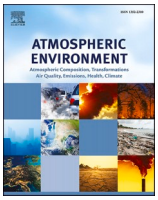
Atmospheric Environment 313 (2023) 120043

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv



A similarity distance-based space-time random forest model for estimating $PM_{2.5}$ concentrations over China

Shikang Guan, Xiaotong Zhang *, Wenbo Zhao, Yanjun Duan, Shuyue Yang, Yunjun Yao, Kun Jia

State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing, 100875, China

HIGHLIGHTS

- A similarity distance-based space-time random forest is used for $PM_{2.5}$ estimation.
- The similarity distance indicator contributes to improving the estimation accuracy.
- The model outperforms other widely used regression models in predictive power.
- $PM_{2.5}$ concentrations at 0.1° are generated based on the latest MODIS AOD (C6.1).

ARTICLE INFO

Keywords:
$PM_{2.5}$
Aerosol optical depth
Spatial-temporal heterogeneity
Similarity distance-based space-time random forest model

ABSTRACT

China has experienced persistent fine particulate matter ($PM_{2.5}$) pollution for the past few years, which adversely affects both physical and mental health. The availability of high-accuracy and full-coverage $PM_{2.5}$ products will be of substantial value in formulating effective policies to combat and regulate $PM_{2.5}$ pollution. Therefore, we have developed a similarity distance-based space-time random forest (SDSTRF) model to estimate daily $PM_{2.5}$ concentrations over China by integrating surface measurements, satellite aerosol products, meteorological data, and auxiliary information. The proposed model not only accounts for spatial-temporal heterogeneity, but also uses the similarity distance to avoid errors caused by outliers. It has undergone rigorous validation through three different cross-validation (CV) approaches and has shown high and stable accuracy, particularly in the site-based CV with a coefficient of determination (R^2) of 0.87, and a root mean square error (RMSE) of 10.68 μg/m^3, along with a relative RMSE (rRMSE) of 27.48%. In addition, the leave-out data are predicted to determine if the SDSTRF model is accurate in its predictive power, which achieves an impressive site-based CV R^2 of 0.80 and an RMSE of 12.89 μg/m^3, along with an rRMSE of 33.01%. The results from a variety of validation approaches strongly indicate that the SDSTRF model can provide accurate estimation of $PM_{2.5}$ concentrations at different time scales and outperform many other space-time models by incorporating similarity distance. The proposed model would be a promising application in air pollution studies with remote sensing.

1. Introduction

Numerous epidemiological studies have identified $PM_{2.5}$, which refers to airborne particles with a tiny diameter not exceeding 2.5 μm, as responsible for a diverse range of respiratory diseases (Baccarelli et al., 2008; Haikerwal et al., 2015; Hamra et al., 2014; Pope and Dockery, 2006). Nearly 1 million people died from the severe haze in December 2013, which blanketed almost all regions in north China (Guan et al., 2016; Yue et al., 2020). The shock caused by the lethality of $PM_{2.5}$-induced chronic diseases led to the accelerated construction of monitoring stations in China to measure particulate levels in the air. Gradually, a whole ground-level air pollutant observation network has been established, facilitating air quality assessment. Although it has satisfactory temporal continuity for recording once per hour, the ground-level $PM_{2.5}$ observation network lacks spatial representativeness because the ground stations are sparsely distributed. As a result, while the forecasting performance works well at the local scale, its accuracy diminishes at regional and global scales. Nowadays, there are more than

* Corresponding author.
E-mail address: xtngzhang@bnu.edu.cn (X. Zhang).

https://doi.org/10.1016/j.atmosenv.2023.120043
Received 24 April 2023; Received in revised form 13 August 2023; Accepted 16 August 2023
Available online 19 August 2023

2000 $PM_{2.5}$ stations in China. However, western China is still underrepresented in comparison to eastern China, resulting from the higher cost of stations and sparser population (Li et al., 2020c). In addition to surface $PM_{2.5}$ measurements, remote sensing is an innovative approach to expand station monitoring thanks to its extensive spatial coverage at regional and global scales (Hoff and Christopher, 2009; Lary et al., 2014). Integrating remote sensing data allows for continuous monitoring of $PM_{2.5}$ concentrations in both space and time. This will facilitate policies that can moderate air pollution and safeguard public health.

As of now, numerous studies consider the fusion of satellite observations and surface measurements as the most promising approach to obtain high-accuracy, large-scale, and long-term $PM_{2.5}$ products (Al-Saadi et al., 2005; Chu et al., 2003; Hutchison et al., 2004). The aerosol optical depth (AOD), a measure of light beam attenuation through the atmosphere, exhibits a strong correlation with $PM_{2.5}$ (Kumar et al., 2007; van Donkelaar et al., 2006). As a result, several models have been developed to establish the most accurate relationship between them. These models fall into three categories: simulated (Geng et al., 2015; Li et al., 2020a; Lin et al., 2015), semi-empirical (Chu et al., 2013; Park et al., 2022; Tian and Chen, 2010), and statistical models (Boyouk et al., 2010; Wang, 2003). Using physical and chemical transport models, the simulated models provide complete spatial coverage of $PM_{2.5}$ concentrations on regional or global scales (Drury et al., 2010; Lee et al., 2022; van Donkelaar et al., 2010). However, uncertainties arising from input data (e.g., anthropogenic emissions) and physical models (e.g., aerosol dynamics) have an impact on their accuracy (Kukkonen et al., 2012; Zheng et al., 2009). Based on physical models, the semi-empirical models integrate meteorological variables such as humidity and surface pressure (Emili et al., 2010; Tao et al., 2013). However, their parameters will vary with space and time because $PM_{2.5}$ is a multi-factorial pollutant. At present, the statistical models are most applicable for $PM_{2.5}$ estimates because they have fast and adaptive learning characteristics as well as relatively high accuracy. At first, the most commonly used method is to use the simple linear regression (LR) model to establish the AOD-$PM_{2.5}$ relationship. For example, Wang (2003) employed this model to quantitatively estimate local air quality categories and reported a linear correlation coefficient of 0.7 for AOD-$PM_{2.5}$. Subsequently, the multiple linear regression (MLR) models were established by introducing meteorological data to enhance the relationship (Al-Saadi et al., 2005; Gupta and Christopher, 2009). To further obtain considerable accuracy of the estimates, Boyouk et al. (2010) first developed an improved linear AOD-$PM_{2.5}$ model that accounted for the effects of aerosol particles on the AOD-$PM_{2.5}$ relationship, which varies with relative humidity and satellite height. Although considerable effort has been invested in building the most accurate linear models, the AOD-$PM_{2.5}$ relationships derived from these models rely on a large number of ground observations and the existing data remain insufficient to achieve significant improvements. Furthermore, these methods have only been applied at the local scale due to the nonlinearity and spatial-temporal heterogeneity that exist in the AOD-$PM_{2.5}$ relationship (Chu et al., 2015; Merbitz et al., 2012).

Recent studies have shown that machine learning (ML) can resolve the AOD-$PM_{2.5}$ relationship more effectively than traditional linear models (Li et al., 2017b). There have been a variety of ML models used to fit the nonlinear relationship, such as the support vector machine (Dong et al., 2016; Weizhen et al., 2014), random forest (Brokamp et al., 2018; Guo et al., 2021a; Park et al., 2020), and gradient boosting regressor (Chen et al., 2019; Gui et al., 2020). Considering the fact that the nonlinear relationships in AOD-$PM_{2.5}$ are different in both space and time domains (Kumar, 2010; Miller, 2004), a single domain of interest can't fully explain this characteristic of the relationship. Tremendous investigations have already been conducted to address this issue by incorporating various spatial and temporal variations into estimation models for local fitting. In an effort to account for the spatial heterogeneity in AOD-$PM_{2.5}$, Hu et al. (2013), for example, proposed the geographically weighted regression (GWR) model by solving weights for each local sample. Moreover, some more complex models take temporal variation into account to enhance the relationship. For instance, Hu et al. (2014) introduced a two-stage model that incorporates both temporal and spatial information using linear mixed effects (LMEs) and GWR, respectively. Since the spatial and temporal variations interact with each other (Choi et al., 2009), the space and time information should be embedded in ML simultaneously, which enables the nonlinear model to gain the capacity to handle the complex relationship between spatial heterogeneity and temporal variation. The space-time random forest (STRF) model (Wei et al., 2019) and the geo-intelligent deep belief network (Geoi-DBN; Li et al., 2017a) have successfully adopted this strategy and achieved a high level predictive performance. This approach is expected to have considerable potential to generate high-accuracy and space-time continuous $PM_{2.5}$ products.

At present, several $PM_{2.5}$ datasets have been released by fusing surface measurements, aerosol products, and other ancillary data based on the proposed methods. Xue et al. (2019) generated a daily $PM_{2.5}$ product at 0.1° resolution for China from 2000 to 2016, using a model with the high-dimensional expansion of $PM_{2.5}$ predictors. Wei et al. (2021) generated a ChinaHigh$PM_{2.5}$ dataset with 1-km resolution from 2000 to 2018 with the proposed Space-Time Extra-Trees model. In addition, an open-access air pollution database, Tracking Air Pollution in China, keeps track of $PM_{2.5}$ concentrations in near-real time and provides 1-km and 10-km products from 2000 to the present (Geng et al., 2021; Liu et al., 2022; Xiao et al., 2022). All of these works have achieved excellent performance and are of vast value, but a persistent challenge remains in underestimating $PM_{2.5}$ concentrations on high pollution days, mainly due to the limited amount of samples available for cases with extremely high pollution levels (e.g., $PM_{2.5} > 150$ μg/m^3) and the high nonlinearity in AOD-$PM_{2.5}$ (Wei et al., 2020). Therefore, the estimation model for such a complete $PM_{2.5}$ dataset needs to be further improved, specifically with high accuracy, full spatial coverage, and a broad time span.

In order to improve the accuracy of the $PM_{2.5}$ estimates, this study introduces a similarity distance-based space-time random forest (SDSTRF) model. It is constructed and evaluated by integrating surface measurements, aerosol products, meteorological data, and auxiliary information from China. The model performance is validated by using sample-, time-, and site-based cross-validation (CV) approaches. This study provides a novel perspective to investigate the spatial-temporal heterogeneity of $PM_{2.5}$ with high precision.

2. Study area and data

2.1. Study area

Mainland China is chosen as the focal area of interest, where most cities have become dominated by $PM_{2.5}$ pollution. According to the data in *the Report on the State of the Environment in China 2022*, the annual average $PM_{2.5}$ level is 31 μg/m^3 nationwide, which is slightly better than the acceptable standard recommended by WHO (i.e., 35 μg/m^3). However, considering the densely populated areas, the exposure to $PM_{2.5}$ for a single person is far more severe than the target (Wang et al., 2019). As a result, a great deal of importance should be attached to monitoring $PM_{2.5}$ concentrations across mainland China.

Fig. 1 depicts the distribution of ground stations across mainland China used for the analysis. We randomly divided 80% of all stations into training sites while the remaining 20% were assigned as test sites. A total of 1566 $PM_{2.5}$ monitoring stations were finally collected for the study.

