Environ Monit Assess (2022) 194: 284
https://doi.org/10.1007/s10661-022-09934-5

Understanding the distribution and drivers of $PM_{2.5}$ concentrations in the Yangtze River Delta from 2015 to 2020 using Random Forest Regression

Zhangwen Su · Lin Lin · Yimin Chen · Honghao Hu

Received: 5 August 2021 / Accepted: 5 March 2022 / Published online: 16 March 2022

Abstract Understanding the drivers of $PM_{2.5}$ is critical for the establishment of $PM_{2.5}$ prediction models and the prevention and control of regional air pollution. In this study, the Yangtze River Delta is taken as the research object. Spatial cluster and outlier method was used to analyze the temporal and spatial distribution and variation of surface $PM_{2.5}$ in the Yangtze River Delta from 2015 to 2020, and Random Forest was utilized to analyze the drivers of $PM_{2.5}$ in this area. The results indicated that (1) based on the spatial cluster distribution of $PM_{2.5}$, the northwest and north of Yangtze River Delta region were mostly highly concentrated and surrounded by high concentrations of $PM_{2.5}$, while lowly concentrated and surrounded by low concentrations areas were distributed in the southern; (2) the relationship between $PM_{2.5}$ concentrations and drivers in the Yangtze River Delta was modeled well and the explanatory rate of drivers to $PM_{2.5}$ were more than 0.9; (3) temperature, precipitation, and wind speed were the main driving forces of $PM_{2.5}$ emission in the Yangtze River Delta. It should be noted that the repercussion of wildfire on $PM_{2.5}$ was gradually prominent. When formulating air pollution control measures, the local government normally considers the impact of weather and traffic conditions. In order to reduce $PM_{2.5}$ pollution caused by biomass combustion, the influence of wildfire should also be taken into account, especially in the fire season. Meanwhile, high leaf area was conducive to improving air quality, and the increasing green area will help reduce air pollutants.

Keywords Yangtze River Delta · $PM_{2.5}$ drivers · Spatial distribution · Random Forest · Spatial autocorrelation

Introduction

$PM_{2.5}$ (equivalent diameter of particulate matter ≤ 2.5 μm in aerodynamics) is a major urban pollutant, which contributes to the occurrence of many diseases from cardiovascular and respiratory (Dominici et al., 2006; Peng et al., 2009), and has even been linked to fetal development (Guo et al., 2018). Although $PM_{2.5}$ has declined worldwide over the past 30 years, there is still no region below the recommended annual target ($PM_{2.5} < 10$ μg/m^3) set by the World Health Organization (Archer et al., 2020). Frequent wildfires in many regions all over the world have led to widespread air pollution in

Supplementary information The online version contains supplementary material available at https://doi.org/10.1007/s10661-022-09934-5.

Z. Su (✉) · Y. Chen · H. Hu
College of Applied Chemical Engineering, Zhangzhou Institute of Technology, Zhangzhou 363000, China
e-mail: FujianSZW@126.com

L. Lin
Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD 20740, USA

recent years (Pang et al., 2020). This crisis is more alarming in the circumstances of the COVID-19 pandemic, because more inhalation of smoke particles increases the incidence of lung injury, which further aggravates the spread of the epidemic (Matz et al., 2020).

The serious $PM_{2.5}$ pollution has aroused great concern of the Chinese government and the public recently, and triggered many researches on the mechanism of the impingement of air pollution on the ecological environment and human health (Guo et al., 2018; Pang et al., 2020; Xue et al., 2017). An accurate estimation of $PM_{2.5}$ exposure is hence essential for assessing its impact on health risks. However, a well-known problem with $PM_{2.5}$ data from air pollutant monitors in ground is that it does not take into account the spatial change of $PM_{2.5}$ concentrations. In previous studies, the information of $PM_{2.5}$ concentrations is often simply taken from a monitoring station or defined as a measurement index that is the mean of measurements obtained from several monitors (Liu et al., 2021; Miskell et al., 2017), which results in the loss of $PM_{2.5}$ spatial variation information. Therefore, it is of great importance to accurately estimate the spatiotemporal variation of surface $PM_{2.5}$ concentrations at a large scale.

The Yangtze River Delta urban agglomeration has currently the most complete industrial system, the best urbanization foundation in China, and its economic aggregate accounts for nearly a quarter of the whole country, playing a crucial role in the economic and social development of China (Ma et al., 2019a, b). With the continuous improvement of industrial and economic development, the high-density population and road traffic network make the pollution buffer distance among cities smaller, thereby bringing serious air pollution problems to this region (Yun et al., 2019).

Many studies have been carried out on the $PM_{2.5}$ exposure and its drivers in the Yangtze River Delta (YRD) region, which provide the theoretical basis for the scientific prevention and restriction of regional air pollutants (Guo et al., 2018; Pang et al., 2020; Xu et al., 2020; Xue et al., 2017; Yun et al., 2019). Socioeconomic factors as the direct pathogenic factors for environmental pollution have long been the focuses of researches on $PM_{2.5}$ influence factors (Zhou et al., 2021). Environmental factors such as air temperature, wind speed, humidity, and topography are considered to be the indirect contributors to the concentrations of $PM_{2.5}$ in the air because they can affect the airflow and the spread of particle matters (Xu et al., 2020). However, this response relationship in the context of climate change has been changing: Yun et al. (2019) found that environmental factors jointly contributed to $PM_{2.5}$ pollution in the YRD; Xu et al. (2020) also noted that the influence of environmental factors on the $PM_{2.5}$ concentrations was greater than the socioeconomic factors in the YRD region. Meteorological factors and land use are the common environmental drivers for predicting $PM_{2.5}$ exposure in the YRD (Liu et al., 2021; Yun et al., 2019; Zhou et al., 2021), while the impact of smoke produced by wildfire on $PM_{2.5}$, in particular, is rarely studied, to our knowledge. Recent theoretical developments have revealed that about 50% of the carbon emissions worldwide are linked to wildfires, and approximately 3.3 million people worldwide died prematurely from poor air quality, with 5–8% of which are attributed to air pollution from fire emissions (Andela et al., 2017; Matz et al., 2020). Therefore, it is necessary to bring wildfire occurrence information into the analysis of $PM_{2.5}$ driving factors.

A deep understanding of the spatio-temporal distribution of $PM_{2.5}$ concentrations and its drivers will help to establish an effective prediction model and improve the accuracy of $PM_{2.5}$ concentration prediction, which is of great significance to the establishment of air pollution prevention and control measures. The main objectives of this study are as follows: (1) to identify the spatio-temporal distribution of $PM_{2.5}$ concentrations in the YRD in the period of 2015 to 2020; (2) to understand the relative importance of environmental and human drivers affecting the $PM_{2.5}$ exposure, and their effects on $PM_{2.5}$ concentrations; and (3) to provide theoretical support for the effective prevention and control of $PM_{2.5}$ emission in the YRD.

Materials and methods

Study area

The YRD region on the eastern coast of China is within 118° 33′ ~ 123° 10′ E and 28° 0′ ~ 33° 52′ N, covering an area of 211,700 km^2 (Fig. 1). Under the influence of the subtropical monsoon climate, its

