RESEARCH ARTICLE

National Science Review
8: nwaa307, 2021
https://doi.org/10.1093/nsr/nwaa307
Advance access publication 5 January 2021

EARTH SCIENCES

Robust prediction of hourly $PM_{2.5}$ from meteorological data using LightGBM

Junting Zhong[1,2], Xiaoye Zhang[1,3,*], Ke Gui[1,*], Yaqiang Wang[1], Huizheng Che[1], Xiaojing Shen[1], Lei Zhang[1], Yangmei Zhang[1], Junying Sun[1] and Wenjie Zhang[1]

[1] State Key Laboratory of Severe Weather and Key Laboratory of Atmospheric Chemistry of China Meteorological Administration, Chinese Academy of Meteorological Sciences, Beijing 100081, China; [2] School of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing 100049, China and [3] Center for Excellence in Regional Atmospheric Environment, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China

[*] **Corresponding authors.** E-mails: xiaoye@cma.gov.cn; guik@cma.gov.cn

Received 24 July 2020; **Revised** 9 November 2020; **Accepted** 23 December 2020

ABSTRACT

Retrieving historical fine particulate matter ($PM_{2.5}$) data is key for evaluating the long-term impacts of $PM_{2.5}$ on the environment, human health and climate change. Satellite-based aerosol optical depth has been used to estimate $PM_{2.5}$, but estimations have largely been undermined by massive missing values, low sampling frequency and weak predictive capability. Here, using a novel feature engineering approach to incorporate spatial effects from meteorological data, we developed a robust LightGBM model that predicts $PM_{2.5}$ at an unprecedented predictive capacity on hourly ($R^2 = 0.75$), daily ($R^2 = 0.84$), monthly ($R^2 = 0.88$) and annual ($R^2 = 0.87$) timescales. By taking advantage of spatial features, our model can also construct hourly gridded networks of $PM_{2.5}$. This capability would be further enhanced if meteorological observations from regional stations were incorporated. Our results show that this model has great potential in reconstructing historical $PM_{2.5}$ datasets and real-time gridded networks at high spatial-temporal resolutions. The resulting datasets can be assimilated into models to produce long-term re-analysis that incorporates interactions between aerosols and physical processes.

Keywords: $PM_{2.5}$, spatial features, hourly prediction, high accuracy, gridded networks

INTRODUCTION

Fine particulate matter ($PM_{2.5}$) consists of suspended and inhalable particles that generate environmental and health effects [1–7]. Suspended $PM_{2.5}$ is the primary cause of visibility reduction in China and parts of the United States. When settling on the ground or water, these particles can exert different impacts on ecosystems depending on their chemical composition, including affecting ecological diversity, depleting soil nutrients and acidizing lakes and streams [8,9]. Inhalable $PM_{2.5}$ can penetrate the respiratory system to aggravate respiratory symptoms [10,11] and increase mortality from cardiovascular and respiratory diseases after long-term exposure to $PM_{2.5}$ [5,6,12]. In addition to the profound impacts on the environment and health, interactions between aerosols and radiation also affect climate change directly or indirectly in the long term [1,13–17]. To evaluate the long-term impacts of $PM_{2.5}$ on the atmospheric environment, human health and climate change, it is critical to obtain historical $PM_{2.5}$ datasets at high spatial-temporal resolutions. Nevertheless, the national hourly $PM_{2.5}$ monitoring network from the Ministry of Ecology and Environment was not established until 2013. As a result of the limited $PM_{2.5}$ observations, retrieving historical $PM_{2.5}$ datasets is becoming a research hotspot.

