approach used in this study showed comparable predictive abilities to other neural network approach and machine learning algorithms (Di et al., 2016; Reid et al., 2015), but it was more user-friendly. Apart from the different methods we used, we also had the ability to incorporate the most recent ground-level measured $PM_{2.5}$ data, which led to substantial improvements in spatial coverage across China. Compared with previous ground monitoring network of CNEMC, the current one has expanded from 943 to 1,479 monitoring stations in mainland China. Most of the new stations are located in Western and Central China, rather than coastal areas of South-Eastern China. The locations of the new stations are shown in Figure S3 in the Supplementary Material. In the previous CNEMC network, many fewer stations were available in Western China, where lower levels of $PM_{2.5}$ air pollution were observed, than Eastern China (Zhang et al., 2016). Thus, in-situ $PM_{2.5}$ data obtained from the expanded CNEMC network are likely to be better-suited to capturing overall population exposures to $PM_{2.5}$ air pollution in China.

Other land-use variables (forest cover and water cover) and population data were used by previous studies for model development (Fang et al., 2016; Ma et al., 2015; Ma et al., 2014). Compared to the annual land cover data available during 2005-2012, the NDVI data used in our model are monthly data available over the whole study period, which can capture more variability in $PM_{2.5}$. We found adding water cover data did not improve the final model, as most of monitoring stations are located in city areas with no water areas nearby. We did not add population data in our model, considering it would be highly correlated with urban cover data in our study.

The North China Plain has been identified as area with the heaviest PM air pollution in China (Wang et al., 2015). Its severe air pollution has been attributed to the dense local steel and power industries, and the air quality has also been affected by surrounding provinces including Henan and Shandong (Wang et al., 2014). The high level of $PM_{2.5}$ in Sichuan Basin was not only associated with the rapid economic growth and urbanization but also the unique local topography (Li et al., 2015a). The climate of the Sichuan Basin is characterized with low wind speed and high humidity, which does not facilitate the dispersion of air pollutants.

The time trends of $PM_{2.5}$ in China illustrated in this study are consistent with a previous study that the peak of $PM_{2.5}$ occurred in 2008 and kept declining after wards (Ma et al., 2015). The Chinese government took a series of strict measures to control air quality during the Beijing

Olympic Games in 2008, and the subsequent benefits of these actions have been reported by many studies (Li et al., 2016). After Beijing Olympic Games, China took further measures to control air pollution. For example, the goal of preventing and controlling air pollution was included in the 12th National Five-Year Plan and the first National Action Plan on Air Pollution and Control was released in 2013 (Chen et al., 2013).

Based on historical levels of $PM_{2.5}$ estimated in this study, it could be inferred that China has made considerable progress in air quality control via strict legislation, regulation and enforcement over a relatively short period of time (Li et al., 2016). However, challenges remain to meet the goal of clean air (Wang and Hao, 2012). Currently, more than 90% of the Chinese population are experiencing unhealthy air according to US EPA standard (Rohde and Muller, 2015). In most parts of China, levels of $PM_{2.5}$ far exceed the WHO standard (Jindal, 2007; Zhang et al., 2016). Air pollution is even more severe in mega cities of China characterized with dense industries and population, such as Beijing, Tianjin, Shanghai, and Chongqing (Chan and Yao, 2008).

There are some limitations in our study. Like some of the previous studies (Hu et al., 2014a; Li et al., 2015b; Ma et al., 2015), we estimated the historical levels of $PM_{2.5}$ air pollution in China based on the $PM_{2.5}$-AOD association. However, due to unavailability of ground measuring data, we could not validate the $PM_{2.5}$-AOD association before 2014. Our historical estimates should be interpreted with due caution for that reason. To account for the spatial variations of $PM_{2.5}$-AOD associations, $PM_{2.5}$ was first predicted at the provincial level and then combined into the national level. The drawback of this approach leads to discontinuities at some provincial boundaries. Finally, due to cloud cover, missing values of AOD are problematic and could be highly prevalent in some seasons and regions (Just et al., 2015).

5 CONCLUSIONS

Novel statistical models with high accuracy and reliability were developed to estimate $PM_{2.5}$ concentrations. Taking advantage of the most recent in-situ $PM_{2.5}$ data and expanded network, many more ground measurements of $PM_{2.5}$ were available in central and western China, making our estimates more representative of the overall historical level of $PM_{2.5}$ air pollution in China. The results of this study could help to evaluate the long-term effects of $PM_{2.5}$ air pollution and disease burden attributed to $PM_{2.5}$ exposures. The study could also provide valuable information and evidence for the future prevention and control of air pollution in China.