2.2. Data

The data used for analysis consist of ground observations, MODIS AOD products, meteorological data, and $PM_{2.5}$-related auxiliary information such as MODIS normalized difference vegetation index (NDVI) and digital elevation model (DEM). Table 1 provides a comprehensive

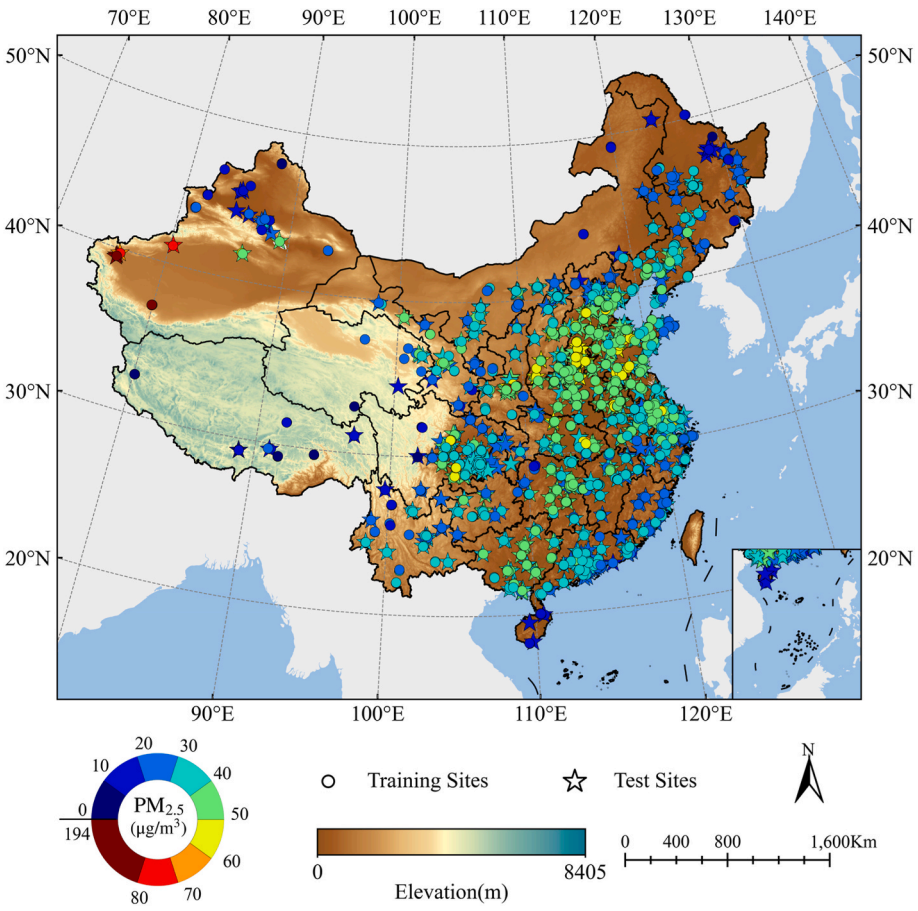


Fig. 1. The distribution of ground $PM_{2.5}$ stations (the circles represent the sites designated for training and validation, while the stars indicate the test sites used to assess the model's predictive power) across mainland China. The colormap table of stations reflects the average $PM_{2.5}$ levels from 2018 to 2022 at each respective location.

Table 1
Summary of the data sources used in this study.

Dataset	Variable	Content	Spatial Resolution	Temporal Resolution	Unit
$PM_{2.5}$	$PM_{2.5}$	Fine particulate matter	-	Hourly	ug/m^3
MO(Y)D04	AOD	Aerosol optical depth at 550 nm	10 km × 10 km	Daily	-
MO(Y)D13	NDVI	Normalized difference vegetation index	1 km × 1 km	8-day	-
SRTM	DEM	Digital elevation model	250 m × 250 m	-	m
ERA5	SP	Surface pressure	0.25° × 0.25°	1-h	hPa
	TEM	2m air temperature	0.25° × 0.25°	1-h	K
	WS	Wind speed	0.25° × 0.25°	1-h	m/s
	WD	Wind direction	0.25° × 0.25°	1-h	degrees
	RH	2m relative humidity	0.25° × 0.25°	1-h	%
	BLH	Boundary layer height	0.25° × 0.25°	1-h	m

overview of all the variables utilized in the study.

2.2.1. Surface $PM_{2.5}$ measurements

The hourly surface $PM_{2.5}$ measurements, which have already been calibrated and quality controlled prior to access, are collected between January 1, 2018 and December 31, 2022. The measurements that remains constant for more than consecutive 3 h or are missing for more than 12 h in a day are removed at each station due to probable instrument failure (Rohde and Muller, 2015). The hourly $PM_{2.5}$ data from the same day, corresponding to the satellite transit time, are then selected and added to the average for modeling. The mean value for each station is calculated from 2018 to 2022, reflecting the average levels of $PM_{2.5}$ at the site and its surroundings (see Fig. 1).

2.2.2. Aerosol products

Considerable work has been undertaken to derive $PM_{2.5}$ concentrations utilizing the MODIS atmosphere L2 aerosol products, which are collected by the Terra and Aqua satellites (Ma et al., 2016; Rohde and Muller, 2015). The products at 10 km spatial resolution are included in this study, and their contained AOD datasets, obtained from the combination of the dark target (Levy et al., 2013) and deep blue (Hsu et al., 2013) algorithms, are applied to build the model for $PM_{2.5}$ estimates. The datasets from the same day are mosaicked together as input to the built model. For each pixel, if only one is valid, it is selected. In the case of pixels corresponding to both datasets, the valid values are added to the average as final input. However, if there is no valid value, that pixel is regarded as missing and can't be used in the estimation model.

2.2.3. Meteorological data

ERA5 is the fifth generation ECMWF atmospheric reanalysis and serves as the successor of ERA-Interim. It assimilates a wealth of historical observations related to the atmosphere, ocean, and land (Hersbach et al., 2020). Meanwhile, it is capable of providing meteorological data which covers a large part of the globe in 1 h at a spatial resolution of

0.25° longitude × 0.25° latitude. ERA5 data have been extensively used in $PM_{2.5}$ estimation and shown superior performance compared to other global atmospheric reanalysis datasets, such as MERRA-2, ERA-Interim, and FNL (Cho et al., 2023; Guo et al., 2021b; Zuo et al., 2023). The performance comparison of estimating $PM_{2.5}$ concentration with MERRA-2 and ERA5 is displayed in Fig. S1. In this study, the following meteorological variables are used: surface pressure (SP, unit: kPa), air temperature at 2 m height (TEM, unit: K), wind speed (WS, unit: m/s), wind direction (WD, unit: degrees), relative humidity (RH, unit: %), and boundary layer height (BLH, unit: m).

2.2.4. Auxiliary data

The MODIS NDVI products are used as a variable related to land cover in the estimation model for $PM_{2.5}$. The NDVI products are sourced from the same database as the AOD products. Each data represents that the type of land cover remains unchanged within a 8-day period. In addition, the Shuttle Radar Topography Mission (SRTM) DEM with a spatial resolution of 250 m is adopted as a terrain-based variable for $PM_{2.5}$ mitigation (Saide et al., 2011).

3. Methodology

3.1. Random forest

The random forest (RF) model is composed of many individual decision trees, and it has been widely used in classification and regression problems due to its superiority in fast convergence and accurate fitting results. In addition, each decision tree of the RF is highly independent, so it can build decision trees in parallel so as to develop the RF quickly (Chen et al., 2017). The key steps involved in the establishment of the RF model are as follows (Breiman, 2001): 1). Randomly extract the same quantity of samples from the original datasets by the bootstrap sample method several times; 2). Randomly extract partial features from the whole for each tree node, then sort these candidate features according to the rules selected by the decision tree growth algorithm, and then select a feature as an attribute to split the node. The decision tree can grow after this process is complete, which is the most important step in building an RF model; 3). Not prune each tree to maximize its growth, and finally form a random forest by integrating all decision trees. The RF model is widely used in various fields, and it has become increasingly popular in the fields of remote sensing and atmospheric science. It has been previously reported for applications in $PM_{2.5}$ estimation (Brokamp et al., 2018; Hu et al., 2017). However, RF requires space-time information to tackle spatial-temporal heterogeneity for higher estimation accuracy.

3.2. Similarity distance-based space-time random forest (SDSTRF) model

3.2.1. Space-time information

There is a significant effect of spatial-temporal heterogeneity on $PM_{2.5}$, and tremendous efforts have been made to resolve this problem by taking space-time information into consideration (Hu et al., 2013, 2014). For instance, Gongbo et al. (2018) tried to input the geographical location and the day of the year (DOY) directly into the machine learning model. However, the coupling effect between space and time information is not considered in this method. The space-time information for one pixel can be extracted from its adjacent pixels, since nearby things are more correlated with each other. Many desirable models have been developed by this method, such as GWR, the two-stage, GTWR, Geo-i DBN, and STRF. Among them, the STRF model has an excellent performance in fitting effect and generalization capacity by calculating the inverse distance weight (IDW) of the central pixel from adjacent pixels. The SDSTRF model extracts the space-time information in a similar approach. For a given pixel, the process of extracting spatial (S_{ij}) and temporal (T_{ij}) information can be described as follows:

$$S_{ij} = \frac{\Sigma_m^M \Sigma_n^N \frac{1}{ds_{mn}^2} Ps_{mn}}{\Sigma_m^M \Sigma_n^N \frac{1}{ds_{mn}^2}} \tag{1}$$

$$T_{ij} = \frac{\Sigma_l^L \frac{1}{dt_{ij}^2} Pt_{ij}^l}{\Sigma_l^L \frac{1}{dt_{ij}^2}} \tag{2}$$

where ds and dt are the distances in space and time, the M and N are the m rows and n columns of pixels adjacent to the j_{th} column in the i_{th} row, and the L is the l prior days at the same location. This method collects adjacent pixel values to generate the spatial vector of square windows in an odd shape (e.g., 3 × 3 and 5 × 5) and the temporal vector of the same length with spatial window size for weighting, which are P_s and P_t respectively.

3.2.2. Similarity distance

Using the STRF model to capture the distribution characteristics of $PM_{2.5}$ in space and time can not only account for the global variation of $PM_{2.5}$, but also achieve a good fitting effect at the local scale. However, two shortcomings undermine the performance of the model. The first is that the AOD products themselves are largely missing, which significantly affects the $PM_{2.5}$ estimation. The second is that the AOD products and the meteorological products have huge resolution discrepancies, which means that it is inevitable that the part of adjacent pixels from the central pixel exists as outliers when resampling to a consistent resolution. The final fitting effect and generalization capacity will drop if these outliers are input into the model. Therefore, the Jaccard similarity coefficient is introduced in the model construction to identify and exclude outliers. We use the complementary of the indicator as an assessment of the differences between pixels, which is the similarity distance. The similarity distance (E_{sd}) is expressed as:

$$E_{sd}(x,y) = 1 - E_j(x,y) = 1 - \frac{x_j * y}{x_j^2 + y^2 - x_j * y} \tag{3}$$

where E_j is the Jaccard similarity coefficient, x_j is the vector of adjacent pixels and y is the central pixel. The E_{sd} ranges from 0.1 to 0.9, with increments of 0.2 for each iteration to improve training efficiency. If any adjacent pixel falls outside the similarity distance, the geographical weighting should not be performed, and that pixel should be excluded from the spatiotemporal vectors. In such cases, the IDW is performed only on the adjacent pixels within the similarity distance to obtain S_{adj} and T_{adj} to improve the performance of the final model, which in turn determines the optimal setting of the similarity distance (see Fig. S2).

3.2.3. Model construction

Data preprocessing was performed to yield a spatially and temporally uniform input dataset for the SDSTRF model. As a first step, the correlation analysis was conducted on AOD and $PM_{2.5}$ and obtained a correlation coefficient of 0.473 (R), which indicates that there is not a strong enough correlation between them in China, and other related variables should be incorporated to strengthen the relationship. Therefore, variables that contribute significantly to $PM_{2.5}$ concentrations were chosen (i.e., AOD, SP, TEM, WS, WD, RH, BLH, NDVI, and DEM; see Fig. S3). Secondly, all involved variables were uniformly reprojected to the same coordinate system as the AOD data and then resampled to a spatial resolution of 0.1° through the bilinear interpolation method. For consistency in temporal resolution, the selected meteorological variables were averaged on a daily scale aligned with the satellite transit time. The NDVI values were updated promptly whenever changes occurred within an 8-day interval in the corresponding grid cell, while the DEM remained constant throughout the specified time period. Finally, to match the surface $PM_{2.5}$ measurements each day, data for these variables were extracted from the grid cells in which the monitoring stations were located. In the meantime, the ground observations

of $PM_{2.5}$ were averaged for the same grid cell. After completing the data preprocessing, a total of 639,649 records were collected for all days from 2018 to 2022.

The estimation of the multiple regression model tends to be distorted due to the high correlation between the explanatory variables (e.g., $R > 0.8$), which is called multicollinearity (Kalnins, 2018). There may be an issue of multicollinearity between SP and DEM ($R = -0.938$). In order to determine the true relationship between $PM_{2.5}$ and its explanatory variables, the variance inflation factor (VIF) is adopted as a method for diagnosing multicollinearity between chosen variables. It is assumed that there is a collinearity problem if the VIF value is greater than 10 (Sheather, 2009). All the variables meet the criterion, including SP and DEM with VIF values of 8.984 and 9.011 respectively. This indicates that these variables can be selected as the suitable inputs for the model in $PM_{2.5}$ estimates (see Table 2).