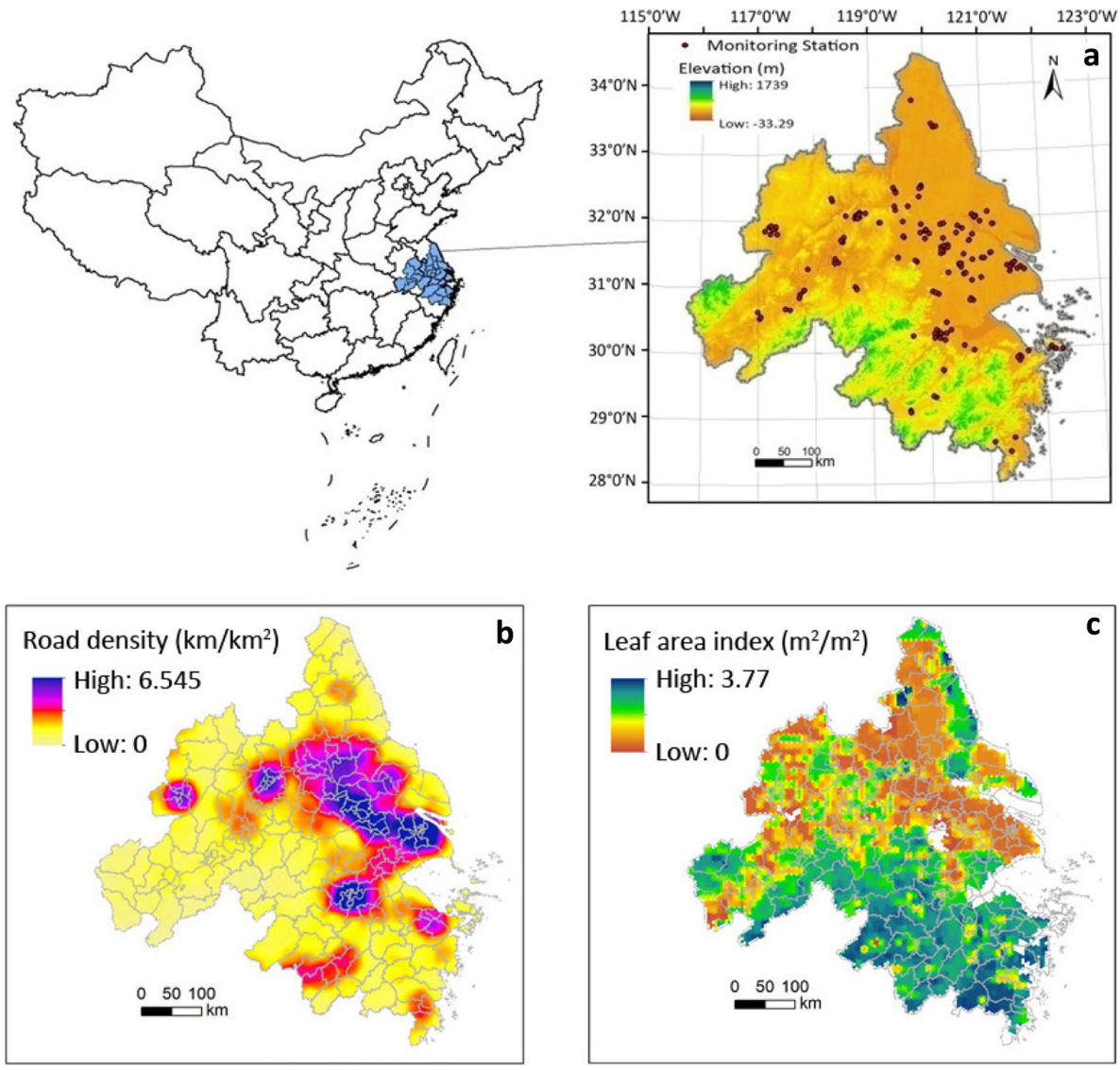

Fig. 1 Study area of the YRD region. Distribution of (**a**) elevation and monitor stations (black points), (**b**) road density, and (**c**) leaf area index in the YRD region

annual average precipitation and temperature are 804 ~ 2057 mm and 9.3 ~ 17.3 °C, respectively (Xue et al., 2017). As an alluvial plain of the Yangtze River to the Pacific Ocean, the YRD region is characterized of a high-density river network, and low-altitude terrain whose altitude is mostly less than 10 m and gradually decreases from southwest to northeast with low hills scattering around. The region embraces 26 major cities and has high rates of economic growth and urbanization (Xu et al., 2020). Thus, the population is dense, and the urban heat island effect is significant. $PM_{2.5}$ emissions in this region maintain high levels (Xue et al., 2017).

Data collection and process

$PM_{2.5}$ data

As of 2020, 165 ground environmental monitoring sites were built by the Chinese government in this

region (Fig. 1a). The $PM_{2.5}$ concentrations data was collected from the monitoring daily data from 2015 to 2020 of the ground air quality monitoring station established by China's Ministry of Ecology and Environment. To ensure consistency on the temporal scale for all variables, these data were integrated into the annual average $PM_{2.5}$ concentrations. Before discussing the spatial distribution of $PM_{2.5}$ exposure in the YRD, the discrete values of ground observations were converted into continuous surface raster data by spatial interpolation. The spatial interpolation techniques, including ordinary kriging (OK), universal kriging (UK) and inverse distance weighting (IDW) (Jerrett et al., 2005; Ma et al., 2019a, b; Mercer et al., 2011; Xu et al., 2019), have been widely adopted in the field of the estimation for $PM_{2.5}$ concentrations, and can be found in ArcGIS10.4 software.

Environmental factors

Meteorological data were from ERA5-Land data available at Copernicus Climate Change Service (C3S) (https://cds.climate.copernicus.eu). ERA5-Land data is a reanalysis dataset produced by the ECMWF ERA5 reanalysis model, which combines observations from around the world into a globally complete and consistent dataset using the laws of physics. This dataset includes 50 dynamic monthly indicators representing temperature, wind speed, precipitation, and vegetation since 1981, with a spatial resolution of $0.1° \times 0.1°$ (about 9×9 km), which describes the past and the present climate conditions (Zhang et al., 2021). Detailed information, code, and summary statistics of all the variables are given in Tables S1 and S2, respectively.

Leaf area index (LAI) refers to the multiple of the total plant leaf area per unit land area. It determines the size of the interface for the exchange of energy (including radiation) and mass between the canopy and the atmosphere. This is an important structural parameter of the ecosystem, which is used to provide quantitative information for the description of material and energy exchange on plant canopy surface. LAI can make an influence on $PM_{2.5}$ through various mechanisms (Berg & McColl, 2021; Velásquez Ciro et al., 2021). This data was also obtained from C3S Data Platform.

Anthropic factors

According to the existing researches on drivers of $PM_{2.5}$, we considered road and railway density, population density, and the proportion of land use (farmland and forest land) as human-influencing factors of $PM_{2.5}$ (Joharestani et al., 2019; Liu et al., 2021; Xu et al., 2020; Yun et al., 2019).

Road and rail information used in this study is 1:1 million terrain feature vector data provided by the National Geomatics Center of China, which was processed by line density analysis in ArcGIS 10.4 to obtain road and rail density raster data. The population density with a resolution of 30 arc-seconds (approximately 1 km at the equator) was derived from the global population density data provided by WorldPop. By calculating the number of people in each pixel and proofreading it with the official population estimation data of the United Nations, the annual global population density data from 2000 to 2020 were generated (these data can be downloaded from https://www.worldpop.org/) (Lloyd et al., 2017). This data has been used in many studies (Liu et al., 2020; Nethery et al., 2021), such as in correcting some remote sensing data including nighttime light data (Liu et al., 2020).

Land-use variable has always been a conventional option in the research on $PM_{2.5}$ drivers, which represents the degree of landscape modification by human. C3S Data Platform provides access to the land-use information used in this research. The dataset provides a global land cover raster data from 2015 to 2020, by dividing the land surface into 22 categories following the Land Cover Classification System (LCCS) of the Food and Agriculture Organization (FAO) of the United Nations. The advantages of long-term consistency, annual renewal, and high resolution (300 m) on a global scale enable it for a wide range of applications and scientific researches, including land accounting, forest monitoring, and ecological environments (Pesaresi et al., 2016).

Wildfire factors

In the present study, we used the active fire data during 2015–2020 released by the National Aeronautics and Space Administration (NASA) (https://earthdata.nasa.gov), which is a 1-km global daily fire product (MCD14ML) retrieved from Moderate-resolution

Imaging Spectroradiometer (MODIS) on the Terra and Aqua satellites. Active fire data has been used in wildfire researches worldwide, and the data information includes fire-point geographic coordinates, detection time, credibility, and information on other detected fire pixels. Due to the impact of cloud and smoke cover of satellite retrieval on fire points, we only utilized the fire points that have been detected with 75% confidence in the dataset (Ferreira et al., 2020). Furthermore, fire points in urban and rural areas, construction sites, and farmland were discarded based on the global 300-m resolution land cover dataset from 2015 to 2020 (https://cds.climate.copernicus.eu) (Su et al., 2021).

Scale of study cell

In order to have a unified spatial scale, ArcGIS10.4 software was employed to divide the study area into 9040 hexagonal grid cells, each with an area of approximately 21.65 km^2. Compared with the traditional rectangular grid, the main advantage of a hexagonal grid is that the coverage area of the generated cell is more uniform, the distance from the geometric center of mass to each edge is the same, and the distortion is also avoided (Ferreira et al., 2020). Subsequently, the annual average $PM_{2.5}$ concentrations of all cells in the research area and the corresponding fire density, environment, and human factors were extracted using "zonal statistics as table" in ArcGIS 10.4.

Data analysis

$PM_{2.5}$ spatial distribution

Spatial autocorrelation analysis was adopted to explain the spatial correlation of $PM_{2.5}$ in all cells of the study area. In our research, we utilized the "Spatial autocorrelation (Moran's I)" tool in ArcGIS 10.4 software to conduct this analysis. The analysis result provides three indicators: Global Moran's I, *z*-score and *p*-values. The Global Moran's I is used to express the spatial correlation degree of $PM_{2.5}$ in all cells; *z*-score and *p*-values are employed to express the significant level of spatial correlation. The calculation process is as follows (Getis & Ord, 1992; Mitchell, 2005):

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}(x_i - \overline{X})(x_j - \overline{X})}{\sum_{i=1}^{n}(x_i - \overline{X})^2} \tag{1}$$

where I is the Global Moran's I, x_i and x_j represent the attribute x of cell i and j, with $i, j = 1,2,\ldots,n$, and n is the total number of cells. $\overline{X}$ is the mean for the attribute x of corresponding cell, $w_{i,j}$ is the spatial weight between cell i and j; $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}$ is the aggregation of all spatial weights. Generally, Moran's $I > 0$ indicates a positive spatial correlation, and the larger the value is, the more obvious is; Moran's $I < 0$ indicates a negative spatial correlation, and the smaller the value is, the greater the spatial difference is; otherwise, when Moran's $I = 0$, the spatial distribution is random.