Acknowledgements

YG was supported by a Career Development Fellowship of Australian National Health and Medical Research Council (NHMRC #APP1107107). SL was supported by an Early Career Fellowship of NHMRC (#APP1109193) and Seed Funding from the NHMRC Centre of Research Excellence–Centre for Air quality and health Research and evaluation (APP1030259). GC was supported by China Scholarship Council (CSC). L.D.K. was partly supported by the NHMRC Centre of Research Excellence–Centre for Air quality and health Research and evaluation (#APP1030259).

Conflict of interests

The authors have declared that no competing interests exist.

Reference:

Breiman, L., 2001. Random forests. Machine learning 45, 5-32.

Brokamp, C., Jandarov, R., Rao, M., LeMasters, G., Ryan, P., 2017. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. Atmospheric Environment 151, 1-11.

Chan, C.K., Yao, X., 2008. Air pollution in mega cities in China. Atmospheric Environment 42, 1-42.

Chen, G., Knibbs, L.D., Zhang, W., Li, S., Cao, W., Guo, J., Ren, H., Wang, B., Wang, H., Williams, G., Hamm, N.A.S., Guo, Y., 2017a. Estimating spatiotemporal distribution of PM1 concentrations in China with satellite remote sensing, meteorology, and land use information. Environmental Pollution (2017), https://doi.org/10.1016/j.envpol.2017.10.011.

Chen, G., Zhang, W., Li, S., Williams, G., Liu, C., Morgan, G.G., Jaakkola, J.J., Guo, Y., 2017b. Is short-term exposure to ambient fine particles associated with measles incidence in China? A multi-city study. Environ Res 156, 306-311.

Chen, G., Zhang, W., Li, S., Zhang, Y., Williams, G., Huxley, R., Ren, H., Cao, W., Guo, Y., 2017c. The impact of ambient fine particles on influenza transmission and the modification effects of temperature in China: a multi-city study. Environ Int 98, 82-88.

Chen, Z.-Y., Zhang, T.-H., Zhang, R., Zhu, Z.-M., Ou, C.-Q., Guo, Y., 2018. Estimating PM2. 5 concentrations based on non-linear exposure-lag-response associations with aerosol optical depth and meteorological measures. Atmospheric Environment 173, 30-37.

Chen, Z., Wang, J.-N., Ma, G.-X., Zhang, Y.-S., 2013. China tackles the health effects of air pollution. The Lancet 382, 1959-1960.

Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., Schwartz, J., 2016. Assessing PM2. 5 exposures with high spatiotemporal resolution across the continental United States. Environ Sci Technol 50, 4712-4721.

Diggle, P.J., Ribeiro, P.J., 2007. An overview of model-based geostatistics. Model-based Geostatistics, 27-45.

Fang, X., Zou, B., Liu, X., Sternberg, T., Zhai, L., 2016. Satellite-based ground PM 2.5 estimation using timely structure adaptive modeling. Remote Sensing of Environment 186, 152-163.

Feng, S., Gao, D., Liao, F., Zhou, F., Wang, X., 2016. The health effects of ambient PM2.5 and potential mechanisms. Ecotoxicol Environ Saf 128, 67-74.

Friedl, M.A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., Huang, X., 2010. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. Remote Sensing of Environment 114, 168-182.

Furrer, R., Nychka, D., Sain, S., Nychka, M.D., 2009. Package ‘fields’. R Foundation for Statistical Computing, Vienna, Austria. http://www. idg. pl/mirrors/CRAN/web/packages/fields/fields. pdf (last accessed 22 December 2012).

Gasparrini, A., 2011. Distributed lag linear and non-linear models in R: the package dlnm. Journal of statistical software 43, 1.

Gasparrini, A., 2014. Modeling exposure-lag-response associations with distributed lag non-linear models. Stat Med 33, 881-899.

Guo, Y., Zeng, H., Zheng, R., Li, S., Barnett, A.G., Zhang, S., Zou, X., Huxley, R., Chen, W., Williams, G., 2016. The association between lung cancer incidence and ambient air pollution in China: A spatiotemporal analysis. Environ Res 144, 60-65.

Gupta, P., Christopher, S.A., 2009. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. Journal of Geophysical Research: Atmospheres 114.

Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L., Strickland, M., Liu, Y., 2017. Estimating PM2. 5 Concentrations in the Conterminous United States Using the Random Forest Approach. Environ Sci Technol.

Hu, X., Waller, L.A., Lyapustin, A., Wang, Y., Liu, Y., 2014a. 10-year spatial and temporal trends of PM2.5 concentrations in the southeastern US estimated using high-resolution satellite data. Atmospheric Chemistry and Physics 14, 6301-6314.

Hu, X., Waller, L.A., Lyapustin, A., Wang, Y., Liu, Y., 2014b. Improving satellite -driven PM2. 5 models with Moderate Resolution Imaging Spectroradiometer fire counts in the southeastern US. Journal of Geophysical Research: Atmospheres 119.