After finishing the selection of input variables, the model can be constructed, the structure of which is shown as follows:

$$PM_{2.5} = f\left(S_{adj}, T_{adj}, AOD, SP, TEM, WS, WD, RH, BLH, NDVI, DEM\right) \quad (4)$$

Fig. 2 exhibits the schematic of the SDSTRF model for estimating $PM_{2.5}$ concentrations.

3.3. Model evaluation

An evaluation of the estimation accuracy for the developed model is performed using the 10-CV approach (Rodriguez et al., 2010). The samples are equally and randomly split into ten folds, of which nine folds serve as training samples and the remaining fold serves as validation samples. By repeating this process 10 times, each fold is validated. In general, there are three types of CV approaches, including sample-, time-, and site-based CV, where the partition criteria are the original samples, the valid sample days, and the cells in which the monitoring stations are located, respectively. Their results should be indicative of overall, temporal, and spatial predictive performance. Historical validation is also conducted by predicting $PM_{2.5}$ concentrations for a specific period (e.g., July to December 2022). Meanwhile, an external validation approach is introduced to show the extent to which the spatial and temporal variations affect the estimation model performance. In addition, a quantitative evaluation of the models is conducted using the following statistical indicators: linear regression equation (slope, unitless; intercept, μg/m^3), coefficient of determination (R^2, unitless), root mean square error (RMSE, unit: μg/m^3), relative root mean square error (rRMSE, unit: %), and mean bias error (MBE, unit: μg/m^3).

4. Results and analysis

4.1. Validation at different spatial scales

4.1.1. Validation at the national scale

A comparison of $PM_{2.5}$ estimation results from three different models (RF, STRF, and SDSTRF) with those derived from ground measurements at the national scale is conducted using the sample-, time-, and site-based CV approaches. The density scatterplots in Fig. 3 present the training and 10-CV results. The model training results indicate that the three models have similar performance, achieving an identical R^2 of 0.98 (RF is 0.97). The corresponding RMSE are 4.75, 4.48, and 4.41 μg/m^3, with rRMSE of 12.12%, 11.45%, and 11.27%, respectively. These results suggest that the RF, STRF, and SDSTRF models all effectively capture the essential features of the dataset. For sample-, time-, and site-based CV approaches, the SDSTRF model shows relatively better performance than the other two models in terms of overall R^2, RMSE, and rRMSE. A higher R^2 of 0.83, a lower RMSE of 11.83 μg/m^3, and a lower rRMSE of 30.29% are achieved by the SDSTRF model for sample-based CV, as compared to either the STRF model ($R^2 = 0.82$, RMSE = 12.06 μg/m^3, and rRMSE = 30.89%) or the RF model ($R^2 = 0.80$, RMSE = 12.81 μg/m^3, and rRMSE = 32.81%). For time-based CV, the SDSTRF model has an R^2 of 0.84, an RMSE of 11.68 μg/m^3, and an rRMSE of 29.79%, which are superior to the STRF and RF models (STRF: $R^2 = 0.83$, RMSE = 11.92 μg/m^3, and rRMSE = 30.43%; RF: $R^2 = 0.81$, RMSE = 12.62 μg/m^3, and rRMSE = 32.19%). For site-based CV, the SDSTRF model achieves a significant improvement over the STRF and RF models with an R^2 of 0.87, an RMSE of 10.68 μg/m^3, and an rRMSE of 27.48%. In general, SDSTRF and STRF are significantly superior to RF due to taking space-time information into account. In addition, the site-based CV approach achieves better performance than other approaches indicating that this cross-validation strategy can tackle spatial heterogeneity well. The overall statistical indicators of the SDSTRF model are relatively better than the STRF for excluding outlier disturbances in spatial and temporal variations.

Table 2
Collinearity analysis among all of selected variables.

Variable	AOD	SP	TMP	WS	WD
VIF	1.256	**8.984**	1.815	1.306	1.132
Variable	RH	BLH	NDVI	DEM	-
VIF	1.693	1.936	1.594	**9.011**	-

4.1.2. Validation at individual stations

The national scale reflects the overall estimation performance of the model, while the individual-station scale illustrates its local performance. In light of this, the SDSTRF model is also validated at individual stations (Fig. 4). Those stations from the site-based CV with no more than 10 samples were ignored in order to show statistical significance. As displayed in Fig. 4, the daily $PM_{2.5}$ estimates are consistent with observations at most monitoring stations across China. There is an average R^2 of 0.78, and 72% of the monitoring stations show high estimation accuracy with $R^2 > 0.7$, particularly those in central and northern China. Regarding the uncertainty indicators, the mean values are 8.97 μg/m^3 for RMSE, 24.53% for rRMSE and 0.19 μg/m^3 for MBE, respectively. ~88% of the monitoring stations have an average of RMSE, an average of rRMSE, and an absolute of MBE less than 15 μg/m^3, 40% and 5 μg/m^3, respectively, especially those in central and southern China. However, in the northwest region of China, lower prediction accuracies and larger estimation errors are observed. These issues can likely be attributed to the sparse distribution of monitoring stations in that area, which affects the model's ability to capture local variations accurately. Additionally, the proposed model tends to slightly underestimate high $PM_{2.5}$ concentrations and slightly overestimate low concentrations. There are three potential ways to address this issue. Firstly, it is essential to intensify the collection of in-situ measurements in this region. Secondly, mechanistic models can provide valuable insights to achieve a deeper understanding of the factors that influence $PM_{2.5}$ concentrations in this region. Lastly, the development of a more advanced model with excellent generalization ability could make a significant contribution to solving this problem and providing reliable $PM_{2.5}$ estimates. In conclusion, considering the overall estimation results, the SDSTRF model provides accurate estimates of daily $PM_{2.5}$ concentrations at the majority of monitoring stations.

4.2. Validation at different time scales

4.2.1. Validation at the daily scale

The adaptability of the SDSTRF model at the daily scale is also evaluated. Fig. 5 shows the model performance as a function of the DOY from available stations in China. With regard to the statistical significance, days with less than 10 samples were ignored. As shown in Fig. 5, the SDSTRF model performs well on most days with an average R^2 of 0.74, and on approximately 64% of these days, the R^2 value exceeds 0.7. Over the course of the year, the model prediction error indicators RMSE

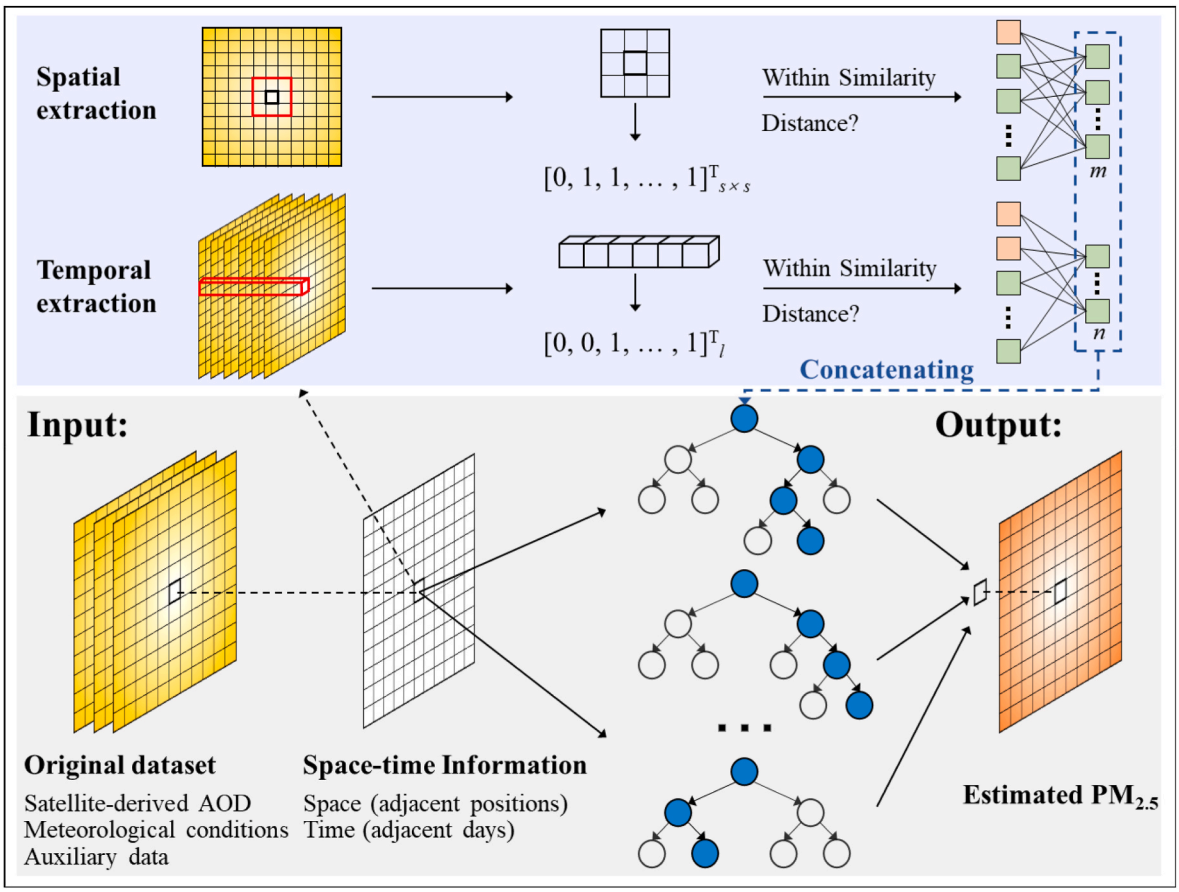


Fig. 2. Schematic of the similarity distance-based space-time random forest (SDSTRF) model for $PM_{2.5}$ concentration estimation.

and MBE exhibit similar time series patterns. The RMSE initially decreases until around day 280 and then gradually increases, while the MBE fluctuates within a small range, then shows large deviations after day 280. A mean value of RMSE and an absolute value of MBE less than 15 and 5 μg/m^3 are observed on approximately 88% and 91% of these days, respectively. On the other hand, rRMSE remains relatively stable at an average level of 24.45%, which indicates that the SDSTRF model maintains a consistent level of performance across different periods of the year. In addition, it is generally observed that R^2 is high at the beginning and end of the year as well as overall large RMSE and MBE values. The reason for this is that $PM_{2.5}$ concentrations are always higher than normal due to increased pollutant emissions from human activities. By way of contrast, the middle of the year is less polluted, which results in a lower R^2 as well as overall smaller RMSE and MBE values. It is evident from these results that, on most days of the year, the SDSTRF model accurately estimates $PM_{2.5}$ concentrations.

4.2.2. Validation at the seasonal and synthetic time scales

An evaluation of the SDSTRF model is also performed on a seasonal basis to examine its performance over a longer period of time. Fig. 6 shows the SDSTRF model's performance on the site-based CV dataset for different seasons during 2018–2022. The small sample size is apparent in summer and winter, due to heavy cloud cover in summer and high frequency of snow and ice in winter. The SDSTRF model provides high estimation accuracy in spring, summer, autumn, and winter, which yields R^2 values of 0.80, 0.76, 0.84, and 0.85, as well as RMSE values of 12.85, 7.43, 9.18, and 14.15 μg/m^3, respectively. Among the four seasons, winter shows the best performance of the model, with the highest R^2 (0.85), relatively good rRMSE (26.01%), and the best fitting line (slope = 0.78, intercept = 12.32 μg/m^3). During summer, the model performs poorly with the lowest R^2 (0.76), relatively poor rRMSE (29.78%) and slope (0.68). However, summer has the smallest RMSE (7.43 μg/m^3) due to the least air pollution. The RMSE values in spring and winter are approximately twice that of summer for the reason that winter experiences the most severe pollutant emissions, followed by spring, resulting from natural conditions and human activities. In brief, the newly developed SDSTRF model can handle the seasonal variation of $PM_{2.5}$ well, despite the differences in model performance.

The validation against ground measurements is also conducted on the synthetic time scales of month, season, and year. For the monthly scale (Fig. 7a), the data for the evaluation are averaged from valid $PM_{2.5}$ days at individual monitoring stations. The SDSTRF model has good accuracy at the monthly scale with an R^2 of 0.85, an RMSE of 8.26 μg/m^3, and an rRMSE of 21.62%. Seasonal data for the estimation are averaged from at least 2 months at each station (Fig. 7b), and it has an equal R^2 (0.85) and a decreased RMSE (7.35 μg/m^3). The annual data for the estimation are averaged from a whole year at each station (Fig. 7c). The estimation results at the annual scale are in reasonable agreement with the ground observations, with an R^2 of 0.81, an RMSE of 5.23 μg/m^3, and an rRMSE of 14.18%. On the basis of these results, the SDSTRF model is accurate at capturing the long-term variation of $PM_{2.5}$ across China.