With broad spatial coverage and relatively long observation periods (∼20 years), satellite-retrieved aerosol optical depth (AOD), a measure of the aerosol extinction of the solar beam, has been increasingly used to estimate large-scale $PM_{2.5}$ concentrations for the past two decades [18–24]. For example, Huang *et al.* [20] predicted monthly $PM_{2.5}$ concentrations on the North China Plain (NCP) from Multi-Angle Implementation of Atmospheric Correction (MAIAC) AOD using random forest that improved monthly prediction R^2 (coefficient of determination) to 0.74. Additionally, using the MAIAC AOD, Xiao *et al.* [25] constructed monthly $PM_{2.5}$ datasets from 2000 to 2018 to evaluate their spatial changes. Wei *et al.* [22] estimated daily $PM_{2.5}$

across China using the space-time random forest approach with daily prediction R^2 at 0.55. However, satellite-based AOD has some inherent limitations that are difficult to overcome. For example, a large proportion of non-random missing AOD due to cloud cover significantly affects data availability and generates biases [18,26]. Previous studies estimated that the missing AOD from MODIS (Moderate Resolution Imaging Spectroradiometer) accounted for 70%–90% of the total retrieval [26,27]. Apart from missing data, the sampling frequency of satellite-based AOD is typically limited to a maximum of twice a day. This kind of frequency makes it impossible for hourly $PM_{2.5}$ assessments and leads to the underrepresentation of the average daily AOD. For model accuracy, the predictive accuracy for samples outside the training period is significantly lower than the validation accuracy, indicating that these models' predictive capability is relatively weak. For example, although the R^2 of 10-fold cross-validation (CV) on a daily scale can exceed 0.85, the R^2 of the prediction is no more than 0.58 [21,22].

Compared with satellite-based AOD, horizontal visibility and other variables from surface meteorological observations have unique advantages in retrieving historical $PM_{2.5}$. Surface meteorological observations that can be traced back to the 1950s have much more extended observation periods than satellite-based AOD observations. Surface observations are not disturbed by cloud cover and can continuously record hourly meteorological variables, which overcome the shortcomings of satellite-based AOD data that have massive missing values and low sample frequency. Unlike $PM_{2.5}$ stations that are mainly located in cities, meteorological stations are distributed more evenly and densely. There are 2450 national meteorological stations and over 60 000 regional stations in China. This considerable magnitude has great potential to retrieve historical $PM_{2.5}$ datasets at high spatial-temporal resolutions using visibility-dominated surface meteorological variables. For example, using daily meteorological observations, Gui *et al.* [28] have constructed a virtual ground-based $PM_{2.5}$ network with XGBoost and achieved a better predictive capability with R^2 values of 0.60 and 0.80 on daily and monthly scales, respectively. This work shows that visibility and other meteorological variables are promising for filling gaps in AOD-based $PM_{2.5}$ [28]. Therefore, surface meteorological observations will continue to be used for retrieving $PM_{2.5}$ but on an hourly scale in this study. We will employ a novel feature engineering approach to incorporate spatial effects and build a robust model of which the prediction capacity will improve significantly. The state-of-the-art machine-learning algorithm, LightGBM, will be used in this study to train the model based on over 30 million samples from meteorological observations at 2450 national stations from 2016 to 2018 (Fig. 1). The model performance will be evaluated using 10-fold CV. After validation, we will assess the predictive capability of this model using more than 10 million meteorological samples in 2019. Additionally, we will attempt to construct a densely gridded $PM_{2.5}$ network by taking advantage of extracted spatial features.

RESULTS AND DISCUSSION

Model evaluation from hourly to yearly scales

In contrast to many machine learning models that are black boxes [29], LightGBM models can explain their predictions in a way that humans can understand. The decision-making process of our model was visualized using a feature-importance plot and a digraph representation of a specified tree. Figure S1 shows the relative importance of all the features used to train the model. Visibility from the nearest meteorological station is the most important feature that accounts for ∼7% of the overall importance. Approximately 6% of the overall importance is composed of distance that also significantly affects the model from a spatial perspective. Temporal features and other spatial features are also incorporated into the model, with relative importance ranging from 2% to 5%. Compared with the models in previous studies [20,22,26,28], our model does not heavily rely on one feature, e.g. AOD or visibility, but is able to integrate the influence of different features, particularly the spatial features that fully represent dimensional heterogeneity. For example, without visibility from the nearest station, the R^2 value of observed and predicted $PM_{2.5}$ in 2019 only decreases from 0.75 to 0.72 (Fig. S2). In contrast, the R^2 value decreases more significantly to 0.65 when spatial features of visibility from surrounding stations are excluded (Fig. S2). To further gain an understanding of the decision process, we retrace the process of partitioning the notes on the training dataset by visualizing 1 of 1000 trees randomly in our model (Fig. S3). The tree node was split into child nodes based on the spatial visibility and then further divided based on the spatial relative humidity (RH) or visibility of the nearest station. It is clearly demonstrated how meteorological features, temporal features and spatial features play a role in our model.