Hu, X.F., Waller, L.A., Lyapustin, A., Wang, Y.J., Al-Hamdan, M.Z., Crosson, W.L., Estes, M.G., Estes, S.M., Quattrochi, D.A., Puttaswamy, S.J., Liu, Y., 2014c. Estimating ground-level PM2.5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. Remote Sensing of Environment 140, 220-232.

Jin, L., Luo, X., Fu, P., Li, X., 2016. Airborne particulate matter pollution in urban China: A chemical mixture perspective from sources to impacts. National Science Review, nww079.

Jindal, S., 2007. Air quality guidelines: Global update 2005, Particulate matter, ozone, nitrogen dioxide and sulfur dioxide. Indian Journal of Medical Research 126, 492-494.

Jinnagara Puttaswamy, S., Nguyen, H.M., Braverman, A., Hu, X., Liu, Y., 2014. Statistical data fusion of multi-sensor AOD over the continental United States. Geocarto International 29, 48-64.

Just, A.C., Wright, R.O., Schwartz, J., Coull, B.A., Baccarelli, A.A., Tellez-Rojo, M.M., Moody, E., Wang, Y., Lyapustin, A., Kloog, I., 2015. Using high-resolution satellite aerosol optical depth to estimate daily PM2. 5 geographical distribution in Mexico City. Environ Sci Technol 49, 8576-8584.

Kan, H., Chen, B., Hong, C., 2009. Health impact of outdoor air pollution in China: current knowledge and future research needs. Environ Health Perspect 117, A187.

Kloog, I., Nordio, F., Coull, B.A., Schwartz, J., 2012. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal PM2.5 exposures in the Mid-Atlantic states. Environ Sci Technol 46, 11913-11921.

Lee, H., Liu, Y., Coull, B., Schwartz, J., Koutrakis, P., 2011. A novel calibration approach of MODIS AOD data to predict PM2. 5 concentrations. Atmospheric Chemistry and Physics 11, 7991.

Levy, R., Hsu, C., 2015. MODIS Atmosphere L2 Aerosol Product, NASA MODIS Adaptive Processing System. Goddard Space Flight Center, USA, doi 10.

Li, S., Williams, G., Guo, Y., 2016. Health benefits from improved outdoor air quality and intervention in China. Environmental Pollution 214, 17-25.

Li, Y., Chen, Q.L., Zhao, H.J., Wang, L., Tao, R., 2015a. Variations in PM10, PM2.5 and PM1.0 in an Urban Area of the Sichuan Basin and Their Relation to Meteorological Factors. Atmosphere 6, 150-163.

Li, Y., Lin, C., Lau, A.K., Liao, C., Zhang, Y., Zeng, W., Li, C., Fung, J.C., Tse, T.K., 2015b. Assessing long-term trend of particulate matter pollution in the Pearl River Delta region using satellite remote sensing. Environ Sci Technol 49, 11670-11678.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2, 18-22.

Lin, H., Tao, J., Du, Y., Liu, T., Qian, Z., Tian, L., Di, Q., Rutherford, S., Guo, L., Zeng, W., 2016. Particle size and chemical constituents of ambient particulate pollution associated with cardiovascular mortality in Guangzhou, China. Environmental Pollution 208, 758-766.

Liu, P., Wang, X., Fan, J., Xiao, W., Wang, Y., 2016. Effects of Air Pollution on Hospital Emergency Room Visits for Respiratory Diseases: Urban-Suburban Differences in Eastern China. Int J Environ Res Public Health 13.

Liu, S., Krewski, D., Shi, Y., Chen, Y., Burnett, R.T., 2007. Association between maternal exposure to ambient air pollutants during pregnancy and fetal growth restriction. Journal of Exposure Science and Environmental Epidemiology 17, 426.

Liu, Y., Cao, G., Zhao, N., Mulligan, K., Ye, X., 2018. Improve ground-level PM 2.5 concentration mapping using a random forests-based geostatistical approach. Environmental Pollution 235, 272-282.

Liu, Y., Paciorek, C.J., Koutrakis, P., 2009. Estimating regional spatial and temporal variability of PM2. 5 concentrations using satellite data, meteorology, and land use information. Environ Health Perspect 117, 886.

Ma, Z., Hu, X., Sayer, A.M., Levy, R., Zhang, Q., Xue, Y., Tong, S., Bi, J., Huang, L., Liu, Y., 2015. Satellite-Based Spatiotemporal Trends in PM Concentrations: China, 2004-2013. Environ Health Perspect.

Ma, Z., Hu, X., Sayer, A.M., Levy, R., Zhang, Q., Xue, Y., Tong, S., Bi, J., Huang, L., Liu, Y., 2016. Satellite-based spatiotemporal trends in PM2. 5 concentrations: China, 2004–2013. Environ Health Perspect 124, 184.

Ma, Z.W., Hu, X.F., Huang, L., Bi, J., Liu, Y., 2014. Estimating Ground-Level PM2.5 in China Using Satellite Remote Sensing. Environ Sci Technol 48, 7436-7444.