In summary, the SDSTRF model can estimate $PM_{2.5}$ concentrations well on diverse time scales. It is extremely valuable for analyzing $PM_{2.5}$ concentrations over time across China. Moreover, it can provide reliable validation support for locations and time without ground observations.

4.3. Mapping of the ground-level $PM_{2.5}$ concentrations over China

4.3.1. Seasonal mapping

Based on the developed SDSTRF model, daily $PM_{2.5}$ concentrations are predicted and averaged over the season. Fig. 8 exhibits the $PM_{2.5}$

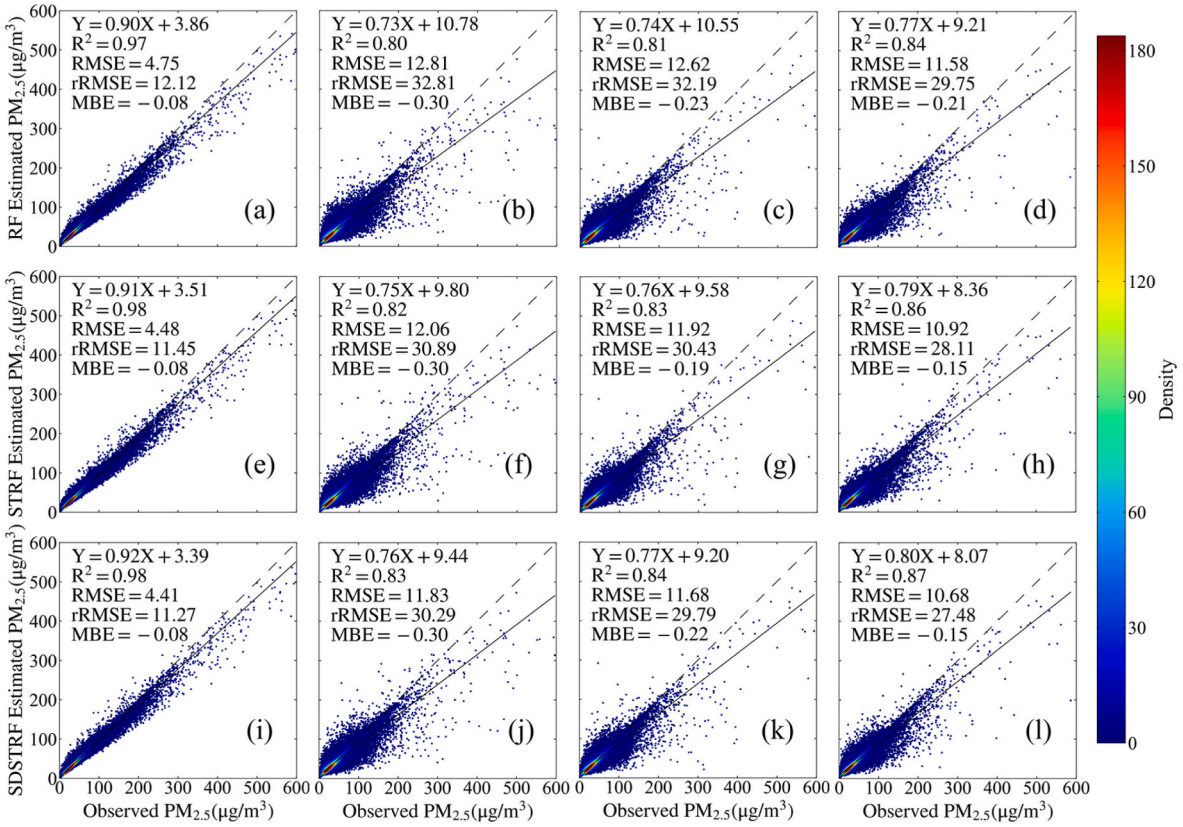


Fig. 3. Training (N = 366,204) and sample-based (N = 127,926), time-based (N = 91,552), and site-based (N = 101,367) cross-validation results for the original RF (a–d), STRF (e–h), and SDSTRF (i–l) models at the national scale.

distribution in different seasons across China in 2020. Spatially, the results cover most regions, but missing values can be observed in southwest China during the whole year and in northeast China during winter. Among the four seasons in 2020, $PM_{2.5}$ pollution is lowest in summer with an average of 31.35 μg/m^3, within a range of 20.72 μg/m^3 to 143.13 μg/m^3. In contrast, winter experiences the highest concentration with an average of 43.45 μg/m^3, in a span of 22.07 μg/m^3 to 229.05 μg/m^3, especially in central and eastern China. For example, cities such as Chengdu, Chongqing, and Wuhan are representative megacities located in midwestern China that suffer from severe $PM_{2.5}$ pollution, mainly due to the close topography, unfavourable dispersion conditions, and high emission intensity of air pollution in winter (Liao et al., 2018). Moreover, these cities are regions characterized by high population densities, a large number of motor vehicles, and numerous industries activities, leading to significant emissions of air pollutants. In addition, it is worth mentioning that Xinjiang Province also suffers from extreme $PM_{2.5}$ pollution in the spring, primarily due to frequent sandstorms.

4.3.2. Annual variation of the estimated $PM_{2.5}$ concentrations over regional hotspots

Annual mean $PM_{2.5}$ concentrations are obtained using a similar method as in the previous section. It shows a variation of decline on the overall level from 2018 to 2022, i.e., 46.91, 46.74, 44.90, 43.60, and 43.31 μg/m^3, for the implementation of pollution control policies and potential contribution of COVID-19 (Yin et al., 2021). The predicted annual mean concentrations are in excess of WHO acceptable levels (i.e., 35 μg/m^3). This suggests that the control of $PM_{2.5}$ pollution in China still cannot be taken lightly.

The North China Plain (NCP), Yangtze River Delta (YRD), Pearl River Delta (PRD), and Sichuan Basin (SCB) regions, which suffer from severe $PM_{2.5}$ pollution, are selected to conduct the regional variation analysis. Fig. 9 shows the estimated $PM_{2.5}$ distribution for these hotspots during 2018–2022. Fig. 9(a–e) show that $PM_{2.5}$ pollution in the NCP region is decreasing from year to year. The highest concentrations are found in southeast Beijing, all of Tianjin, south Hebei, and northeast Shandong. The annual mean concentrations in this region are 42.98, 40.83, 39.83, 38.27, and 37.54 μg/m^3, respectively. Fig. 9(f-j) and Fig. 9(k-o) show that $PM_{2.5}$ pollution is decreasing at the overall level for the YRD and PRD regions. For the YRD region, the highest concentrations are found in northwest and southeast Jiangsu, north Anhui, and all of Shanghai. For the PRD region, the highest concentrations are found in southern Guangdong. The detailed annual variations of these two hotpots are listed in Table 3. As shown in Fig. 9(p-t), $PM_{2.5}$ pollution in the SCB region is overall increasing from year to year. The highest concentrations of $PM_{2.5}$ are found in central Sichuan and southwest Chongqing. The annual average concentrations in this region are 35.90, 36.59, 37.24, 37.33, and 36.74 μg/m^3, respectively. According to these results, it appears that the $PM_{2.5}$ product derived from the proposed SDSTRF model is valuable for air quality studies in areas of high public concern.

5. Discussion

5.1. Predictive power of the SDSTRF model

In this study, the performance of the newly developed SDSTRF model on estimating $PM_{2.5}$ concentrations is evaluated by integrating station measurements, AOD products, meteorological data and auxiliary information. Compared with traditional $PM_{2.5}$ estimation models, the SDSTRF model uses a local fitting strategy to capture the $PM_{2.5}$ variations in space and time, and it achieves an excellent performance on estimation accuracy. Based on a variety of validation approaches, the model has been found to be robust and stable by incorporating similarity distance into the STRF model, which eliminates the outlier disturbances

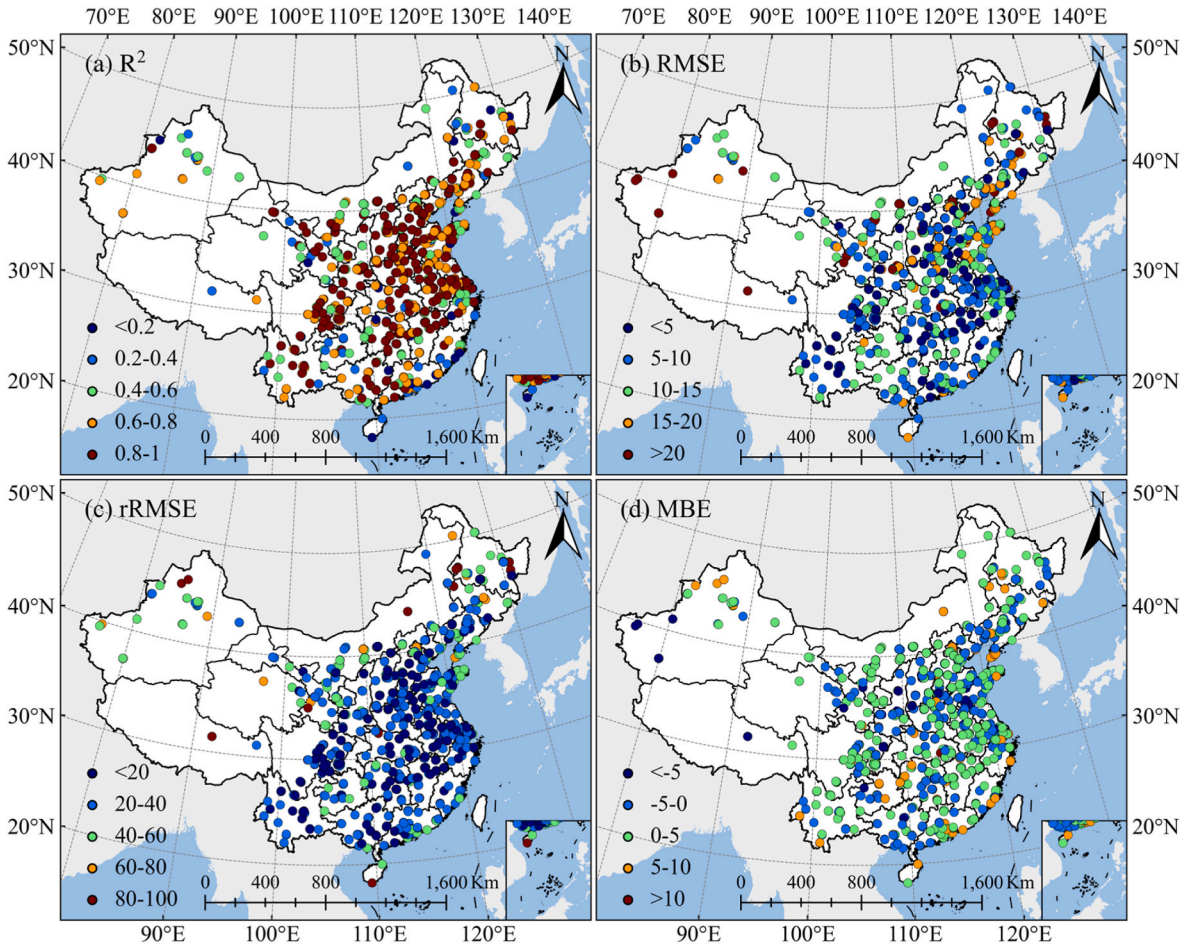


Fig. 4. Spatial distributions of the SDSTRF model performance at individual stations for the site-based cross-validation (a) R^2, (b) RMSE, (c) rRMSE, and (d) MBE during 2018–2022.