Our model's performance was evaluated with 10-fold CV using 31 863 778 hour-by-hour training data across China from 2016 to 2018. As shown

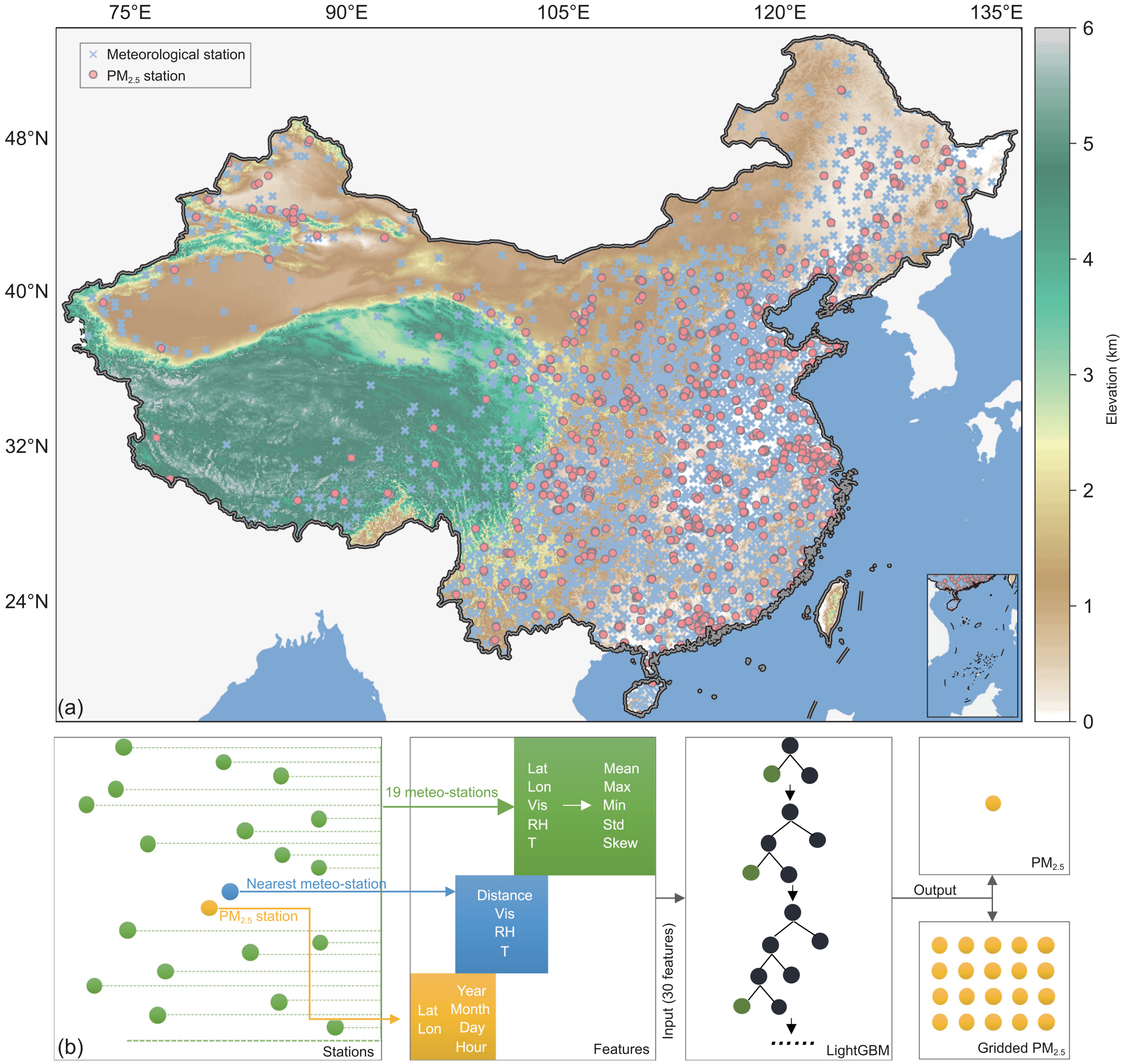

Figure 1. (a) Spatial distribution of 1440 $PM_{2.5}$ stations and 2450 national meteorological stations across China and (b) a conceptual scheme for the extraction of spatial features and the development of our LightGBM model. Review drawing number: GS(2020)6868.

in Fig. 2, the overall R^2 and root-mean-square error (RMSE) for hourly $PM_{2.5}$ estimations are 0.80 and 19.80 $\mu g\,m^{-3}$, respectively, which are in close agreement with the fitting results with the R^2 value of 0.80 and the RMSE value of 19.60 $\mu g\,m^{-3}$ (Fig. S4). This finding indicates that this model can effectively avoid overfitting and achieve high and stable accuracy in estimating hourly $PM_{2.5}$ concentrations. The performance of our model is even better for $PM_{2.5}$ estimations on larger timescales. For daily estimations with 1 454 688 samples, the overall R^2 and RMSE are 0.89 and 12.78 $\mu g\,m^{-3}$, respectively. For monthly and yearly estimations, the overall R^2 values increase to 0.94 and 0.98, respectively, and the RMSE values decrease to 6.78 $\mu g\,m^{-3}$ and 2.16 $\mu g\,m^{-3}$, respectively. To better present our model's performance, we compared our CV scores with those in recent studies that predicted $PM_{2.5}$ across China. As shown in Table 1, our model outperformed all of the other models in the R^2 and RMSE for model validation from daily to yearly scales and allowed for unprecedented hour-by-hour evaluation with R^2 (0.80) even better than most of the other R^2 values on a daily scale (0.41∼0.85) [20–22,25,28,30–38].