Pope, C.A., Dockery, D.W., 2006. Health effects of fine particulate air pollution: Lines that connect. Journal of the Air & Waste Management Association 56, 709-742.

Reid, C.E., Jerrett, M., Petersen, M.L., Pfister, G.G., Morefield, P.E., Tager, I.B., Raffuse, S.M., Balmes, J.R., 2015. Spatiotemporal prediction of fine particulate matter during the 2008 Northern California wildfires using machine learning. Environ Sci Technol 49, 3887-3896.

Rohde, R.A., Muller, R.A., 2015. Air Pollution in China: Mapping of Concentrations and Sources. PLoS One 10, e0135749.

Sayer, A., Munchak, L., Hsu, N., Levy, R., Bettenhausen, C., Jeong, M.J., 2014. MODIS Collection 6 aerosol products: Comparison between Aqua's e -Deep Blue, Da usage recommendations. Journal of Geophysical Research: Atmospheres 119.

Van Donkelaar, A., Martin, R.V., Brauer, M., Boys, B.L., 2015. Use of satellite observations for long-term exposure assessment of global concentrations of fine particulate matter. Environ Health Perspect 123, 135.

Wang, L., Wei, Z., Yang, J., Zhang, Y., Zhang, F., Su, J., Meng, C., Zhang, Q., 2014. The 2013 severe haze over southern Hebei, China: model evaluation, source apportionment, and policy implications. Atmospheric Chemistry and Physics 14, 3151-3173.

Wang, S., Hao, J., 2012. Air quality management in China: Issues, challenges, and options. Journal of Environmental Sciences 24, 2-13.

Wang, Y.Q., Zhang, X.Y., Sun, J.Y., Zhang, X.C., Che, H.Z., Li, Y., 2015. Spatial and temporal variations of the concentrations of PM10, PM2.5 and PM1 in China. Atmospheric Chemistry and Physics 15, 13585-13598.

Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators 52, 394-403.

Yang, G., Wang, Y., Zeng, Y., Gao, G.F., Liang, X., Zhou, M., Wan, X., Yu, S., Jiang, Y., Naghavi, M., Vos, T., Wang, H., Lopez, A.D., Murray, C.J.L., 2013. Rapid health transition in China, 1990–2010: findings from the Global Burden of Disease Study 2010. The Lancet 381, 1987-2015.

You, W., Zang, Z., Zhang, L., Li, Y., Wang, W., 2016. Estimating national-scale ground-level PM25 concentration in China using geographically weighted regression based on MODIS and MISR AOD. Environmental Science and Pollution Research 23, 8327-8338.

Zhang, H., Hoff, R.M., Engel-Cox, J.A., 2009. The relation between Moderate Resolution Imaging Spectroradiometer (MODIS) aerosol optical depth and PM2. 5 over the United States: a geographical comparison by US Environmental Protection Agency regions. Journal of the Air & Waste Management Association 59, 1358-1369.

Zhang, T., Liu, G., Zhu, Z., Gong, W., Ji, Y., Huang, Y., 2016. Real-time estimation of satellite-derived PM2. 5 based on a semi-physical geographically weighted regression model. Int J Environ Res Public Health 13, 974.

Table 1. The results of 10-fold cross-validation in each province of China

Province	Random forests model		GAM		Non-linear exposure-lag-response model	
	CV R^2	RMSE	CV R^2	RMSE	CV R^2	RMSE
Hebei	90%	20.7	60%	30.7	54%	34.4
Beijing	90%	19.6	66%	27.7	60%	30.4
Tianjin	88%	20.4	60%	25.7	49%	29.1
Henan	86%	19.2	52%	22.4	46%	23.7
Hubei	86%	14.6	60%	13.3	55%	14.5
Jilin	86%	15.5	44%	17.4	45%	18.2
Sichuan	84%	13.9	58%	10.7	56%	10.6
Jiangsu	84%	15.0	51%	14.7	46%	15.2
Heilongjiang	83%	18.8	45%	18.8	44%	18.3
Chongqing	83%	13.3	53%	9.5	54%	9.3
Shanghai	82%	16.1	43%	15.4	46%	14.3
Shandong	82%	21.0	53%	20.4	48%	22.0
Hunan	82%	14.5	45%	12.2	45%	12.4
Guangxi	81%	13.0	48%	9.5	51%	9.3
Shanxi	81%	19.7	47%	21.9	39%	23.9
Liaoning	80%	16.6	43%	19.5	34%	20.9
Zhejiang	80%	13.1	47%	10.6	48%	10.6
Shaanxi	80%	18.3	54%	19.3	50%	19.1
Anhui	76%	18.0	43%	15.7	39%	16.3
Guizhou	75%	12.6	34%	7.4	39%	7.2
Jiangxi	75%	14.6	32%	12.3	33%	12.0
Guangdong	72%	12.0	41%	7.8	45%	7.5
Xinjiang	72%	24.9	55%	27.7	49%	25.6
Inner Mongol	70%	15.9	38%	15.1	33%	16.1
Gansu	66%	18.9	33%	18.0	29%	16.9
Fujian	65%	9.8	24%	6.5	29%	6.8
Ningxia	63%	19.6	28%	20.2	27%	18.7
Yunnan	51%	13.1	26%	8.3	34%	7.7
Qinghai	46%	19.1	24%	13.2	23%	12.7
Tibet	36%	13.4	28%	5.6	26%	5.8