In addition, the *z*-score of global autocorrelation statistical data is calculated as follows:

$$z_I = \frac{I - E[I]}{\sqrt{V[I]}} \tag{2}$$

where $E[I] = -\frac{1}{n-1}$ and $V[I] = E[I^2] - E[I]^2$ represent the expectation and variance of the global Moran's I, individually.

The local spatial autocorrelation was used to describe the spatial association mode of $PM_{2.5}$ in different spatial positions, and the Clustering/Outlier analysis (Anselin Local Moran's I) was employed in ArcGIS 10.4 software. The analysis has the ability to classify spatial clusters of high-value or low-value cells, and can also identify spatial outliers (high values are surrounded by low values or low values are surrounded by high values). Clustering/Outlier analysis can also get Local Moran's I, accompanied by *z*-score and *p*-value representing the statistical significance of Local Moran's I. The calculation process is as follows (parameter interpretation is the same as formula (1)) (Anselin, 1995; Mitchell, 2005):

$$I_i = \frac{x_i - \overline{X}}{S_i^2} \sum_{j=1, j\neq i}^{n} w_{i,j}(x_j - \overline{X}),\ S_i^2 = \frac{\sum_{j=1, j\neq i}^{n} w_{i,j}(x_j - \overline{X})^2}{n-1} \tag{3}$$

In addition, the *z*-score of statistical data of local spatial autocorrelation is calculated as follows:

$$z_{Ii} = \frac{I_i - E[I_i]}{\sqrt{V[I_i]}} \quad (4)$$

where $E[I_i] = -\frac{\sum_{j=1,j\neq i}^{n} w_{i,j}}{n-1}$ and $V[I_i] = E[I_i^2] - E[I_i]^2$ represent the expectation and variance of the Local Moran's I, respectively.

Random Forest Regression

Random Forest (RF) Regression was conducted for the analysis of the importance of variables in this study. RF regression is a nonparametric technique obtained from the regression tree. Its principle is that N sample units are randomly and replaceably extracted from the original data to generate a regression tree; m (< M) variables are randomly selected at each node and used as candidate variables for segmentation nodes. Then, the results of each regression tree are integrated to generate predicted values and automatically calculate the relative importance of each independent variable (Cutler et al., 2007).

An important advantage of the RF is that there is no need to cross-validate it or use independent tests to obtain an unbiased estimate of the error. RF can be evaluated internally, which means that an unbiased estimate of the error can be established during the generation process. RF is an optimized version of Bagging based on a tree model. In the Bagging method, about 1/3 of the samples will not appear in the training sample set collected by Bootstrap each time, so it will not participate in the establishment of tree. This 1/3 data is called out-of-bag (OOB), which is used to replace the test set error estimation method (Liaw & Wiener, 2002). Without verifying the dataset, the OOB prediction error can be calculated (the predicted values of the sample points that are not used in tree generation can be estimated by the generated tree, and the OOB prediction error can be obtained by comparing predictor with the real values). This technology has been proved to have high prediction accuracy, high tolerance to outliers and "noises" (Breiman, 2001), and has been widely used in many different research fields in the past (Chen et al., 2021; Cutler et al., 2007), especially in the research related to the air pollution in recent years (Zhan et al., 2018). In this study, variables listed in Table S1 were used in the RF regression, and RF regression was carried out with the "*randomForest*" packages in R software.

When using RF for data fitting and prediction, the number of trees (*ntree*) and the number of each random variable at each node (*mtry*) must be parameterized. Liaw and Wiener (2002) suggests choosing *mtry* = M/3 for RF regression, where M is the number of variables. The parameter *ntree* is calculated after determining *mtry*. The *ntree* specifies the number of decision trees contained in a random forest. When the errors in the model are stable, the minimum value of *ntree* is used as the parameter to train the model (the calculation result is shown in Fig. S1).

Indicators of regression evaluation

In this study, the performance of the fitting models was evaluated by the indicators that have been widely used in previous studies, including RMSE (root mean squared error), MAE (mean absolute error), MSR (mean of square residual), and coefficient of determination (R square) of models (Su et al., 2021; Xue et al., 2017). RMSE is the square root of the ratio between the sum of squares for error between predicted values and true values and the number of observations, and MAE is the average of absolute values of errors between observed values and true values. These two indicators were used to describe the errors between predicted values and true values. The difference between them is that RMSE amplifies and severely punishes large errors due to its process of the square. MSR measures the fitting degree of the model by measuring the ratio between the residual square and the sample size. The smaller value of RMSE, MAE, and MSR, the better accuracy of the prediction model will be in describing experimental data. R^2 represents the percentage of the total variation in the observed value that is explained by the regression model.

Results

Comparison between the prediction of different interpolation and observation

In order to evaluate the accuracy of the interpolation algorithm, the tenfold cross-validation (CV) method was employed to verify the interpolation effect of OK, UK, and IDW, and the results are presented in Table 1. We created ten training sets and

Table 1 The linear correlation coefficient between the observed values and the predicted values from IDW, UK, and OK interpolation for test sets

Test sets	2015			2016			2017			2018			2019			2020		
	R^2_{IDW}	R^2_{UK}	R^2_{OK}	R^2_{IDW}	R^2_{UK}	R^2_{OK}	R^2_{IDW}	R^2_{UK}	R^2_{OK}	R^2_{IDW}	R^2_{UK}	R^2_{OK}	R^2_{IDW}	R^2_{UK}	R^2_{OK}	R^2_{IDW}	R^2_{UK}	R^2_{OK}
T1	0.883	0.877	0.827	0.842	0.77	0.762	0.849	0.768	0.753	0.819	0.732	0.657	0.858	0.74	0.797	0.883	0.869	0.846
T2	0.889	0.875	0.803	0.826	0.828	0.765	0.856	0.719	0.811	0.897	0.857	0.857	0.929	0.878	0.884	0.913	0.918	0.9
T3	0.853	0.819	0.702	0.776	0.753	0.7	0.93	0.856	0.814	0.755	0.809	0.741	0.892	0.887	0.85	0.941	0.889	0.912
T4	0.899	0.77	0.773	0.842	0.834	0.778	0.842	0.746	0.778	0.887	0.845	0.81	0.841	0.816	0.777	0.917	0.918	0.897
T5	0.824	0.833	0.783	0.862	0.816	0.795	0.817	0.812	0.777	0.881	0.854	0.812	0.936	0.897	0.868	0.938	0.851	0.87
T6	0.848	0.729	0.700	0.819	0.786	0.727	0.823	0.682	0.810	0.849	0.794	0.745	0.670	0.659	0.607	0.898	0.87	0.879
T7	0.877	0.802	0.823	0.865	0.872	0.851	0.886	0.765	0.772	0.872	0.837	0.685	0.912	0.890	0.892	0.861	0.859	0.836
T8	0.815	0.73	0.686	0.839	0.836	0.834	0.853	0.777	0.782	0.646	0.752	0.694	0.895	0.824	0.854	0.938	0.883	0.877
T9	0.896	0.872	0.834	0.8	0.767	0.735	0.861	0.613	0.797	0.829	0.787	0.719	0.955	0.903	0.887	0.9	0.871	0.85
T10	0.858	0.813	0.745	0.782	0.766	0.762	0.868	0.833	0.669	0.837	0.808	0.805	0.844	0.823	0.76	0.766	0.82	0.709

Table 2 Global Moran's I of $PM_{2.5}$ exposure in the four study areas

Year	Global Moran's I	*z*-score	*p*-value
2015	0.958636	211.041863	<0.0001
2016	0.967905	213.165213	<0.0001
2017	0.972172	214.063621	<0.0001
2018	0.97572	214.79421	<0.0001
2019	0.964589	212.350493	<0.0001
2020	0.958581	211.026694	<0.0001

the corresponding test sets, and then used the test data to extract the interpolation results of the training sets. The linear correlation coefficient between the observed values and the predicted values from interpolation for test sets described that IDW interpolation is better than UK and OK methods (Table 1). We believe that the IDW interpolation method can better estimate the $PM_{2.5}$ concentration in the study area, and can reflect the change of $PM_{2.5}$ exposure. Therefore, we only consider $PM_{2.5}$ grid data obtained by the IDW interpolation method as the dependent variable for modeling analysis in our study.

Spatial distribution analysis of $PM_{2.5}$

The global spatial autocorrelation analysis showed that the Global Moran's I from 2015 to 2020 were greater than 0, and all of them passed the 5% significance level test (Table 2), indicating that the spatial positive correlation of $PM_{2.5}$ exposure in this period. Meanwhile, the spatial correlation of $PM_{2.5}$ exposure in 2018 was the strongest, followed by 2017, while the weakest spatial correlation appeared in 2020.