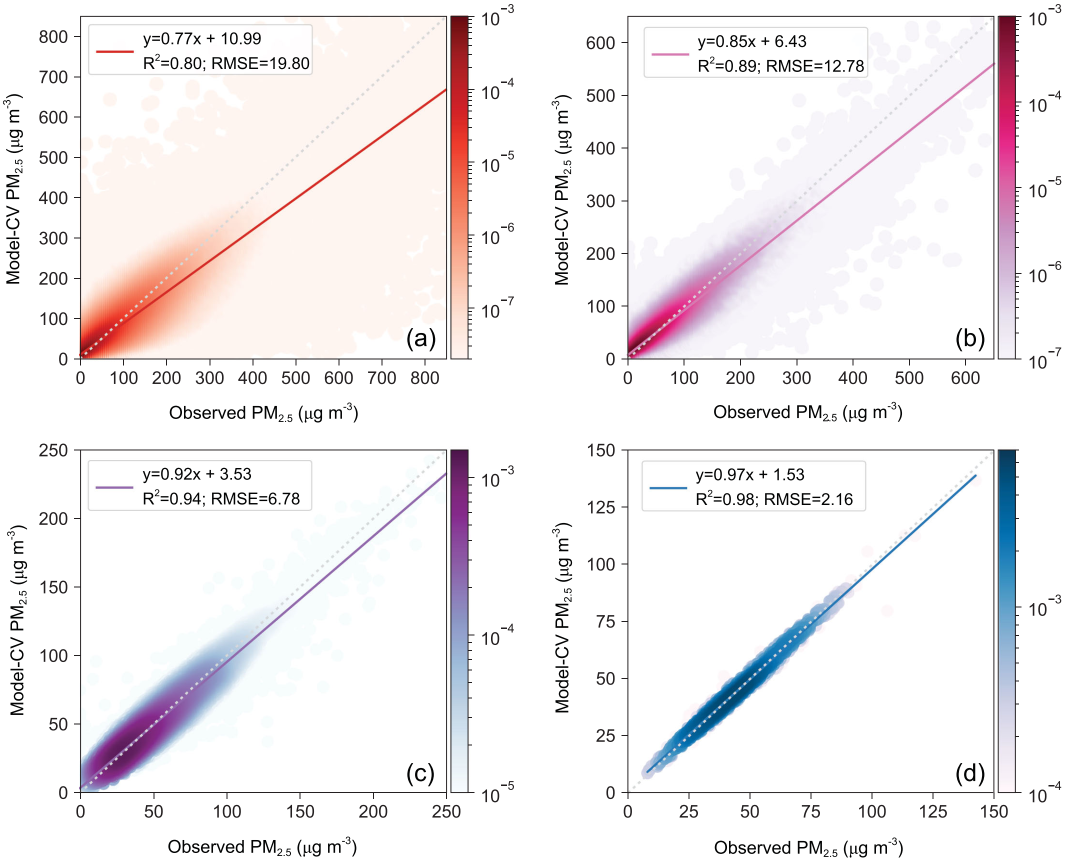

Figure 2. Density scatterplots of 10-fold CV results for (a) hourly (N = 31 863 778), (b) daily (N = 1 454 688), (c) monthly (N = 49 886) and (d) yearly (N = 1440) $PM_{2.5}$ from 2016 to 2018 across China (colors show probability distribution densities).

Robust prediction of hourly $PM_{2.5}$ across China

Our model's predictive capability, which is crucial for retrieving historical $PM_{2.5}$ datasets, was evaluated using 10 522 939 'unseen' samples from 2019. Figure 3 presents the overall correlations of model-predicted $PM_{2.5}$ and observed $PM_{2.5}$ on different timescales. For hourly $PM_{2.5}$ prediction, the overall R^2 and RMSE are 0.75 and 19.19 μg m^{-3}, respectively, which are closely consistent with the 10-fold CV results (Fig. 2a). This excellent relevance indicates that our model has a robust predictive capability that can construct hourly historical $PM_{2.5}$ datasets feasibly and accurately. The predictive power of this model is even better for $PM_{2.5}$ prediction on larger timescales. For daily estimations with 477 867 samples, the overall R^2 and RMSE are 0.84 and 13.82 μg m^{-3}, respectively. For monthly and yearly estimations, the overall R^2 values increase to 0.88 and 0.87, respectively, and the RMSE values decrease to 8.39 μg m^{-3} and 5.55 μg m^{-3}, respectively. To better evaluate the predictive power of our model, we compared our predictive scores with those in recent studies (Table 1). As mentioned above, the R^2 and RMSE of the predictions are significantly worse than those of the 10-fold CV, particularly for the models based on satellite-retrieved AOD. The best predictive R^2 of AOD-based models is only 0.58 on a daily scale, which indicates that there will be potential biases that cannot be ignored when we estimate $PM_{2.5}$ datasets using those models. Compared with the other models in Table 1, our model can provide unprecedented hour-by-hour predictions on $PM_{2.5}$ and gains considerable advantages in predictive capacity from daily to yearly scales. These advantages mainly result from the incorporation of spatial features from 19 surrounding meteorological stations. If these spatial features are removed, the predictive capacity of our model is significantly reduced, with the R^2 values decreasing to 0.61 and 0.72 on hourly and daily scales, respectively. This kind of performance is only slightly better than that of models in previous studies.

To evaluate the predictive capacity of our model in different regions of China, we obtained the spatial distribution of R^2 values between observed $PM_{2.5}$ and model-predicted $PM_{2.5}$ on an hourly scale

Table 1. Statistics for the comparison of the validation performance and predictive capability of different models from hourly to yearly scales.