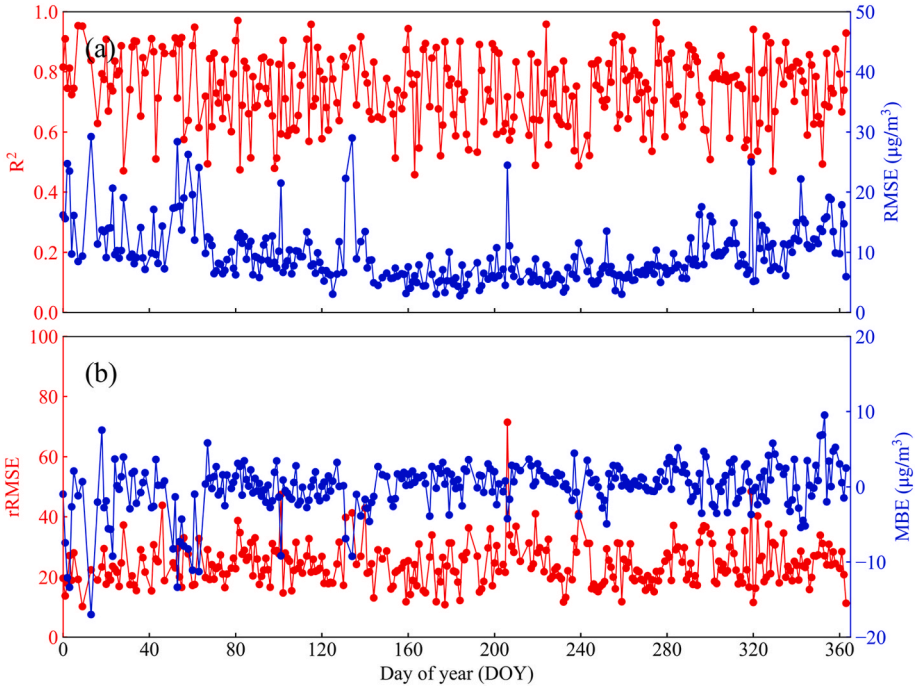


Fig. 5. Time series of the daily performance from the SDSTRF model for (a) R^2 (red) and RMSE (blue) and (b) rRMSE (red) and MBE (blue) at validation stations in 2019 across China.

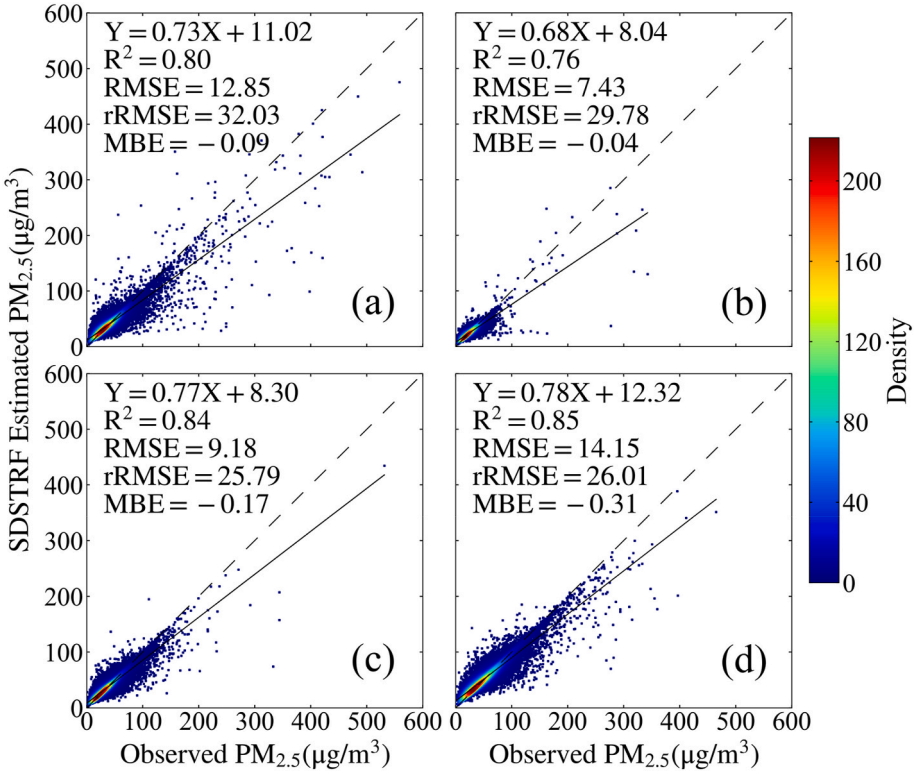


Fig. 6. Site-based cross-validation results from the SDSTRF model for (a) spring (N = 27,195), (b) summer (N = 16,346), (c) autumn (N = 28,016), and (d) winter (N = 20,042) in China during 2018–2022.

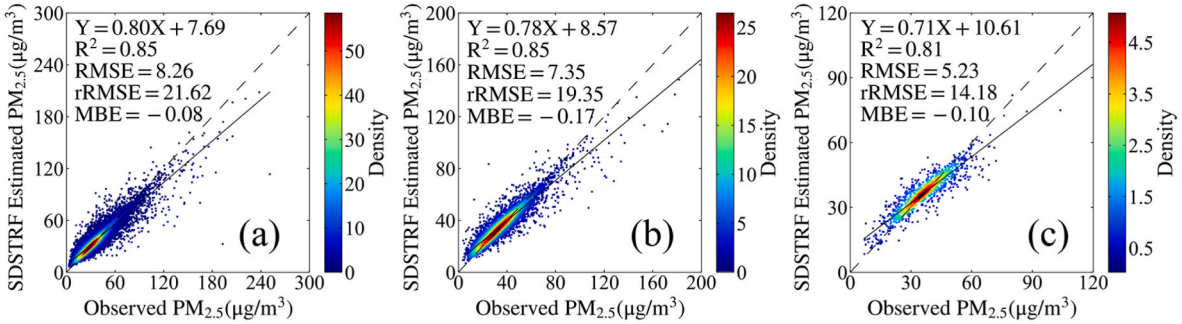


Fig. 7. Site-based cross-validation results from the SDSTRF model at (a) monthly (N = 11,712), (b) seasonal (N = 4143), and (c) annual (N = 1131) scales during 2018–2022.

caused by the resampling and matching of data. The indicator of similarity distance is designed to eliminate errors and retain the most useful adjacent information for the central cell. If the SDSTRF model can predict the untrained data accurately as well, it will be a highly prospective model for $PM_{2.5}$ estimates.

To determine if the SDSTRF model is accurate in its predictive power, the samples selected from the test sites are input into the best models from the site-based CV approaches. Fig. 10 shows the test results of three models derived from the site-based CV approach. The overall prediction accuracy of the three different models decreases using the test samples. The SDSTRF model performs best with a 0.80 R^2, a 12.89 μg/m^3 RMSE, and a 33.01% rRMSE compared with the STRF (R^2 = 0.79, RMSE = 13.15 μg/m^3, and rRMSE = 33.70%) and RF (R^2 = 0.71, RMSE = 15.50 μg/m^3, and rRMSE = 39.91%) models. This indicates that the SDSTRF model is robust under the influence of spatial-temporal heterogeneity. However, the SDSTRF model underpredicts the higher $PM_{2.5}$ concentrations slightly with an MBE less than 0 (i.e., −0.92 μg/m^3). Previous studies have commonly reported this problem (Li et al., 2020b, 2021; Wei et al., 2020). The chief reason for this is that there are only 1% of samples from those days when the daily $PM_{2.5}$ level is extremely high (e.g., $PM_{2.5}$ > 150 μg/m^3). Our model, however, provides better predictions for high $PM_{2.5}$ levels with a steeper 0.74 slope and a lower 10.79 μg/m^3 intercept.

Historical validation is also conducted to assess the predictive power of the SDSTRF model by using the second half-year data of 2022 as a distinct test set (i.e., July to December). The validation results suggest that the SDSTRF model accurately captures more than 59% of historical daily $PM_{2.5}$ concentrations, and exhibits overall minimal estimation uncertainties (RMSE = 14.70 μg/m^3, rRMSE = 41.07%, and MBE = −4.81 μg/m^3). Importantly, the SDSTRF model still outperforms the STRF and RF models (refer to Fig. S4a and Fig. S4b). This robust predictive power indicates that the SDSTRF model is adept at estimating historical $PM_{2.5}$ concentrations across mainland China.

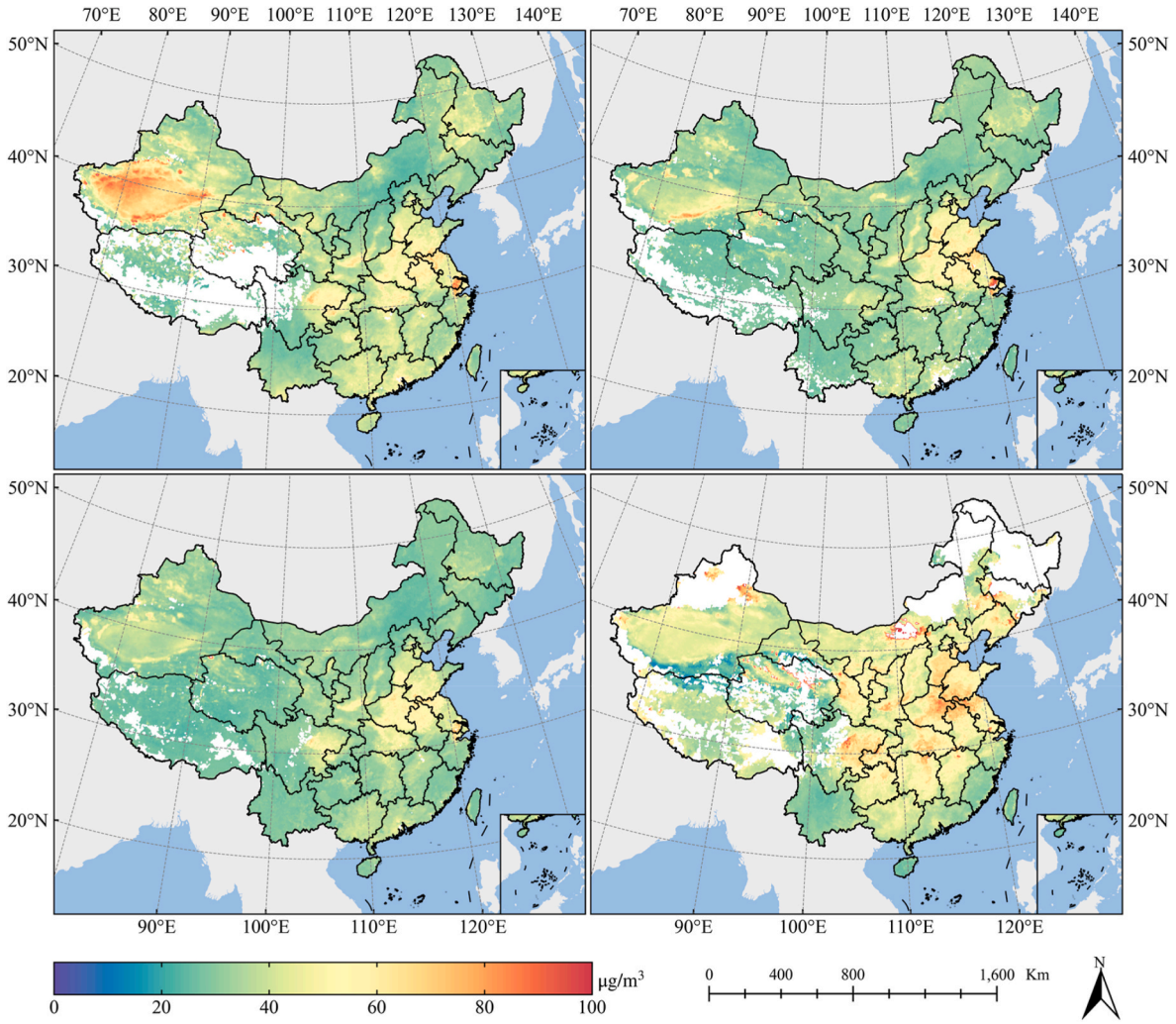


Fig. 8. Seasonal mean distribution of the estimated $PM_{2.5}$ concentrations for (a) spring, (b) summer, (c) autumn, and (d) winter in 2020 across China.