In order to further explore the spatial correlation, difference, and aggregation distribution between $PM_{2.5}$ in each cell and surrounding areas, the Local Moran's I scatter diagram and Local Indicators of Spatial Association (LISA) aggregation diagram were obtained to analyze the spatial pattern of $PM_{2.5}$ exposure in the YRD (Fig. 2). A total of 97.94–99.21% cells were concentrated in HH and LL quadrants (Fig. 2a–f), while almost none of the cell fell into HL and LH quadrants during 2015–2020. The numbers of cells in the HH and LL quadrants are roughly equal in 2015–2017; most of the cells are concentrated in the HH quadrant in 2018–2020. However, the scatter plot did not provide us with more detailed information

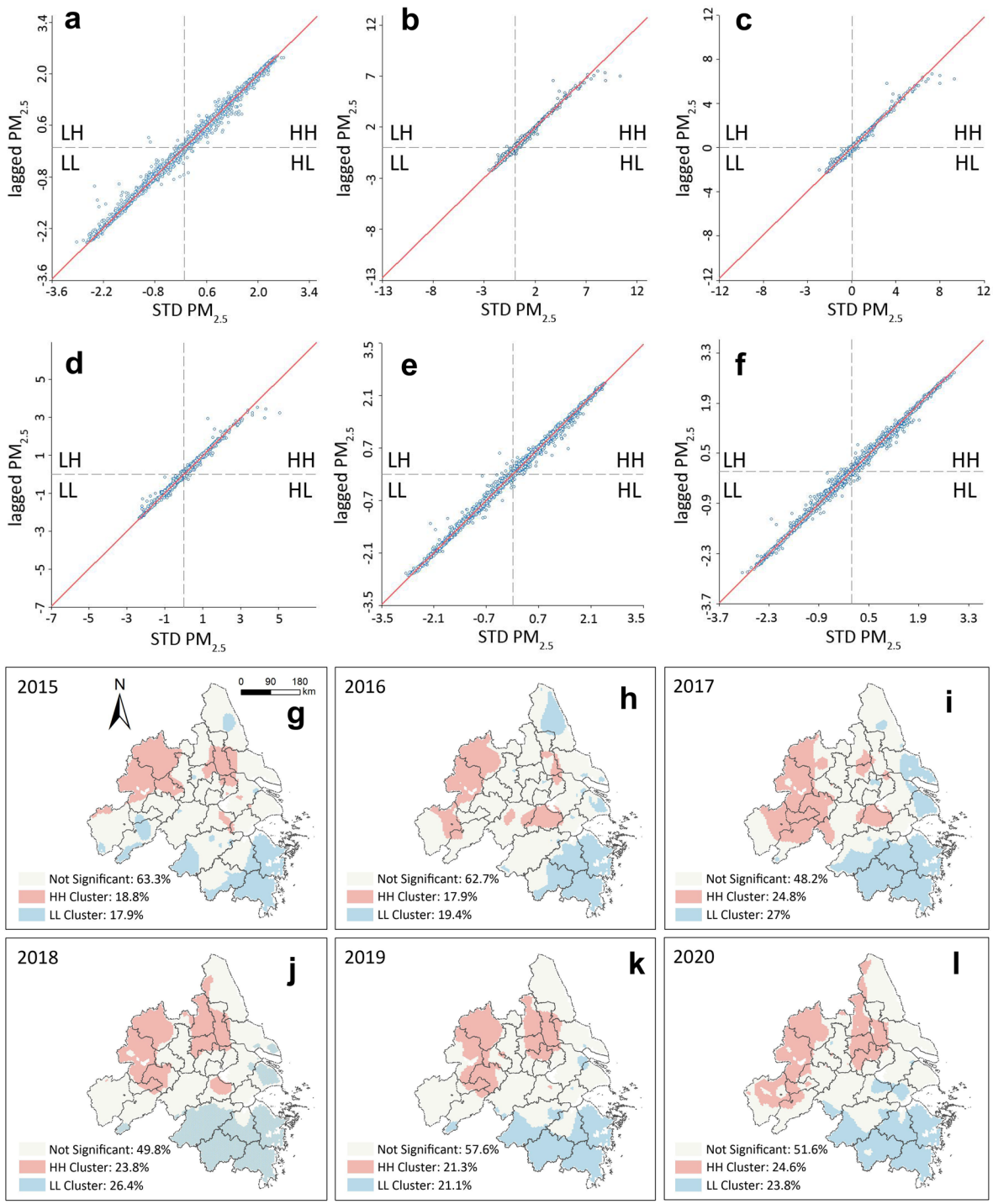

Fig. 2 Spatial distribution of $PM_{2.5}$ in the study area based on cluster and outlier analysis (Ansel in Local Moran's I). Scatter plot of Moran's I (**a**–**f**) and LISA agglomeration (**g**–**l**) of $PM_{2.5}$ in the study area. In **a**–**f**, the number of scattered points is 9040; the abscissa is the observed value of $PM_{2.5}$ of a spatial unit (after standardization), and the ordinate is the "lagged" value of the spatial unit, that is, the average value of the observed $PM_{2.5}$ of adjacent units (after standardization). In **g**–**l**, the level of significance is p-value < 0.05

about the statistical significance of spatial correlation for all cells.

In order to demonstrate the local correlation types and clusters and their statistical significance, the LISA cluster maps of $PM_{2.5}$ exposure during the 2015–2020 period are illustrated in Fig. 2g–l. Not all cells of HH or LL clusters were statistically significant (p-value < 0.05), and there were regions with insignificant spatial correlation during 2015–2020. The periods with the largest area of significant spatial clustering effect appeared in 2017 (51.8%) and 2018 (50.2%). Moreover, the LISA diagram demonstrates that HH clusters were mainly concentrated in the north and northwest, and LL clusters were almost concentrated in the south of the research area. The maximum area of the significant HH concentrations occurred in 2017 and 2020 (24.8% and 24.6% of the research area, respectively).

Analysis of factors affecting $PM_{2.5}$

Comparison of all and optimal combination variables in RF

According to the data characteristics of $PM_{2.5}$ concentrations (continuous dependent variable), the RF regression model was used to study the relationship between $PM_{2.5}$ concentration and its drivers. Many studies on $PM_{2.5}$ emission drivers or model predictors found that $PM_{2.5}$ concentrations were non-linearly associated with drivers (Zhou et al., 2021), bringing challenges to the use of the traditional linear regression model. As shown in Table 3, the RF regression algorithm produced a good fitting in detecting the relationship between $PM_{2.5}$ and its driving factors. The variable explanation rate for all results was more than 90%, for an individual year from 2015 to 2020 (92.19–98.31%) and the whole 6-year period (94.51%). It is worth noting that both the largest variable interpretation rate and the smallest mean of squared residuals occurred in 2020, followed closely by the fitting results for 2019.

RF regression provides the rank of the importance of 11 factors in influencing $PM_{2.5}$ concentrations over the 6 years from 2015 to 2020 (Fig. 3). Figure 3a presents that the scores of the global importance of average temperature (AT), cumulative precipitation (CP), average wind speed (AWS), and road density (DROAD) were much higher than the other variables. In addition, it can be seen from the local importance ranking diagram (Fig. 3c) that factors unimportant in the global ranking were also of the least importance in the local importance ranking, such as proportion of crop (PCORP) and proportion of forest (PFOREST). Based on this analysis, we believe that not all factors contribute equally to the accuracy of RF regression analysis. Some variables with less obvious characteristics may produce larger noise in the regression and bring bigger errors to the precision of the model (Breiman, 2001). Therefore, we further removed the factors with a low contribution to RF regression through the variable selection.

Based on the ranking of importance from high to low after estimating the importance score of all the variables to the dependent variable by RF regression, we discarded the variables that had smaller values and retained the factors that contributed more to the model by the tenfold cross-validation (CV) method. CV method was used to deal with the problem that the individual test results are too

Table 3 Fitting results of $PM_{2.5}$ and drivers based on Random Forest regression

Year	Complete variables		Optimal combination of variables	
	Mean of squared residuals	Variables explained (%)	Mean of squared residuals	Variables explained (%)
2015	1.2510	96.51	1.1367	96.83
2016	3.1576	93.48	2.1904	95.48
2017	4.6360	92.19	3.7074	93.76
2018	1.3065	97.2	0.9826	98.23
2019	0.4423	98.05	0.3251	98.56
2020	0.4302	98.31	0.3169	98.75
2015–2020	4.4114	94.51	3.6382	95.42