Note: GAM is generalized addictive model; CV R^2 is R-squared for cross validation; RMSE is root mean squared prediction error ($\mu g/m^3$)

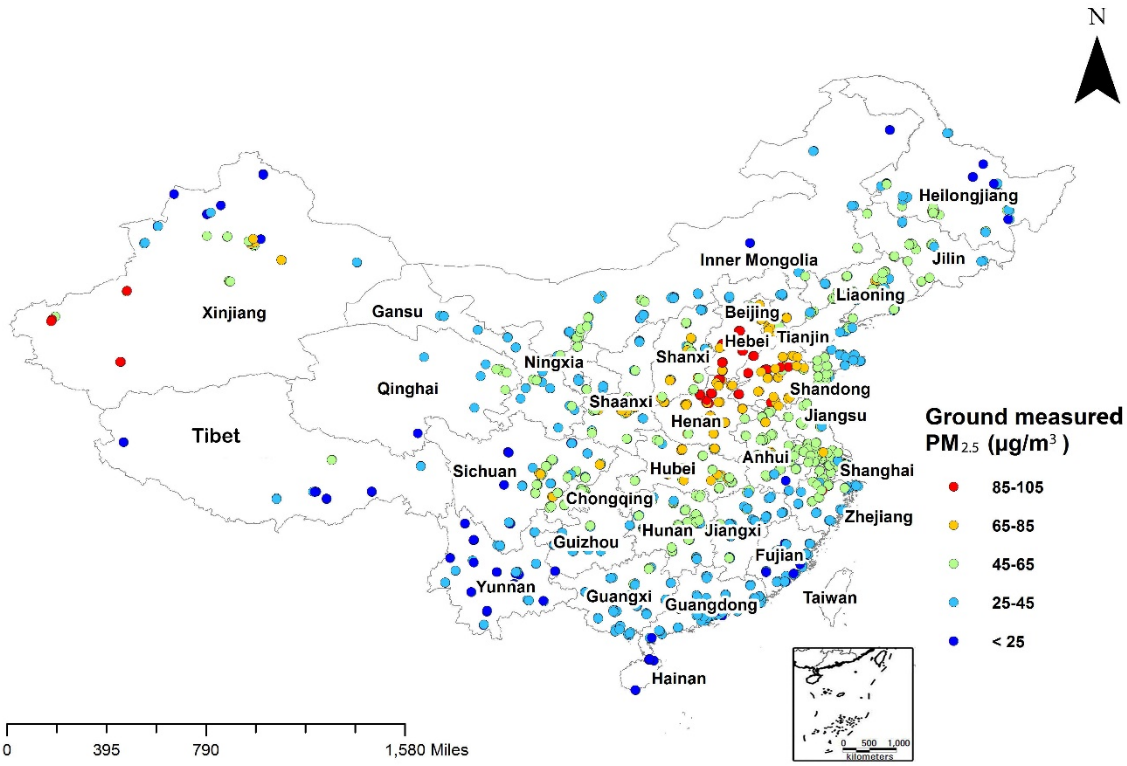

Figure 1. Mean concentrations of ground-level measured $PM_{2.5}$ (μg/m^3) at 1479 stations during 2014-2016.