5.2. Influences of the space-time information on model accuracy

In addition to introducing similarity distance to improve the estimation accuracy, another SDSTRF's superiority is that it fully considers the influences of independent variables on AOD-$PM_{2.5}$. The appropriate meteorological variables are indispensable for $PM_{2.5}$ estimates (Chen et al., 2020), and actually so are spatial and temporal variables. An external test is performed to investigate the extent to which spatial-temporal variations can explain the $PM_{2.5}$ concentrations. Based on the SDSTRF model, Fig. 11 exhibits the test results with different inputs of spatial-temporal variations. Compared to incorporating both spatial-temporal variations into the model (Fig. 10c), the R^2, RMSE, and rRMSE values for the model without space-time information decrease dramatically to 0.71, 15.50 μg/m^3, and 39.91%, respectively. When only spatial (temporal) information is considered in the model, they decrease slightly to 0.77, 13.83 μg/m^3, and 35.38% (0.78, 13.55 μg/m^3, and 34.77%). The spatial and temporal variations are essential for estimating $PM_{2.5}$ concentrations, and temporal variation accounts for more $PM_{2.5}$ concentrations than spatial variation.

5.3. Comparison with recent studies

Previous studies have tried to further extend the predictive power of $PM_{2.5}$ models. With the development of these models, we can obtain the space-time continuous distribution of $PM_{2.5}$ to mitigate its impact on air quality. An analysis of this study in comparison with recent studies on China is presented in Table 4, i.e., GWR, GTWR, Two-stage, Geoi-DBN, and STRF models. With respect to these models, the SDSTRF model using the site-based CV approach captures 87% of the daily variation, which outperforms most of the previous models, such as GWR (R^2 = 0.64), GTWR (R^2 = 0.80), the two-stage (stage-1: R^2 = 0.78; stage-2: R^2 = 0.79), and STRF (R^2 = 0.85). In addition, the SDSTRF model is examined for its predictive power on leave-out data from both temporal and spatial perspectives, with R^2 of 0.59 and 0.80, respectively. The comparison results confirm the SDSTRF model's ability to accurately estimate and predict $PM_{2.5}$ concentrations across China.

6. Summary and conclusions

$PM_{2.5}$ has long-lasting detrimental effects on the environment and human health. This situation can be improved by obtaining $PM_{2.5}$ variations using remote sensing. The similarity distance-based space-time random forest (SDSTRF) model is therefore proposed to provide $PM_{2.5}$ estimates with high accuracy over China. By taking into account the nonlinearity and spatial-temporal variations in AOD-$PM_{2.5}$, the SDSTRF model achieves satisfactory performance with the input of independent variables. For model validation, the sample-, time-, and site-based 10-CV approaches are adopted, with R^2 values of 0.83, 0.84, and 0.87, respectively. Compared with previous space-time models, such as GTWR and STRF, the SDSTRF model has a significant advantage in estimation accuracy. Its predictive power is also evaluated by predicting the leave-out data for the historical data and test sites (historical data: R^2 = 0.59,

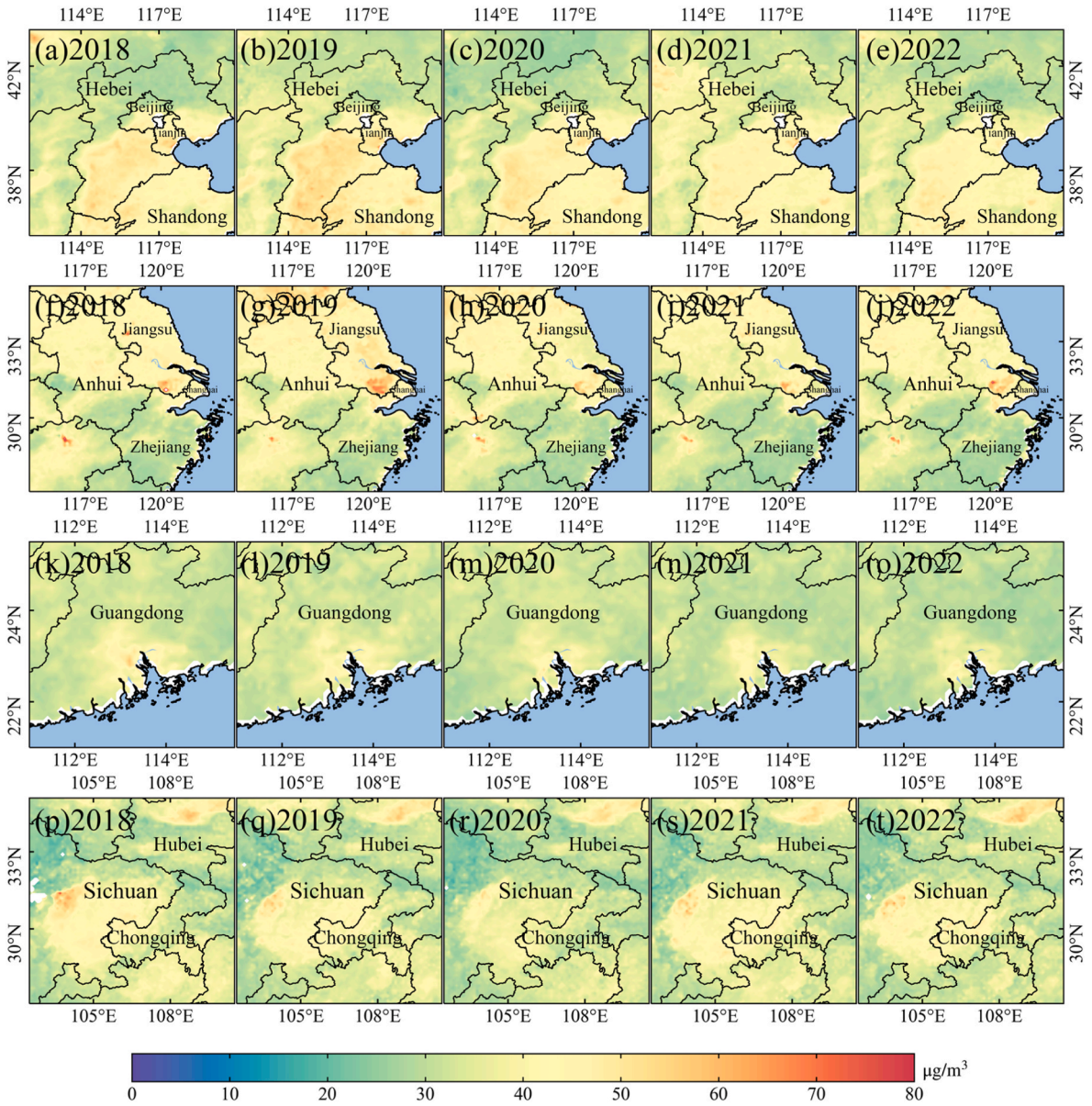


Fig. 9. Annual mean distribution of the estimated $PM_{2.5}$ concentrations for the North China Plain (a–e), Yangtze River Delta (f–j), Pearl River Delta (k–o), and Sichuan Basin (p–t) regions from 2018 to 2022.

Table 3
The annual variation of mean $PM_{2.5}$ concentrations (μg/m^3) in China, the North China Plain (NCP), the Yangtze River Delta (YRD), the Pearl River Delta (PRD), and the Sichuan Basin (SCB) during 2018–2022.

region	2018	2019	2020	2021	2022
China	46.91	46.74	44.90	43.60	43.31
NCP	42.98	40.83	39.83	38.27	37.54
YRD	37.43	39.11	35.54	34.35	35.58
PRD	36.89	38.13	34.98	33.90	34.74
SCB	35.90	36.59	37.24	37.33	36.74

RMSE = 14.70 μg/m^3, and rRMSE = 41.07%; test sites: R^2 = 0.80, RMSE = 12.89 μg/m^3, and rRMSE = 33.01%). Meanwhile, to investigate the extent to which spatial-temporal variations affect the estimation accuracy, an external test is performed by inputting spatial-only or temporal-only variation into the model (spatial-only: R^2 = 0.77, RMSE = 13.83 μg/m^3, and rRMSE = 35.38%; temporal-only: R^2 = 0.78, RMSE = 13.53 μg/m^3, and rRMSE = 34.77%). It is shown that spatial-temporal heterogeneity, especially temporal heterogeneity, strongly affects the $PM_{2.5}$ estimation.

The SDSTRF model is capable of accurately predicting $PM_{2.5}$ concentrations at different time scales. In accordance with this, the $PM_{2.5}$ distributions across China are mapped by season and year. As a final point, the SDSTRF model can generate an accurate and long-term $PM_{2.5}$ product, which will be useful for air quality studies, especially in a large area of interest.

Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by the National Natural Science Foundation of China Major Program [grant number 42192584]; National Natural

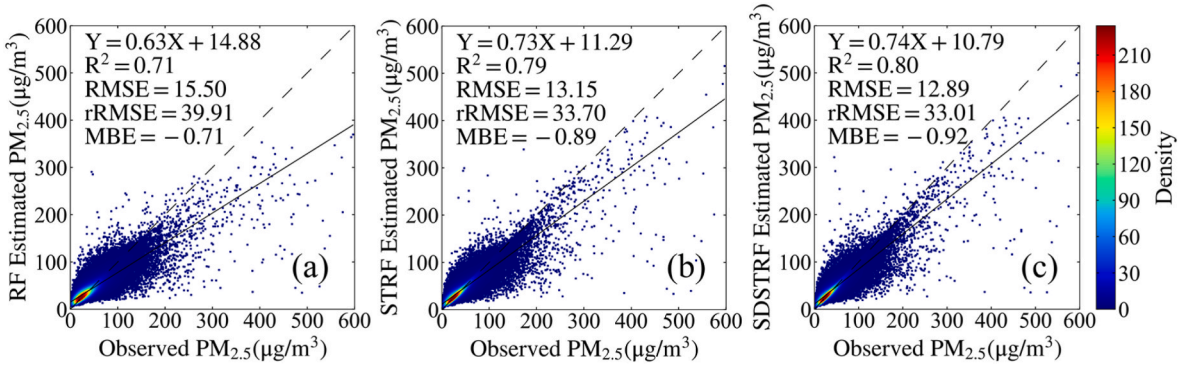


Fig. 10. The results based on the site-based cross-validation approach for assessing the predictive power of (a) the original RF, (b) STRF, and (c) SDSTRF models (N = 181,875) on test sites.

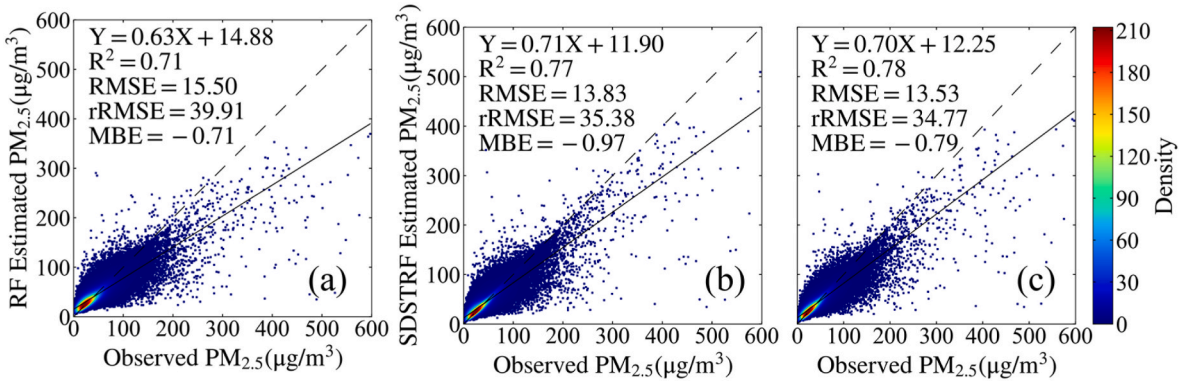


Fig. 11. Test results based on the site-based cross-validation approach for different space-time information inputs: (a) the none of spatial and temporal variation, (b) only spatial variation, and (c) only temporal variation inputs.

Table 4
Statistical indicators to compare the performance of different space-time models in China.

Related study	Model	Spatial resolution	Temporal span	Training		Validation		Prediction	
				R^2	RMSE	R^2	RMSE	R^2	RMSE
Ma et al. (2014)	GWR	10 km	2012–2013	0.71	29.58	0.64	32.98	-	-
He and Huang (2018)	GTWR	3 km	2015	0.85	15.28	0.80	18.00	0.47	37.57
Ma et al. (2016)	Stage-1	10 km	2004–2013	0.81	26.31	0.78	27.99	-	-
	Stage-2			0.82	25.74	0.79	27.42	0.41	-
Li et al. (2017a)	Geoi-DBN	10 km	2015	0.88	13.05	0.88	13.03	-	-
Wei et al. (2019)	STRF	1 km	2015–2016	0.98	5.57	0.85	15.57	0.55	27.38
this study	SDSTRF (Historical validation)	10 km	2018–2022	0.98	4.61	0.85	11.46	0.59	14.70
	SDSTRF (Site-based CV)			0.98	4.41	0.87	10.68	0.80	12.89

Science Foundation of China [grant number 42171320].