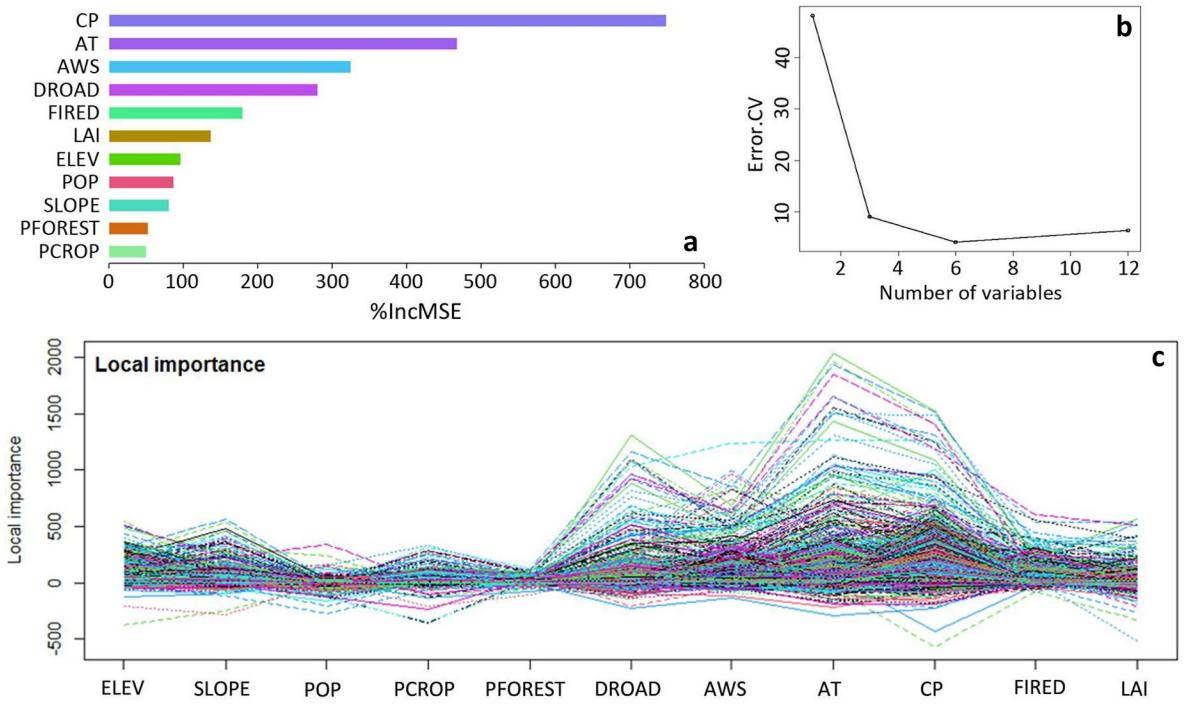

Fig. 3 **a** Global variable importance of $PM_{2.5}$ influencing factors in six years from 2015 to 2020 based on Random Forest. **a** "%IncMSE" is to increase in mean squared error, by randomly assigning values to each predictive variable. If the predictive variable is more important, then the error of the model prediction will increase when its value is randomly replaced. Therefore, the greater the value, the greater the importance of the variable. **b** The optimal number of independents selected by tenfold cross-validation. **c** Local variable importance

unilateral. The calculation of CV provided a curve about cross-validation shown in Figs. 3b and 4c displaying the relationship between the model error and the optimal number of variables used for fitting. The error decreased sharply as the number of variables increased at the beginning, when the number of variables reduced to 6, the decline became neglectable, and then began to climb slowly (Figs. 3b and 4c). The error appeared to be minimal when the six important variables were kept to get the desired regression result. The calculated results demonstrated that better results and less complexity for model dimensions were obtained using these six more important variables (CP, AT, AWS, DROAD, FIRED, LAI) in RF regression models than all variables (Figs. 5 and S2, Table 3). The indexes of fitting values, including variable explained, mean of square residual, RMSE, MAE, and R^2, all revealed the interpretation advantage of the optimal combination for variables to $PM_{2.5}$ by the elimination of unimportant or high-noise variables. Taking the whole 6 years data (2015 to 2020) as an example, the explain rate of overall variance and modeling R^2 in of CP, AT, AWS, DROAD, FIRED, and LAI to $PM_{2.5}$ concentrations were 95.42% and 0.992%, higher than 94.64% and 0.991 provided by all variables. Meanwhile, the computed results of each year are also consistent with those from the 6-year period. After removing the influence of noise from some variables, all the fitting parameters were improved at varying degrees. These results are presented in Fig. S2 and Table 3.

Drivers of $PM_{2.5}$ exposure

Figures 3 and 4 also illustrate the absolute advantage of meteorological factors in affecting $PM_{2.5}$ exposure in the YRD region, whether for the analysis of the 6-year period or the individual annual analysis. As can be seen from Fig. 3, that CP took the leading position in affecting $PM_{2.5}$ concentrations,

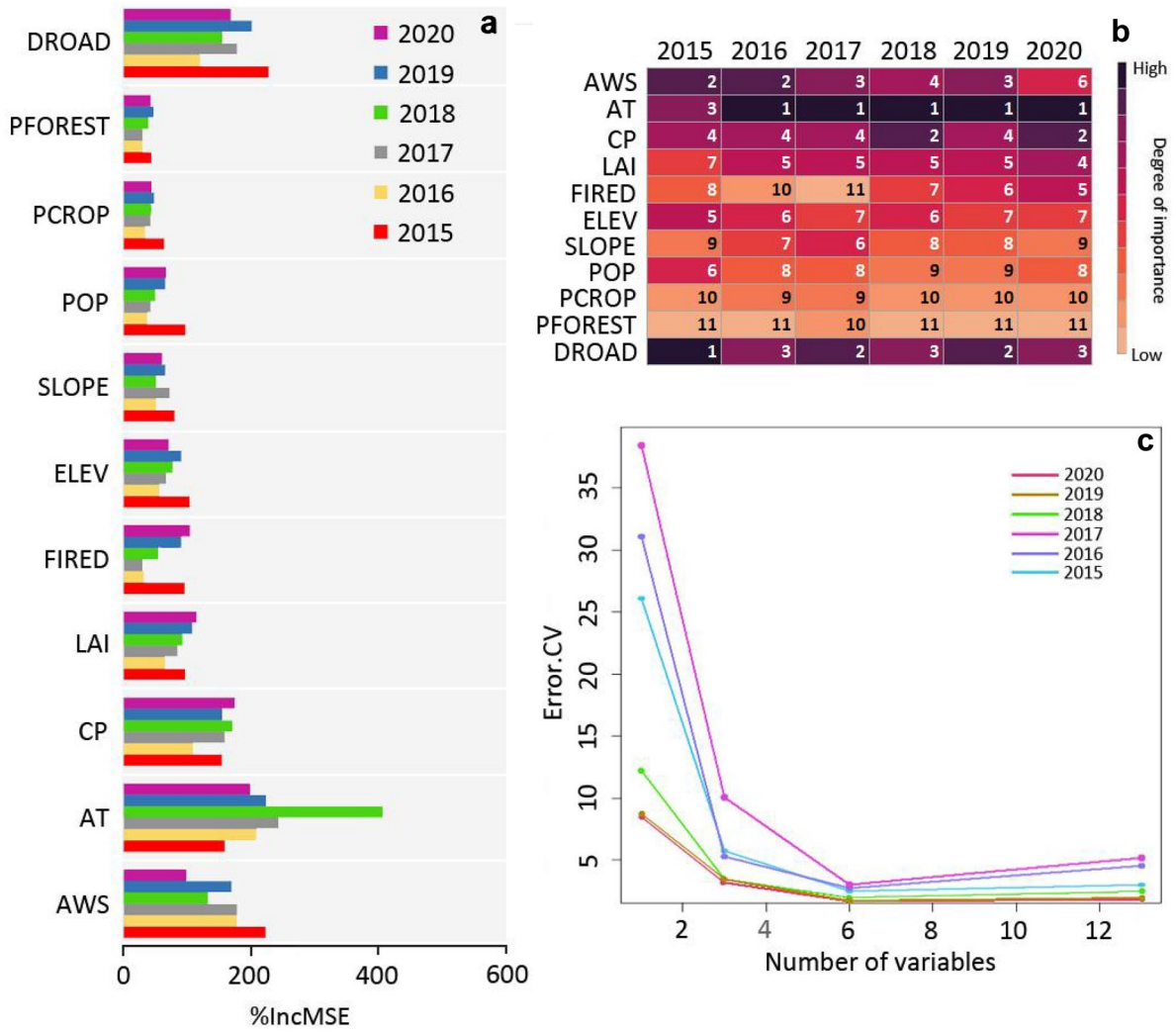

Fig. 4 The effect of annual variable factors by Random Forest on $PM_{2.5}$ concentration. **a** The importance ranking of annual variable factors on $PM_{2.5}$ concentration. **b** The heatmap of annual variable importance. **c** Selection of the optimal number of variables per year

followed closely by AT and AWS in the recent 6 years. DROAD was the only human factor with a large impact on $PM_{2.5}$ in the YRD region. Among the local importance ordering, the above four variables also exhibited the same dominant status trend. As expected, FIRED followed by LAI contributed to the $PM_{2.5}$ concentrations, but to a much lesser extent than the four variables mentioned above.

Figure 4 provides that the information of the influence degree of annual independent variables on $PM_{2.5}$ concentrations was generally consistent with the results of 6-year aggregation. The influence degree of AT, CP, and AWS on $PM_{2.5}$ exposure was very strong yearly. However, it is worth noting that CP had less control over the annual $PM_{2.5}$ concentrations than AT. DROAD was as important to $PM_{2.5}$ as meteorological factors, even more important than AWS and CP. Furthermore, the biggest difference was observed from Fig. 4b that FIRED's contribution to $PM_{2.5}$ was not always very high but signs of growth for the influence on $PM_{2.5}$ over the period 2018 to 2020, which was unlike the results of 6-year consolidated data.