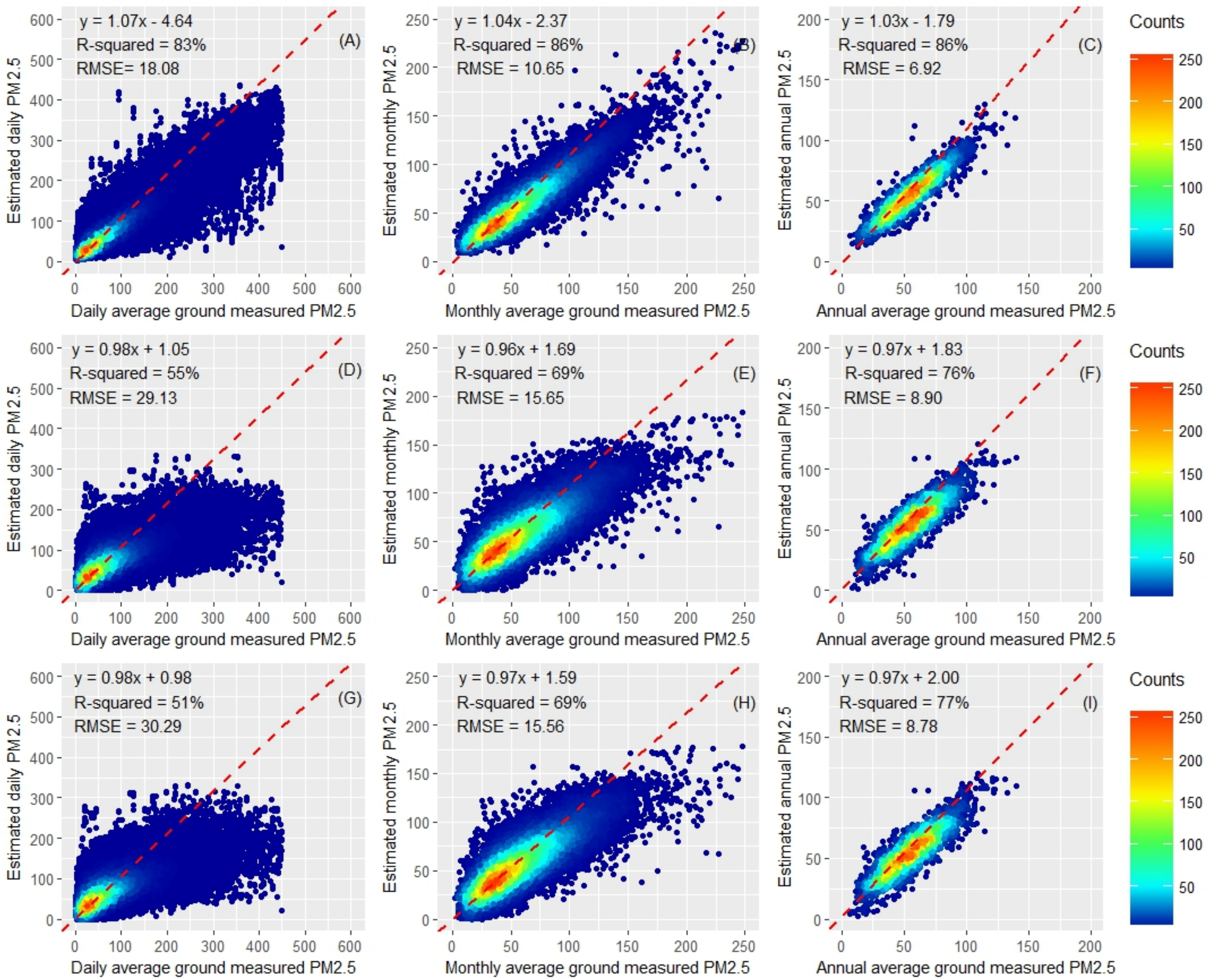

Figure 2. Density scatterplots of model performance and validation. (A), (B) and (C) are daily, monthly and seasonal results for random forests model; (D), (E) and (F) are daily, monthly and seasonal results for generalized additive model (GAM); (G), (H) and (I) are daily, monthly and seasonal results for non-linear exposure-lag-response model. Note: RMSE, root mean squared prediction error ($\mu g/m^3$)