CRediT authorship contribution statement

Shikang Guan: Methodology, Validation, Data curation, Writing – original draft, Writing – review & editing. **Xiaotong Zhang:** Conceptualization, Writing – review & editing, Supervision. **Wenbo Zhao:** Writing – review & editing. **Yanjun Duan:** Writing – review & editing. **Shuyue Yang:** Writing – review & editing. **Yunjun Yao:** Writing – review & editing. **Kun Jia:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors sincerely thank the institutions and researchers who provided the data used in this study and made them available to the public. The authors also sincerely thank the anonymous reviewers and editors for their valuable suggestions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2023.120043.

References

Al-Saadi, J., Szykman, J., Pierce, R.B., Kittaka, C., Neil, D., Chu, D.A., Remer, L., Gumley, L., Prins, E., Weinstock, L., MacDonald, C., Wayland, R., Dimmick, F., Fishman, J., 2005. Improving national air quality forecasts with satellite aerosol observations. Bull. Am. Meteorol. Soc. 86, 1249–1262. https://doi.org/10.1175/BAMS-86-9-1249.

Baccarelli, A., Martinelli, I., Zanobetti, A., Grillo, P., Hou, L.-F., Bertazzi, P.A., Mannucci, P.M., Schwartz, J., 2008. Exposure to particulate air pollution and risk of deep vein thrombosis. Arch. Intern. Med. 168, 920–927. https://doi.org/10.1001/archinte.168.9.920.

Boyouk, N., Léon, J.-F., Delbarre, H., Podvin, T., Deroo, C., 2010. Impact of the mixing boundary layer on the relationship between PM2.5 and aerosol optical thickness. Atmos. Environ. 44, 271–277. https://doi.org/10.1016/j.atmosenv.2009.06.053.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brokamp, C., Jandarov, R., Hossain, M., Ryan, P., 2018. Predicting daily urban fine particulate matter concentrations using a random forest model. Environ. Sci. Technol. 52, 4173–4179. https://doi.org/10.1021/acs.est.7b05381.

Chen, J., Li, K., Tang, Z., Bilal, K., Yu, S., Weng, C., Li, K., 2017. A parallel random forest algorithm for big data in a spark cloud computing environment. IEEE Trans. Parallel Distr. Syst. 28, 919–933. https://doi.org/10.1109/TPDS.2016.2603511.

Chen, Z.-Y., Zhang, T.-H., Zhang, R., Zhu, Z.-M., Yang, J., Chen, P.-Y., Ou, C.-Q., Guo, Y., 2019. Extreme gradient boosting model to estimate PM2.5 concentrations with missing-filled satellite data in China. Atmos. Environ. 202, 180–189. https://doi.org/10.1016/j.atmosenv.2019.01.027.

Chen, Z., Chen, D., Zhao, C., Kwan, M.-p., Cai, J., Zhuang, Y., Zhao, B., Wang, X., Chen, B., Yang, J., Li, R., He, B., Gao, B., Wang, K., Xu, B., 2020. Influence of meteorological conditions on PM2.5 concentrations across China: a review of methodology and mechanism. Environ. Int. 139 https://doi.org/10.1016/j.envint.2020.105558.

Cho, Y., Kim, J., Lee, J., Choi, M., Lim, H., Lee, S., Im, J., 2023. Fine particulate concentrations over East Asia derived from aerosols measured by the advanced Himawari Imager using machine learning. Atmos. Res. 290 https://doi.org/10.1016/j.atmosres.2023.106787.

Choi, J., Fuentes, M., Reich, B.J., 2009. Spatial-temporal association between fine particulate matter and daily mortality. Comput. Stat. Data Anal. 53, 2989–3000. https://doi.org/10.1016/j.csda.2008.05.018.

Chu, D.A., Kaufman, Y.J., Zibordi, G., Chern, J.D., Mao, J., Li, C., Holben, B.N., 2003. Global monitoring of air pollution over land from the earth observing system-terra moderate resolution imaging spectroradiometer (MODIS). J. Geophys. Res. Atmos. 108 https://doi.org/10.1029/2002jd003179.

Chu, D.A., Tsai, T.-C., Chen, J.-P., Chang, S.-C., Jeng, Y.-J., Chiang, W.-L., Lin, N.-H., 2013. Interpreting aerosol lidar profiles to better estimate surface PM2.5 for columnar AOD measurements. Atmos. Environ. 79, 172–187. https://doi.org/10.1016/j.atmosenv.2013.06.031.

Chu, H.-J., Huang, B., Lin, C.-Y., 2015. Modeling the spatio-temporal heterogeneity in the PM10-PM2.5 relationship. Atmos. Environ. 102, 176–182. https://doi.org/10.1016/j.atmosenv.2014.11.062.

Dong, Y., Wang, H., Zhang, L., Zhang, K., 2016. An Improved Model for PM2.5 Inference Based on Support Vector Machine, 2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pp. 27–31.

Drury, E., Jacob, D.J., Spurr, R.J.D., Wang, J., Shinozuka, Y., Anderson, B.E., Clarke, A. D., Dibb, J., McNaughton, C., Weber, R., 2010. Synthesis of satellite (MODIS), aircraft (ICARTT), and surface (IMPROVE, EPA-AQS, AERONET) aerosol observations over eastern North America to improve MODIS aerosol retrievals and constrain surface aerosol concentrations and sources. J. Geophys. Res. Atmos. 115 https://doi.org/10.1029/2009JD012629.

Emili, E., Popp, C., Petitta, M., Riffler, M., Wunderle, S., Zebisch, M., 2010. PM10 remote sensing from geostationary SEVIRI and polar-orbiting MODIS sensors over the complex terrain of the European Alpine region. Rem. Sens. Environ. 114, 2485–2499. https://doi.org/10.1016/j.rse.2010.05.024.

Geng, G., Xiao, Q., Liu, S., Liu, X., Cheng, J., Zheng, Y., Xue, T., Tong, D., Zheng, B., Peng, Y., Huang, X., He, K., Zhang, Q., 2021. Tracking air pollution in China: near real-time $PM_{2.5}$ retrievals from multisource data fusion. Environ. Sci. Technol. 55, 12106–12115. https://doi.org/10.1021/acs.est.1c01863.

Geng, G., Zhang, Q., Martin, R.V., van Donkelaar, A., Huo, H., Che, H., Lin, J., He, K., 2015. Estimating long-term $PM_{2.5}$ concentrations in China using satellite-based aerosol optical depth and a chemical transport model. Rem. Sens. Environ. 166, 262–270. https://doi.org/10.1016/j.rse.2015.05.016.

Gongbo, C., Shanshan, L., Luke, D.K., Nicholas, A.S.H., Wei, C., Tiantian, L., Jianping, G., Hongyan, R., Michael, J.A., Yuming, G., 2018. A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information. Sci. Total Environ. 636, 52–60. https://doi.org/10.1016/J.SCITOTENV.2018.04.251.

Guan, W.-J., Zheng, X.-Y., Chung, K.F., Zhong, N.-S., 2016. Impact of air pollution on the burden of chronic respiratory diseases in China: time for urgent action. Lancet 388, 1939–1951. https://doi.org/10.1016/S0140-6736(16)31597-5.

Gui, K., Che, H., Zeng, Z., Wang, Y., Zhai, S., Wang, Z., Luo, M., Zhang, L., Liao, T., Zhao, H., Li, L., Zheng, Y., Zhang, X., 2020. Construction of a virtual $PM_{2.5}$ observation network in China based on high-density surface meteorological observations using the Extreme Gradient Boosting model. Environ. Int. 141, 105801 https://doi.org/10.1016/j.envint.2020.105801.

Guo, B., Zhang, D., Pei, L., Su, Y., Wang, X., Bian, Y., Zhang, D., Yao, W., Zhou, Z., Guo, L., 2021a. Estimating $PM_{2.5}$ concentrations via random forest method using satellite, auxiliary, and ground-level station dataset at multiple temporal scales across China in 2017. Sci. Total Environ. 778, 146288 https://doi.org/10.1016/j.scitotenv.2021.146288.

Guo, J., Zhang, J., Yang, K., Liao, H., Zhang, S., Huang, K., Lv, Y., Shao, J., Yu, T., Tong, B., Li, J., Su, T., Yim, S.H.L., Stoffelen, A., Zhai, P., Xu, X., 2021b. Investigation of near-global daytime boundary layer height using high-resolution radiosondes: first results and comparison with ERA5, MERRA-2, JRA-55, and NCEP-2 reanalyses. Atmos. Chem. Phys. 21, 17079–17097. https://doi.org/10.5194/acp-21-17079-2021.

Gupta, P., Christopher, S.A., 2009. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: multiple regression approach. J. Geophys. Res. 114 https://doi.org/10.1029/2008jd011496.

Haikerwal, A., Akram, M., Del Monaco, A., Smith, K., Sim, M.R., Meyer, M., Tonkin, A. M., Abramson, M.J., Dennekamp, M., 2015. Impact of fine particulate matter ($PM_{2.5}$) exposure during wildfires on cardiovascular health outcomes. J. Am. Heart Assoc. 4 https://doi.org/10.1161/JAHA.114.001653.

Hamra, G.B., Guha, N., Cohen, A., Laden, F., Raaschou-Nielsen, O., Samet, J.M., Vineis, P., Forastiere, F., Saldiva, P., Yorifuji, T., Loomis, D., 2014. Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis. Environ. Health Perspect. 122, 906–911. https://doi.org/10.1289/ehp/1408092.

He, Q., Huang, B., 2018. Satellite-based mapping of daily high-resolution ground $PM_{2.5}$ in China via space-time regression modeling. Rem. Sens. Environ. 206, 72–83. https://doi.org/10.1016/j.rse.2017.12.018.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.-N., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146, 1999–2049. https://doi.org/10.1002/qj.3803.

Hoff, R.M., Christopher, S.A., 2009. Remote sensing of particulate pollution from space: have we reached the promised land? J. Air Waste Manag. Assoc. 59, 645–675. https://doi.org/10.3155/1047-3289.59.6.645.

Hsu, N.C., Jeong, M.J., Bettenhausen, C., Sayer, A.M., Hansell, R., Seftor, C.S., Huang, J., Tsay, S.C., 2013. Enhanced Deep Blue aerosol retrieval algorithm: the second generation. J. Geophys. Res. Atmos. 118, 9296–9315. https://doi.org/10.1002/jgrd.50712.

Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L.A., Strickland, M.J., Liu, Y., 2017. Estimating $PM_{2.5}$ concentrations in the conterminous United States using the random forest approach. Environ. Sci. Technol. 51, 6936–6944. https://doi.org/10.1021/acs.est.7b01210.

Hu, X., Waller, L.A., Al-Hamdan, M.Z., Crosson, W.L., Estes Jr., M.G., Estes, S.M., Quattrochi, D.A., Sarnat, J.A., Liu, Y., 2013. Estimating ground-level $PM_{2.5}$ concentrations in the southeastern U.S. using geographically weighted regression. Environ. Res. 121, 1–10. https://doi.org/10.1016/j.envres.2012.11.003.

Hu, X., Waller, L.A., Lyapustin, A., Wang, Y., Al-Hamdan, M.Z., Crosson, W.L., Estes, M. G., Estes, S.M., Quattrochi, D.A., Puttaswamy, S.J., Liu, Y., 2014. Estimating ground-level $PM_{2.5}$ concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. Rem. Sens. Environ. 140, 220–232. https://doi.org/10.1016/j.rse.2013.08.032.

Hutchison, K.D., Smith, S., Faruqui, S., 2004. The use of MODIS data and aerosol products for air quality prediction. Atmos. Environ. 38, 5057–5070. https://doi.org/10.1016/j.atmosenv.2004.06.032.

Kalnins, A., 2018. Multicollinearity: how common factors cause Type 1 errors in multivariate regression. Strat. Manag. J. 39, 2362–2385. https://doi.org/10.1002/smj.2783.