Variable importance provides only an order of how important explanatory variables are in consequence

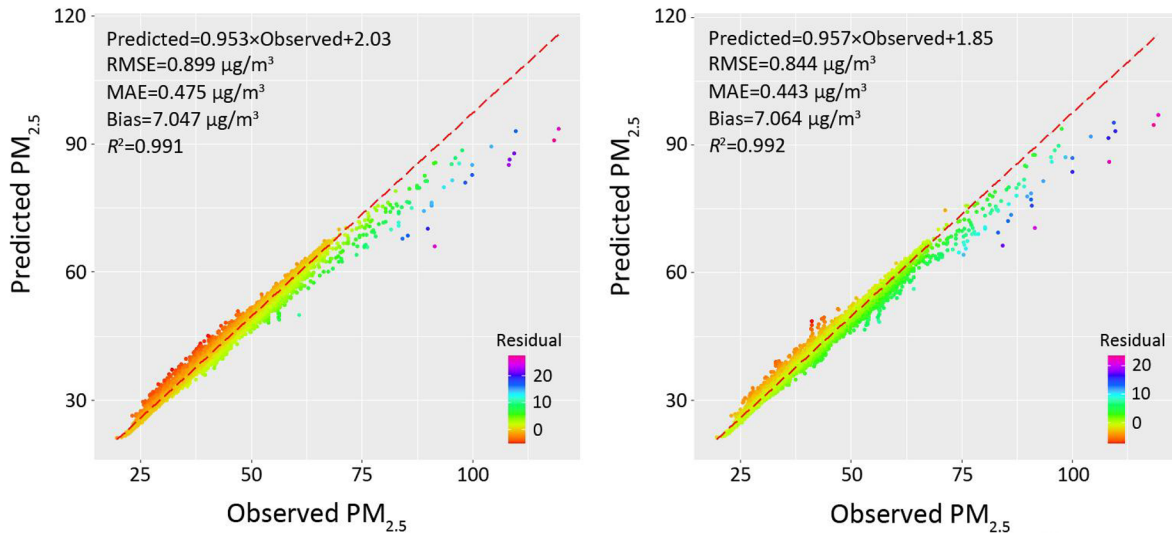

Fig. 5 Scatterplots of the observed and predicted $PM_{2.5}$ from the RF regression modeling (**a**) all the variables and (**b**) optimal variables during 2015–2020

of response variable, while partial dependence plots give a graphical depiction of the marginal effect of a variable on the response (regression), i.e., how each independent variable affects the dependent variables. The partial dependence plots of the six most important variables after variable selection are shown in

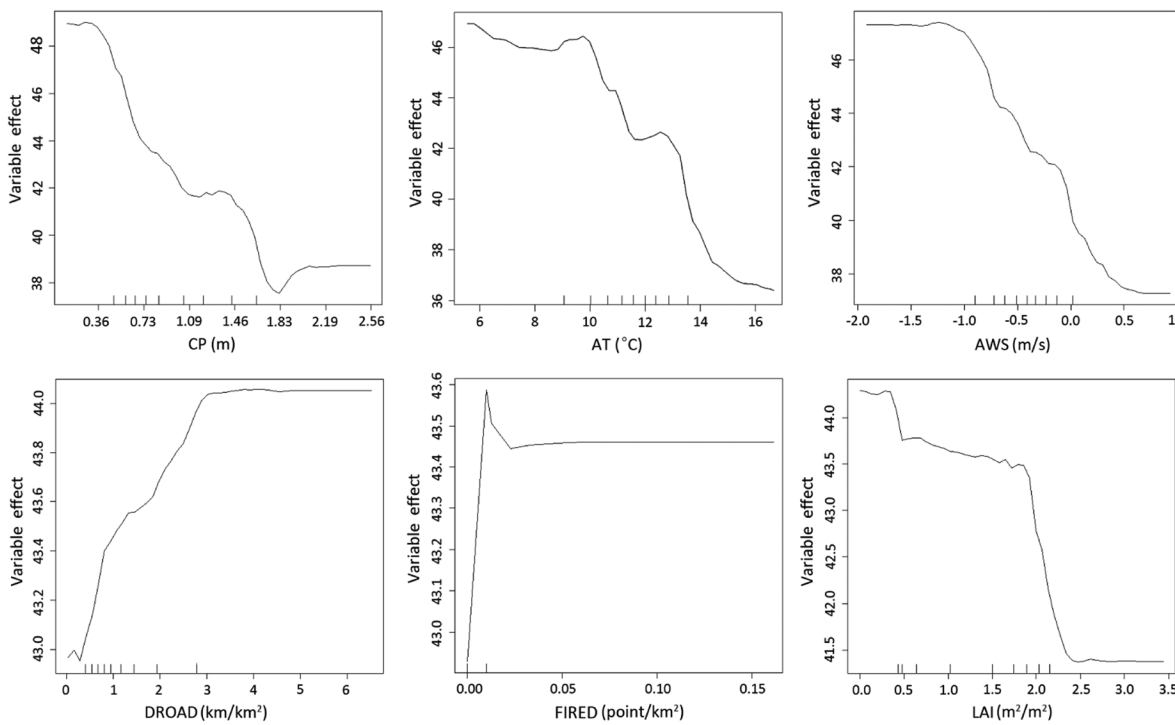

Fig. 6 Partial dependence plots for variables predicting $PM_{2.5}$, selected using Random Forests. Partial plots show the dependence of the concentration of $PM_{2.5}$ on one predictor variable after averaging out the effects of all other predictor variables in the model

Fig. 6. The effect of the six drivers on the range of $PM_{2.5}$ concentrations was different and there was no linear relationship.

CP, AT, and AWS had similar magnitude of influence on $PM_{2.5}$ concentrations over the period 2015 to 2020 in the YRD region (roughly 37–49 μg/m^3, 36–47 μg/m^3, and 37–47.5 μg/m^3, respectively). With the increase of CP value, the exposure of $PM_{2.5}$ experienced a sharp downward trend, after that began to rise slightly at around 1.83 m. Likewise, with the increase of AT, its effect on $PM_{2.5}$ decreased in a fluctuating way. At first, the decline was modest and the mutation appeared at 10℃; after that, the exposure of $PM_{2.5}$ dropped dramatically to 36 μg/m^3. It is noting worthy that the influence range of AWS < 0 on $PM_{2.5}$ concentrations (about 40–46 μg/m^3) was greater than that of AWS > 0 (about 37–40 μg/m^3).

In contrast to the three meteorological factors above, DROAD, LAI, and FIRED did not perform well, and their repercussions on $PM_{2.5}$ exposure were only in a small range (43–44 μg/m^3, 41.5–44 μg/m^3, and 43–43.6 μg/m^3, respectively). Even so, as one of the important variables, the effect of DROAD on $PM_{2.5}$ rose with a large variation observed during the increase of DROAD (Fig. 6). We can also learn that the DROAD in low concentration of $PM_{2.5}$ area was quite low, while some areas with high concentrations of $PM_{2.5}$ fell into high DROAD (Figs. 1b and 2g–l). In the process of increasing LAI, there were two stages of rapid decline (0–0.5 $m^2 \cdot m^{-2}$ and 2–2.5 $m^2 \cdot m^{-2}$, respectively). The $PM_{2.5}$ increased enormously at the low value of FIRED, decreased slightly, and then stabilized by about 43.45 μg/m^3.

Discussion

In this study, spatial autocorrelation techniques were used to analyze the spatial and temporal distribution characteristics of $PM_{2.5}$ concentrations in the YRD region over the period from 2015 to 2020. Spatial clustering analysis shows that $PM_{2.5}$ in the YRD region displays a significant spatial clustering state and HH cluster areas were located in the north and northwest. Some cities from Jiangsu and Anhui province in the northern part of YRD are more heavily industrialized than those cities from Zhejiang in the YRD's south (Chen et al., 2017), which may explain why $PM_{2.5}$ concentrations in the northern part of YRD remained high in recent years. Yang et al. (2020) also confirmed the concentration of high concentration $PM_{2.5}$ in the north and northwest of the Yangtze River Delta and the concentration of low concentration $PM_{2.5}$ in the south through the positive spatial correlation between night time light (NTL) and high concentration $PM_{2.5}$. The combined repercussions from the western Pacific subtropical high and tropical cyclone system are another reason for the distribution of high concentrations of $PM_{2.5}$ in the north and northwest (Liao et al., 2017; Xu et al., 2020). Moreover, the LL cluster areas were found in the south part of the YRD steadily because of low emissions and favorable meteorological conditions for clean air, such as land wind and sea breeze in the southeast YRD coastal cities (She et al., 2017); the forest coverage in the southern YRD, on the other hand, is far higher than the northern region (Fig. 1), conducive to improve air quality (Feng et al., 2017; She et al., 2017).