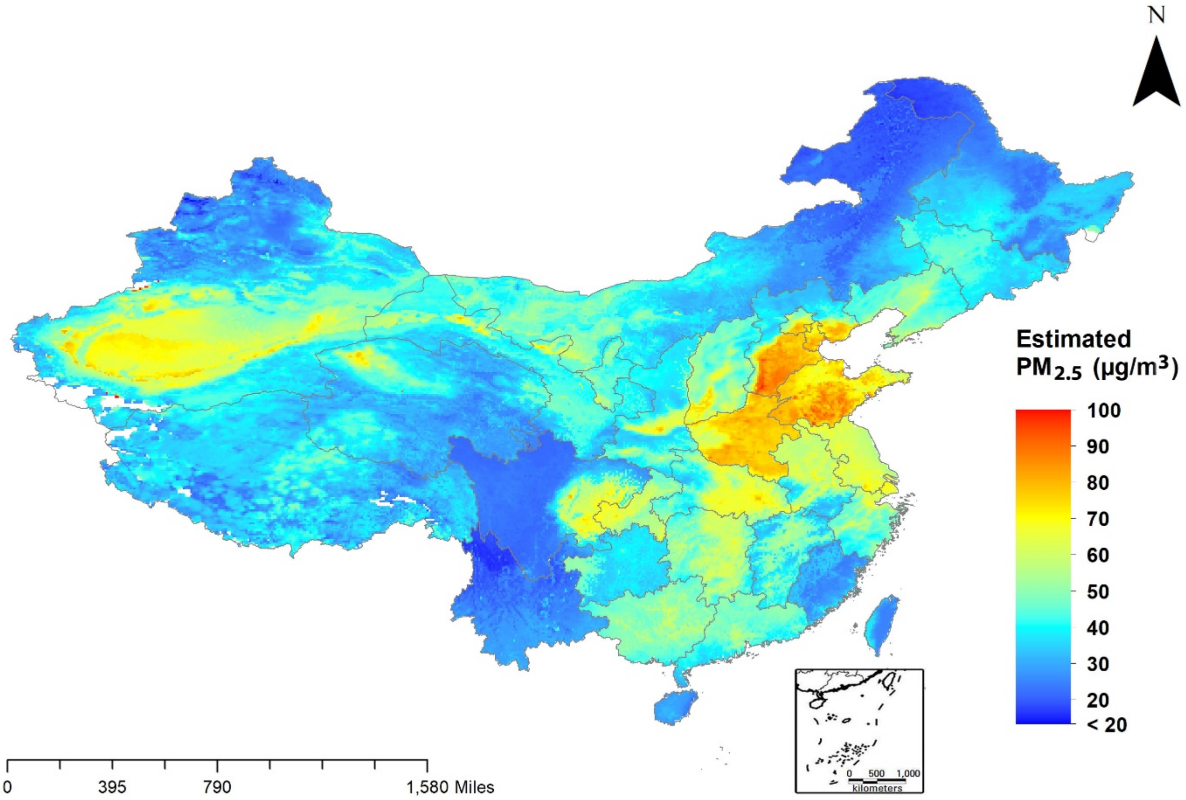

Figure 3. Estimated mean concentrations of $PM_{2.5}$ (μg/m^3) across China during 2005-2016.