Kukkonen, J., Olsson, T., Schultz, D.M., Baklanov, A., Klein, T., Miranda, A.I., Monteiro, A., Hirtl, M., Tarvainen, V., Boy, M., Peuch, V.H., Poupkou, A., Kioutsioukis, I., Finardi, S., Sofiev, M., Sokhi, R., Lehtinen, K.E.J., Karatzas, K., San José, R., Astitha, M., Kallos, G., Schaap, M., Reimer, E., Jakobs, H., Eben, K., 2012. A review of operational, regional-scale, chemical weather forecasting models in Europe. Atmos. Chem. Phys. 12, 1–87. https://doi.org/10.5194/acp-12-1-2012.

Kumar, N., 2010. What can affect AOD-PM(2.5) association? Environ. Health Perspect. 118, A 109–110. https://doi.org/10.1289/ehp.0901732 author reply A 110-111.

Kumar, N., Chu, A., Foster, A., 2007. An empirical relationship between $PM_{2.5}$ and aerosol optical depth in Delhi Metropolitan. Atmos. Environ. 41, 4492–4503. https://doi.org/10.1016/j.atmosenv.2007.01.046, 1994.

Lary, D.J., Faruque, F.S., Malakar, N., Moore, A., Roscoe, B., Adams, Z.L., Eggelston, Y., 2014. Estimating the global abundance of ground level presence of particulate matter (PM2.5). Geospatial health 8, S611–S630. https://doi.org/10.4081/gh.2014.292.

Lee, S., Park, S., Lee, M.-I., Kim, G., Im, J., Song, C.-K., 2022. Air quality forecasts improved by combining data assimilation and machine learning with satellite AOD. Geophys. Res. Lett. 49, e2021GL096066 https://doi.org/10.1029/2021GL096066.

Levy, R.C., Mattoo, S., Munchak, L.A., Remer, L.A., Sayer, A.M., Patadia, F., Hsu, N.C., 2013. The Collection 6 MODIS aerosol products over land and ocean. Atmos. Meas. Tech. 6, 2989–3034. https://doi.org/10.5194/amt-6-2989-2013.

Li, S.X., Chen, L.L., Huang, G., Lin, J.T., Yan, Y.Y., Ni, R.J., Huo, Y.F., Wang, J.X., Liu, M. Y., Weng, H.J., Wang, Y.H., Wang, Z.F., 2020a. Retrieval of surface PM2.5 mass concentrations over North China using visibility measurements and GEOS-Chem simulations. Atmos. Environ. 222 https://doi.org/10.1016/j.atmosenv.2019.117121.

Li, T., Shen, H., Yuan, Q., Zhang, L., 2020b. Geographically and temporally weighted neural networks for satellite-based mapping of ground-level PM2.5. ISPRS J.

Photogrammetry Remote Sens. 167, 178–188. https://doi.org/10.1016/j.isprsjprs.2020.06.019.
Li, T., Shen, H., Yuan, Q., Zhang, L., 2021. A locally weighted neural network constrained by global training for remote sensing estimation of $PM_{2.5}$. IEEE Trans. Geosci. Rem. Sens. 60, 1–13. https://doi.org/10.1109/tgrs.2021.3074569.
Li, T., Shen, H., Yuan, Q., Zhang, X., Zhang, L., 2017a. Estimating ground-level PM2.5by fusing satellite and station observations: a geo-intelligent deep learning approach. Geophys. Res. Lett. 44 (11) https://doi.org/10.1002/2017gl075710, 985-911,993.
Li, T., Shen, H., Zeng, C., Yuan, Q., 2020c. A validation approach considering the uneven distribution of ground stations for satellite-based PM2.5 estimation. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 13, 1312–1321. https://doi.org/10.1109/jstars.2020.2977668.
Li, T., Shen, H., Zeng, C., Yuan, Q., Zhang, L., 2017b. Point-surface fusion of station measurements and satellite observations for mapping PM2.5 distribution in China: methods and assessment. Atmos. Environ. 152, 477–489. https://doi.org/10.1016/j.atmosenv.2017.01.004.
Liao, T., Gui, K., Jiang, W., Wang, S., Wang, B., Zeng, Z., Che, H., Wang, Y., Sun, Y., 2018. Air stagnation and its impact on air quality during winter in Sichuan and Chongqing, southwestern China. Sci. Total Environ. 635, 576–585. https://doi.org/10.1016/j.scitotenv.2018.04.122.
Lin, C., Li, Y., Yuan, Z., Lau, A.K.H., Li, C., Fung, J.C.H., 2015. Using satellite remote sensing data to estimate the high-resolution distribution of ground-level PM2.5. Rem. Sens. Environ. 156, 117–128. https://doi.org/10.1016/j.rse.2014.09.015.
Liu, S., Geng, G., Xiao, Q., Zheng, Y., Liu, X., Cheng, J., Zhang, Q., 2022. Tracking daily concentrations of PM2.5 chemical composition in China since 2000. Environ. Sci. Technol. 56, 16517–16527. https://doi.org/10.1021/acs.est.2c06510.
Ma, Z., Hu, X., Huang, L., Bi, J., Liu, Y., 2014. Estimating ground-level PM2.5 in China using satellite remote sensing. Environ. Sci. Technol. 48, 7436–7444. https://doi.org/10.1021/es5009399.
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, 184–192. https://doi.org/10.1289/ehp.1409481.
Merbitz, H., Fritz, S., Schneider, C., 2012. Mobile measurements and regression modeling of the spatial particulate matter variability in an urban area. Sci. Total Environ. 438, 389–403. https://doi.org/10.1016/j.scitotenv.2012.08.049.
Miller, H.J., 2004. Tobler's first law and spatial analysis. Ann. Assoc. Am. Geogr. 94, 284–289. https://doi.org/10.1111/j.1467-8306.2004.09402005.x.
Park, S., Im, J., Kim, J., Kim, S.-M., 2022. Geostationary satellite-derived ground-level particulate matter concentrations using real-time machine learning in Northeast Asia. Environ. Pollut. 306, 119425 https://doi.org/10.1016/j.envpol.2022.119425.
Park, S., Lee, J., Im, J., Song, C.-K., Choi, M., Kim, J., Lee, S., Park, R., Kim, S.-M., Yoon, J., Lee, D.-W., Quackenbush, L.J., 2020. Estimation of spatially continuous daytime particulate matter concentrations under all sky conditions through the synergistic use of satellite-based AOD and numerical models. Sci. Total Environ. 713, 136516 https://doi.org/10.1016/j.scitotenv.2020.136516.
Pope, C.A., Dockery, D.W., 2006. Health effects of fine particulate air pollution: lines that connect. J. Air Waste Manag. Assoc. 56, 709–742. https://doi.org/10.1080/10473289.2006.10464485.
Rodriguez, J.D., Perez, A., Lozano, J.A., 2010. Sensitivity analysis of kappa-fold cross validation in prediction error estimation. IEEE Trans. Pattern Anal. Mach. Intell. 32, 569–575. https://doi.org/10.1109/TPAMI.2009.187.
Rohde, R.A., Muller, R.A., 2015. Air pollution in China: mapping of concentrations and sources. PLoS One 10, e0135749. https://doi.org/10.1371/journal.pone.0135749.
Saide, P.E., Carmichael, G.R., Spak, S.N., Gallardo, L., Osses, A.E., Mena-Carrasco, M.A., Pagowski, M., 2011. Forecasting urban PM10 and PM2.5 pollution episodes in very stable nocturnal conditions and complex terrain using WRF–Chem CO tracer model. Atmos. Environ. 45, 2769–2780. https://doi.org/10.1016/j.atmosenv.2011.02.001.
Sheather, S., 2009. A Modern Approach to Regression with R. Springer, New York.
Tao, J., Zhang, M., Chen, L., Wang, Z., Su, L., Ge, C., Han, X., Zou, M., 2013. A method to estimate concentrations of surface-level particulate matter using satellite-based aerosol optical thickness. Sci. China Earth Sci. 56, 1422–1433. https://doi.org/10.1007/s11430-012-4503-3.
Tian, J., Chen, D., 2010. A semi-empirical model for predicting hourly ground-level fine particulate matter (PM2.5) concentration in southern Ontario from satellite remote sensing and ground-based meteorological measurements. Rem. Sens. Environ. 114, 221–229. https://doi.org/10.1016/j.rse.2009.09.011.
van Donkelaar, A., Martin Randall, V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., Villeneuve Paul, J., 2010. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. Environ. Health Perspect. 118, 847–855. https://doi.org/10.1289/ehp.0901623.
van Donkelaar, A., Martin, R.V., Park, R.J., 2006. Estimating ground-level PM2.5using aerosol optical depth determined from satellite remote sensing. J. Geophys. Res. 111 https://doi.org/10.1029/2005jd006996.
Wang, J., 2003. Intercomparison between satellite-derived aerosol optical thickness and PM2.5mass: implications for air quality studies. Geophys. Res. Lett. 30 https://doi.org/10.1029/2003gl018174.
Wang, Q., Wang, J., Zhou, J., Ban, J., Li, T., 2019. Estimation of PM2.5-associated disease burden in China in 2020 and 2030 using population and air quality scenarios: a modelling study. Lancet Planet. Health 3, e71–e80. https://doi.org/10.1016/s2542-5196(18)30277-8.
Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., Cribb, M., 2019. Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach. Rem. Sens. Environ. 231 https://doi.org/10.1016/j.rse.2019.111221.
Wei, J., Li, Z., Cribb, M., Huang, W., Xue, W., Sun, L., Guo, J., Peng, Y., Li, J., Lyapustin, A., Liu, L., Wu, H., Song, Y., 2020. Improved 1 km resolution PM2.5 estimates across China using enhanced space–time extremely randomized trees. Atmos. Chem. Phys. 20, 3273–3289. https://doi.org/10.5194/acp-20-3273-2020.
Wei, J., Li, Z., Lyapustin, A., Sun, L., Peng, Y., Xue, W., Su, T., Cribb, M., 2021. Reconstructing 1-km-resolution high-quality PM2.5 data records from 2000 to 2018 in China: spatiotemporal variations and policy implications. Rem. Sens. Environ. 252 https://doi.org/10.1016/j.rse.2020.112136.
Weizhen, H., Zhengqiang, L., Yuhuan, Z., Hua, X., Ying, Z., Kaitao, L., Donghui, L., Peng, W., Yan, M., 2014. Using support vector regression to predict PM10 and PM2.5. IOP Conf. Ser. Earth Environ. Sci. 17 https://doi.org/10.1088/1755-1315/17/1/012268.
Xiao, Q., Geng, G., Liu, S., Liu, J., Meng, X., Zhang, Q., 2022. Spatiotemporal continuous estimates of daily 1km PM2.5 from 2000 to present under the Tracking Air Pollution in China (TAP) framework. Atmos. Chem. Phys. 22, 13229–13242. https://doi.org/10.5194/acp-22-13229-2022.
Xue, T., Zheng, Y., Tong, D., Zheng, B., Li, X., Zhu, T., Zhang, Q., 2019. Spatiotemporal continuous estimates of PM2.5 concentrations in China, 2000–2016: a machine learning method with inputs from satellites, chemical transport model, and ground observations. Environ. Int. 123, 345–357. https://doi.org/10.1016/j.envint.2018.11.075.
Yin, Z., Zhang, Y., Wang, H., Li, Y., 2021. Evident PM2.5 drops in the east of China due to the COVID-19 quarantine measures in February. Atmos. Chem. Phys. 21, 1581–1592. https://doi.org/10.5194/acp-21-1581-2021.
Yue, H., He, C., Huang, Q., Yin, D., Bryan, B.A., 2020. Stronger policy required to substantially reduce deaths from PM2.5 pollution in China. Nat. Commun. 11, 1462. https://doi.org/10.1038/s41467-020-15319-4.
Zheng, J., Zhang, L., Che, W., Zheng, Z., Yin, S., 2009. A highly resolved temporal and spatial air pollutant emission inventory for the Pearl River Delta region, China and its uncertainty assessment. Atmos. Environ. 43, 5112–5122. https://doi.org/10.1016/j.atmosenv.2009.04.060.
Zuo, C., Chen, J., Zhang, Y., Jiang, Y., Liu, M., Liu, H., Zhao, W., Yan, X., 2023. Evaluation of four meteorological reanalysis datasets for satellite-based PM2.5 retrieval over China. Atmos. Environ. 305 https://doi.org/10.1016/j.atmosenv.2023.119795.