In the drivers affecting the $PM_{2.5}$ concentrations of the YRD, it has been widely observed in current researches that meteorological factors are dominant (Xu et al., 2020; Yun et al., 2019). The YRD region has a subtropical monsoon climate with obvious seasonal characteristics for precipitation. Some studies reveal that raindrops can absorb the dust in the air. When raindrops and dust fall to the ground by gravity, the number of particles matter in the air decreases accordingly (Hu et al., 2020). The weakening effect of the increase of rainfall on the concentrations of $PM_{2.5}$ in a certain range was also found in our study. Our results also reveal that temperature played a significant role in $PM_{2.5}$ exposure. The impingement of low temperature on high concentrations of $PM_{2.5}$ was stronger, which decreased with increasing temperature. A similar pattern of results was found by Xu et al. (2020). It is widely accepted that the rising temperature increases the height of the mixing layer which helps the vertical diffusion of the atmosphere, thereby providing more space for the dilution of surface pollutants (Murthy et al., 2020). Simultaneously, the turbulent mixing effect from the thermal and dynamic forces of the underlying surface has a direct impact on the migration and transformation of pollutants in the mixed layer (Ma et al., 2019a, b). Temperature is positively associated with open sources of pollution, including soil dust and transport and industrial emissions, and has been used to successfully account

for the change of $PM_{2.5}$ (Xu et al., 2020). Wind plays a vital role in the transport, dilution, and diffusion of $PM_{2.5}$ (Xue et al., 2017). It is interesting that the influence range of westerly (land breeze) speed on $PM_{2.5}$ concentrations (about 40–46 μg/m^3) was larger than that of easterly (sea breeze) (about 37–40 μg/m^3) in this study. One reasonable explanation for this situation is that the YRD is a coastal region, which is greatly affected by sea and land breeze, and the easterly gradually weakens from east to west with the help of the terrain, while the westerly from inland encounters more resistance because of the hinder of high land. In the process, its effect on $PM_{2.5}$ exposure becomes gradually smaller (Xu et al., 2020).

Road density is the only anthropogenic variable that has a relationship with $PM_{2.5}$ exposure in the YRD area. It is a greatly useful index to characterize the urban traffic operation and measure the level of regional social and economic development and the richness of human activities (Zhang et al., 2015). Although we observed a small range of repercussions on $PM_{2.5}$ concentrations, there was still a positive sign. It is obvious that the developed economy and frequent production activities in the YRD region have resulted in the dense traffic trunk line, which reduces the buffer distance between the cities and makes $PM_{2.5}$ diffusible (Yun et al., 2019). In Figs. 1b and 2g–l, the concentration of $PM_{2.5}$ was not high in some high-density road. At higher $PM_{2.5}$ concentration, the effect of high road density became stable, which is due to the stronger influence of meteorological factors (Fig. 6).

A promising finding is that wildfire played an important role on affecting the $PM_{2.5}$ exposure in the YRD region. It has been confirmed that wildfire has a positive impact on the chemical composition and concentrations of $PM_{2.5}$ emitted in the air. Fire season in the YRD region is from November to April. The dry weather in fall and winter increases the risk of wildfire. The smoke and particulate matter produced by biomass combustion increase the concentrations of $PM_{2.5}$ in some local areas (Hu et al., 2014). High frequency and/or severe wildfire bring smoke from biomass burning, which not only affects the total amount of pollutants in the local air, but also with the help of the wind, often brings troubles to the surrounding areas (Liu et al., 2015). Besides, the relatively late appearance of the effect of wildfire on $PM_{2.5}$ exposure was a characteristic worthy of attention (Fig. 4b). This is an important finding in understanding the significantly negative effects of smoke and particulate matter emitted by large-scale wildfires on air quality in recent years (Fann et al., 2018; Landguth et al., 2020).

Leaf area index can reflect plant coverage, canopy structure changes, plant community vitality, and its environmental effects (Ferreira et al., 2020). The results of our study showed a negative response between LAI and $PM_{2.5}$ concentration. As can be seen from Figs. 1c and 2g–l, areas with low concentrations of $PM_{2.5}$ are areas with high LAI values. A popular explanation is that vegetation can effectively reduce the number of $PM_{2.5}$ sources by fixing the soil. At the same time, new findings verify that larger leaf area, branch, and stem surface enhance the efficiency of intercepting or capturing $PM_{2.5}$ in the subtropical broad-leaved or coniferous and broad-leaved mixed forest, thereby inhibiting effectively the concentrations of $PM_{2.5}$ in the air (Liu et al., 2014; Zhang et al., 2020).

Limitations of this study

However, our research still has some limitations. First of all, the interpolation results from the data of monitoring stations were associated with the number of monitors. The shortage of stations has a certain impact on the interpolation results; at the same time, our investigation of the relationship between wind speed and $PM_{2.5}$ needs to be improved. Second, we are aware of the important influence of the social economy on $PM_{2.5}$, but few social and economic factors were considered in this study. In the future related research, we will make a more comprehensive consideration, including GDP (gross domestic product), and POI (point of interest) density. Additionally, attention should be put on more different categories of land-use information in future research. Finally, the RF regression analysis employed in this study is different from the traditional linear regression model or land-use regression model. The latter analysis models can give the positive and negative correlation of variables. We will hence further improve the analysis method of $PM_{2.5}$ concentrations and establish a more comprehensive prediction model in the future.

Conclusion

In conclusion, we have reasons to believe that (1) from the spatial cluster distribution of $PM_{2.5}$ concentrations, the northwest and north of the YRD region from 2015 to 2020 are mostly highly concentrated and surrounded by high concentrations of $PM_{2.5}$, and these areas will be the focus of work related to pollution control. In the future, the local governments need to prevent $PM_{2.5}$ invasion while strengthening the local air prevention and control; (2) RF regression has a high degree of explanation in the simulation of the relationship between $PM_{2.5}$ concentrations and driving factors in the YRD region; (3) meteorological factors are the main drivers of $PM_{2.5}$ emissions in the YRD region over the period 2015 to 2020. It should be noted that the impact of wildfire on $PM_{2.5}$ concentrations was gradually prominent. When formulating air pollution prevention and control measures in the future, the improvement of wildfire management should be taken seriously by YRD's governments and indigenous people, especially in fire season, which will help to reduce $PM_{2.5}$ pollution caused by biomass combustion. At the same time, higher leaf area contributes to improving air quality, and the increasing green area will help reduce air pollution for environmental protection.

Author contribution ZS conceived and designed the experiments; ZS analyzed the data of experiments; ZS and YC collected the data; ZS and YC wrote the original draft; ZS, LL, YC, and HH critically reviewed and edited the manuscript; ZS and LL supervised the whole research and writing process; ZS provided the support of the funding.

Funding Young and Middle-aged Teacher Education Research Project of Fujian Province (JAT201273); Dr. Lin is supported by the US Department of Commerce under National Oceanic and Atmospheric Administration (NOAA) Grant "NA19NES4320003" (Cooperative Institute for Satellite Earth System Studies—CISESS) at the Earth System Science Interdisciplinary Center (ESSIC), University of Maryland.

Availability of data and material The datasets used or analyzed during the current study are available from the corresponding author on reasonable request.

Code availability The current research uses R statistical software, and the related model code is open source. If there are reasonable requirements, they can be obtained from the correspondent author.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest The authors declare no competing interests.

References

Andela, N., Morton, D. C., Giglio, L., Chen, Y., van der Werf, G. R., Kasibhatla, P. S., Defries, R. S., Collatz, G. J., Hantson, S., Kloster, S., Bachelet, D., Forrest, M., Lasslop, G., Li, F., Mangeon, S., Melton, J. R., Yue, C., & Randerson, J. (2017). A human-driven decline in global burned area. *Science, 356*(6345), 1356.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis, 27*(2), 93–115.

Archer, C., Penny, A. L., Templeman, S., & McKenzie, M. (2020). State of the Tropics 2020 Report.

Berg, A., & McColl, K. A. (2021). No projected global drylands expansion under greenhouse warming. *Nature Climate Change, 11*, 331–337.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5–32.

Chen, Q., Yuan, Y., Huang, X., Jiang, Y., & Tan, H. (2017). Estimation of surface-level $PM_{2.5}$ concentrations using aerosol optical thickness through aerosol type analysis method. *Atmospheric Environment, 159*, 26–33.

Chen, Y., Zheng, W., Li, W., & Huang, Y. (2021). Large group activity security risk assessment and risk early warning based on random forest algorithm. *Pattern Recognition Letters, 144*(4), 1–5.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., & Gibson, J. (2007). Random forests for classification in ecology. *Ecology, 88*, 2783–2792.

Dominici, F., Peng, R. D., Bell, M. L., Pham, L., Mcdermott, A., & Zeger, S. L. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA, 295*, 1127–1134.

Fann, N., Alman, B., Broome, R. A., Morgan, G. G., Johnston, F. H., & Pouliot, G. (2018). The health impacts and economic value of wildland fire episodes in the U.S.: 2008–2012. *Science of the Total Environment, 610–611*, 802–809.

Feng, H., Zou, B., & Tang, Y. (2017). Scale- and region-dependence in landscape-$PM_{2.5}$ correlation: Implications for urban planning. *Remote Sensing-Basel, 9*(9), 918.

Ferreira, L. N., Vega-Oliveros, D. A., Zhao, L., Cardoso, M. F., & Macau, E. E. N. (2020). Global fire season severity analysis and forecasting. *Computers & Geosciences-UK, 134*, 104339.

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis, 24*(3), 189–206.

Guo, T., Wang, Y., Zhang, H., Zhang, Y., Zhao, J., Wang, Q., Shen, H., Wang, Y., Xie, X., Wang, L., Xu, Z., Zhang, Y., Yan, D., He, Y., Yang, Y., Xu, J., Peng, Z., & Ma, X. (2018). The association between ambient $PM_{2.5}$ exposure and the risk of preterm birth in China: A retrospective cohort study. *Science of the Total Environment, 633*, 1453–1459.