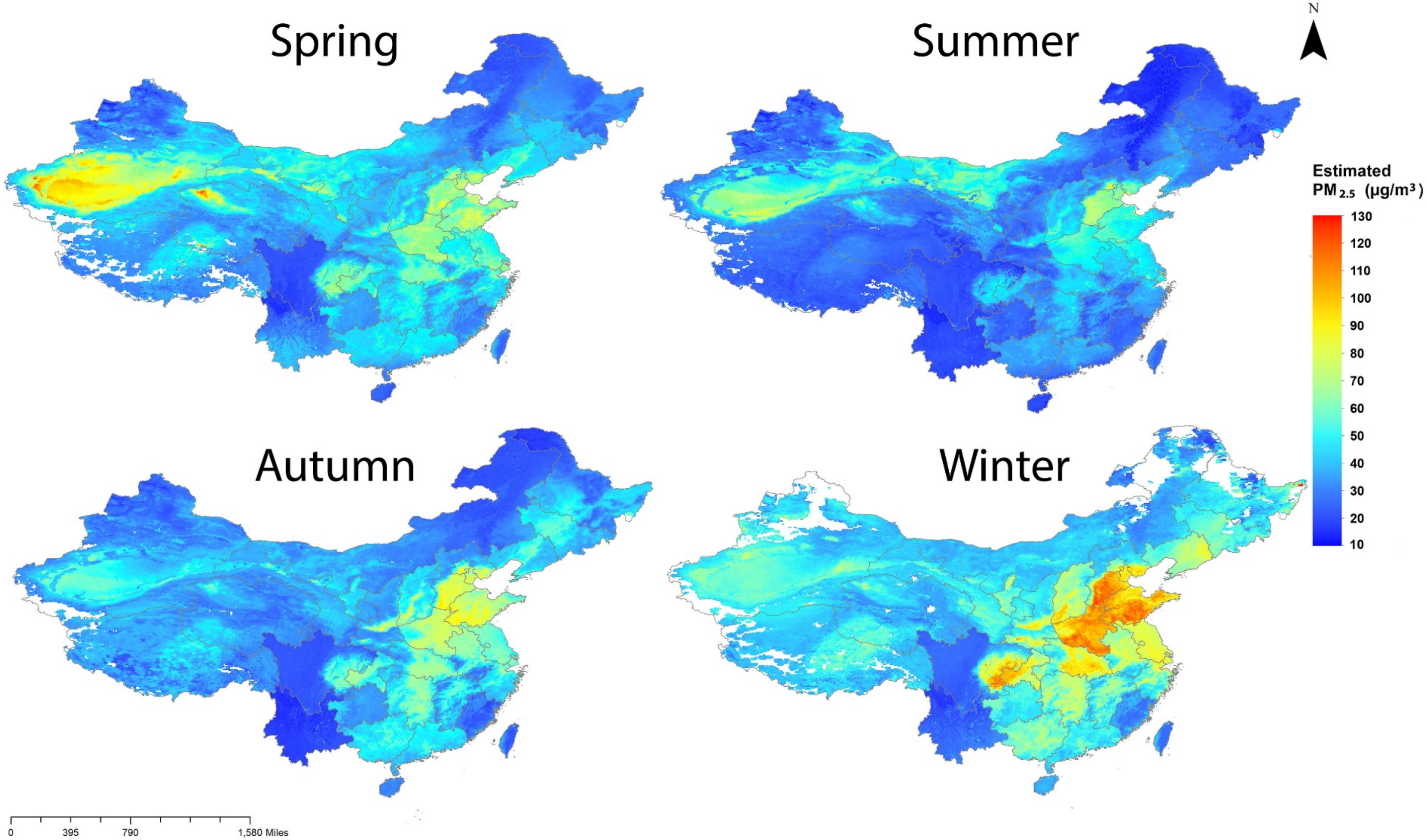

Figure 4. Estimated mean concentrations of $PM_{2.5}$ (μg/m^3) across China in four seasons during the study period.

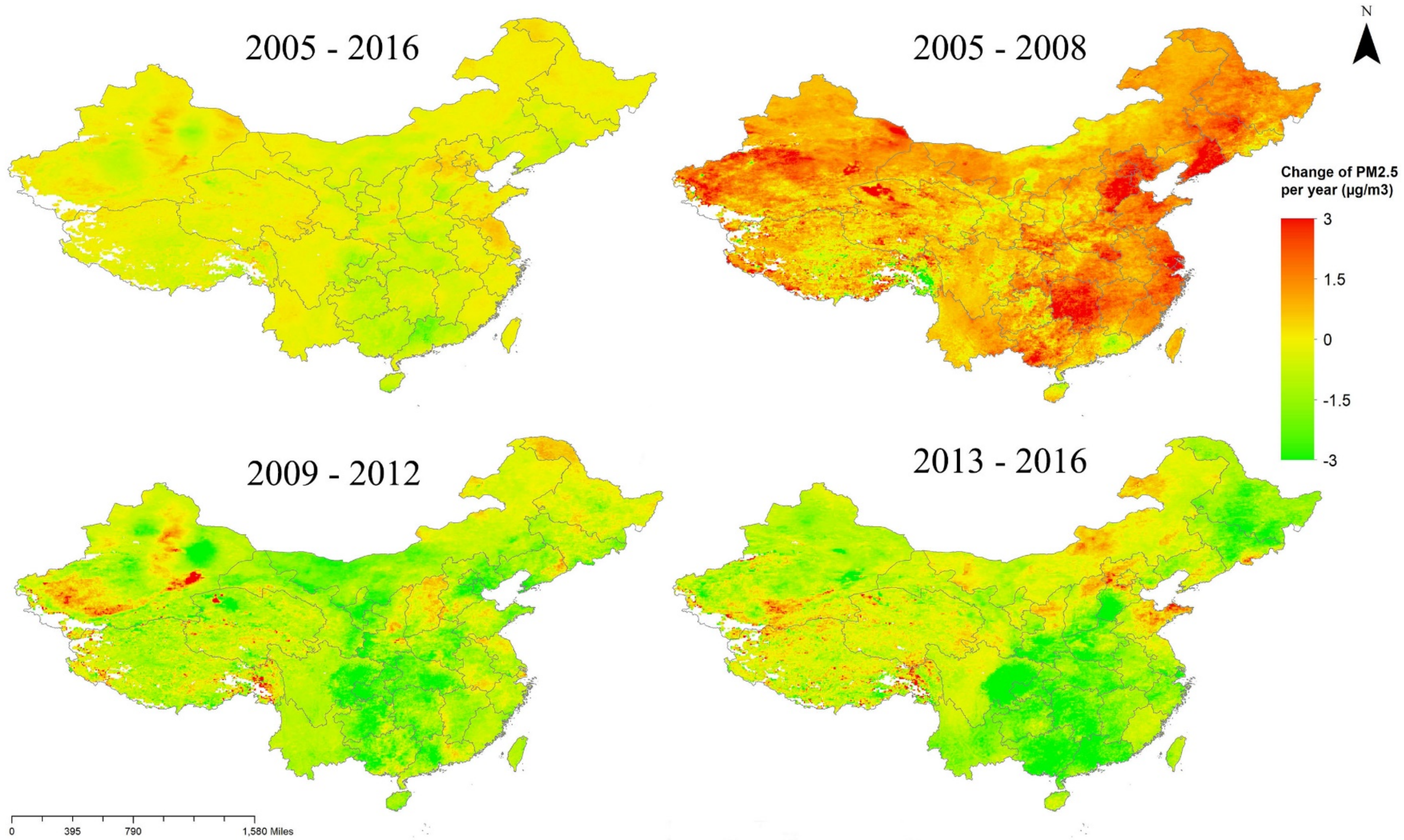

Figure 5. Changes in estimated concentrations of $PM_{2.5}$ ($\mu g/m^3$ per year) over time in China during the study period.