Hu, W., Zhao, T., Bai, Y., & Kong, S. (2020). Importance of regional $PM_{2.5}$ transport and precipitation washout in heavy air pollution in the Twain-Hu Basin over Central China: Observational analysis and WRF-Chem Simulation. *Science of the Total Environment, 758*, 43710.

Hu, X., Waller, L. A., Lyapustin, A., Wang, Y., & Liu, Y. (2014). Improving satellite-driven $PM_{2.5}$ models with Moderate Resolution Imaging Spectroradiometer fire counts in the southeastern U.S.: Improving $PM_{2.5}$ models with fire counts. *Journal of Geophysical Research: Atmospheres, 119*(19), 11375–11386.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., & Giovis, C. (2005). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Science & Environmental Epidemiology, 15*(2), 185–204.

Joharestani, M. Z., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). $PM_{2.5}$ Prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere, 10*(7), 373.

Landguth, E. L., Holden, Z. A., Graham, J., Stark, B., Mokhtari, E. B., Kaleczyc, E., Anderson, S., Urbanski, S., Jolly, M., Semmens, E. O., Warren, D. A., Swanson, A., Stone, E., & Noonan, C. (2020). The delayed effect of wildfire season particulate matter on subsequent influenza season in a mountain west region of the USA. *Environment International, 139*, 105668.

Liao, Z., Gao, M., Sun, J., & Fan, S. (2017). The impact of synoptic circulation on air quality and pollution-related human health in the Yangtze River Delta region. *Science of the Total Environment, 607–608*, 838–846.

Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. *R News., 2*(3), 18–22.

Liu, J. C., Pereira, G., Uhl, S. A., Bravo, M. A., & Bell, M. L. (2015). A systematic review of the physical health impacts from non-occupational exposure to wildfire smoke. *Environmental Research, 136*, 120–132.

Liu, N., Zou, B., Li, S., Zhang, H., & Qin, K. (2021). Prediction of $PM_{2.5}$ concentrations at unsampled points using multiscale geographically and temporally weighted regression. *Environmental Pollution, 284*, 117116.

Liu, P., Wang, Q., Zhang, D., & Lu, Y. (2020). Remote sensing an improved correction method of nighttime light data based on EVI and WorldPop data. *Remote Sensing-Basel, 12*, 3988.

Liu, Z., Wang, Y., Liu, Q., Hu, B., & Sun, Y. (2014). Source apportionment of urban fine particle number concentrations during summertime in Beijing. *Atmospheric Environment, 13*, 1367–1397.

Lloyd, C. T., Sorichetta, A., & Tatem, A. J. (2017). High resolution global gridded data for use in population studies. *Scientific Data, 4*, 170001.

Ma, T., Duan, F., He, K., Qin, Y., Tong, D., Geng, G., Liu, X., Li, H., Yang, S., Ye, S., Xu, B., Zhang, Q., & Ma, Y. (2019a). Air pollution characteristics and their relationship with emissions and meteorology in the Yangtze River Delta region during 2014–2016. *Journal of Environmental Sciences, 83*, 8–20.

Ma, X., Longley, I., Gao, J., Kachhara, A., & Salmond, J. (2019b). A site-optimised multi-scale GIS based land use regression model for simulating local scale patterns in air pollution. *Science of the Total Environment, 685*, 134–149.

Matz, C. J., Marika, E., Xi, G., Racine, J., Pavlovic, R., Rittmaster, R., Henderson, S. B., & Stieb, D. M. (2020). Health impact analysis of $PM_{2.5}$ from wildfire smoke in Canada (2013–2015, 2017–2018). *Science of the Total Environment, 725*, 138506.

Mercer, L. D., Szpiro, A. A., Sheppard, L., Lindström, J., Adar, S. D., Allen, R. W., Avol, E. L., Oron, A. P., Larson, T., Liu, L. J. S., & Kaufman, J. D. (2011). Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NO_x) for theMulti-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Atmospheric Environment, 45*(26), 4412–4420.

Miskell, G., Salmond, J. A., & Williams, D. E. (2017). Use of a handheld low-cost sensor to explore the effect of urban design features on local-scale spatial and temporal air quality variability. *Science of the Total Environment, 619–620*, 480–490.

Mitchell, A. (2005). The ESRI guide to GIS analysis, vol. 2. ESRI Press.

Murthy, B. S., Latha, R., Tiwari, A., Rathod, A., Singh, S., & Beig, G. (2020). Impact of mixing layer height on air quality in winter. *Journal of Atmospheric and Solar-Terrestrial Physics, 197*, 105157.

Nethery, R. C., Rushovich, T., Peterson, E., Chen, J. T., Waterman, P. D., Krieger, N., Waller, L., & Coull, B. A. (2021). Comparing denominator sources for real-time disease incidence modeling: American Community Survey and WorldPop. *SSM – Population Health, 14*, 100786.

Pang, Y., Huang, W., Luo, X., Chen, Q., Zhan, Z., Tang, M., Hong, Y., Chen, J., & Li, H. (2020). In-vitro human lung cell injuries induced by urban $PM_{2.5}$ during a severe air pollution episode: variations associated with particle components. *Ecotoxicology and Environmental Safety, 206*, 111406.

Peng, R. D., Bell, M. L., Geyh, A. S., Mcdermott, A., Zeger, S. L., Samet, J. M., & Dominici, F. (2009). Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution. *Environmental Health Perspectives, 117*, 957–963.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A. J., Freire, S., Stamatia, H., Julea, A., Kemper, T., Pierre, S., & Syrris, V. (2016). Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. EUR 27741. Luxembourg (Luxembourg): Publications Office of the European Union; 2016. JRC97705.

She, Q., Peng, X., Xu, Q., Long, L., Wei, N., Liu, M., Jia, W., Zhou, T., Han, J., & Xiang, W. (2017). Air quality and its response to satellite-derived urban form in the Yangtze River Delta, China. *Ecological Indicators, 75*, 297–306.

Su, Z., Zheng, L., Luo, S., Tigabu, M., & Guo, F. (2021). Modeling wildfire drivers in Chinese tropical forest ecosystems using global logistic regression and geographically weighted logistic regression. *Natural Hazards, 108*, 1317–1345.

Velásquez Ciro, D., Cañón Barriga, J. E., & Hoyos Rincón, I. C. (2021). The removal of $PM_{2.5}$ by trees in tropical Andean metropolitan areas: An assessment of

environmental change scenarios. *Environmental Monitoring and Assessment, 193*(7), 396.

Xu, G., Ren, X., Xiong, K., Li, L., Bi, X., & Wu, Q. (2020). Analysis of the driving factors of $PM_{2.5}$ concentrations in the air: A case study of the Yangtze River Delta, China. *Ecological Indicators, 110*, 105889.

Xu, H., Bechle, M. J., Wang, M., Szpiro, A. A., Vedal, S., Bai, Y., & Marshall, J. D. (2019). National $PM_{2.5}$ and NO_2 exposure models for China based on land use regression, satellite measurements, and universal kriging. *Science of the Total Environment, 655*, 423–433.

Xue, T., Zheng, Y., Geng, G., Zheng, B., Jiang, X., Zhang, Q., & He, K. (2017). Fusing observational, satellite remote sensing and air quality model simulated data to estimate spatiotemporal variations of $PM_{2.5}$ exposure in China. *Remote Sensing-Basel, 9*(3), 221.

Yang, D., Chen, Y., Miao, C., & Liu, D. (2020). Spatiotemporal variation of $PM_{2.5}$ concentrations and its relationship to urbanization in the Yangtze river delta region, China. *Atmospheric Pollution Research, 11*(3), 491–498.

Yun, G., He, Y., Jiang, Y., Dou, P., & Dai, Q. (2019). $PM_{2.5}$ spatiotemporal evolution and drivers in the Yangtze River Delta between 2005 and 2015. *Atmosphere, 10*(2), 55.

Zhan, Y., Luo, Y., Deng, X., Zhang, K., Zhang, M., Grieneisen, M. L., & Di, B. (2018). Satellite-based estimates of daily NO_2 exposure in China using hybrid random forest and spatiotemporal Kriging model. *Environmental Science and Technology, 52*(7), 4180.

Zhang, R., Li, L., Zhang, Y., Huang, F., Li, J., Liu, W., Mao, T., Xiong, Z., & Shangguan, W. (2021). Assessment of agricultural drought using soil water deficit index based on ERA5-land soil moisture data in four southern provinces of China. *Agriculture, 11*(5), 411.

Zhang, X., Lyu, J., Han, Y., Sun, N., Sun, W., Li, J., Liu, C., & Yin, S. (2020). Effects of the leaf functional traits of coniferous and broadleaved trees in subtropical monsoon regions on $PM_{2.5}$ dry deposition velocities. *Environmental Pollution, 265*, 114845.

Zhang, Y., Li, X., Wang, A., Bao, T., & Tian, S. (2015). Density and diversity of OpenStreetMap road networks in China. *Journal of Urban Management, 4*(2), 135–146.

Zhou, W., Wu, X., Ding, S., Ji, X., & Pan, W. (2021). Predictions and mitigation strategies of $PM_{2.5}$ concentrations in the Yangtze River Delta of China based on a novel nonlinear seasonal grey model. *Environmental Pollution, 276*, 116614.