Basic information			Model validation								Predictive capability							
			Hourly		Daily		Monthly		Yearly		Hourly		Daily		Monthly		Yearly	
Primary predictor	Model	References	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
AOD	GWR	[30]	-[a]	-	0.64	32.98	-	-	-	-	-	-	-	-	-	-	-	-
AOD	Stage-1	[32]	-	-	0.78	27.99	-	-	-	-	-	-	-	-	-	-	-	-
AOD	Stage-2		-	-	0.79	27.42	-	-	-	-	-	-	0.41	-	0.73	-	0.79	-
AOD	TSAM	[31]	-	-	-	-	-	-	0.80	22.75	-	-	-	-	-	-	-	-
AOD	GWR	[33]	-	-	0.79	18.60	-	-	-	-	-	-	-	-	-	-	-	-
Visibility	ER	[38]	-	-	0.42	-	-	-	-	-	-	-	0.38	-	-	-	-	-
AOD	Gaussian	[36]	-	-	0.81	21.87	-	-	-	-	-	-	-	-	-	-	-	-
AOD	GRNN	[34]	-	-	0.67	20.93	-	-	-	-	-	-	-	-	-	-	-	-
Visibility	LMEM	[35]	-	-	-	-	0.71	25.62	-	-	-	-	0.60	-	0.71	-	-	-
AOD	GTWR	[37]	-	-	0.80	18.00	-	-	-	-	-	-	0.47	12.03	-	-	-	-
Merra-2 $PM_{2.5}$	RF	[20]	-	-	-	-	0.88	14.89	-	-	-	-	-	-	0.74	17.80	0.76	11.35
AOD	Ensemble	[21]	-	-	-	-	0.79	21.00	-	-	-	-	0.58	29.00	0.76	15.70	-	-
AOD	MLR	[22]	-	-	0.41	20.04	-	-	-	-	-	-	0.38	21.97	-	-	-	-
	GWR		-	-	0.53	23.28	-	-	-	-	-	-	0.44	26.47	-	-	-	-
	Stage-1		-	-	0.65	19.50	-	-	-	-	-	-	0.31	27.73	-	-	-	-
	Stage-2		-	-	0.71	8.59	-	-	-	-	-	-	0.35	27.65	-	-	-	-
	RF		-	-	0.81	17.91	-	-	-	-	-	-	0.53	28.09	-	-	-	-
	STRF		-	-	0.85	15.57	-	-	-	-	-	-	0.55	27.38	-	-	-	-
AOD	Ensemble	[25]	-	-	-	-	0.91	9.30	-	-	-	-	-	-	0.78	14.00	-	-
Visibility	Xgboost	[28]	-	-	0.79	15.75	0.92	6.75	-	-	-	-	0.60	25.34	0.80	14.75	0.83	10.10
Visibility	LightGBM	This study	0.80	19.80	0.89	12.78	0.94	6.78	0.98	2.16	0.75	19.19	0.84	13.82	0.88	8.39	0.87	5.55

[a] '-' indicates no data.

(Fig. 4). Five key polluted regions were selected as focuses based on long-term trends in visibility [39], including (i) the NCP and the Guanzhong Plain (GZP) in northern China; (ii) the Yangtze River Delta (YRD) region and the Two Lakes Basin (TLB) along the middle and lower reaches of the Yangtze River; (iii) the Pearl River Delta (PRD) region in southern China; (iv) the Sichuan Basin (SB) in southwestern China; and (v) the Northeast China Plain (NeCP) [39,40]. As shown in Fig. 4, the predictive capacity of our model is remarkable in the five regions and is much better in more polluted regions. Among these regions, the model presents the most impressive predictive performance on the NCP, with R^2 values generally more than 0.80. Following the NCP, the model also shows high accuracy in $PM_{2.5}$ prediction on the GZP, with R^2 values over 0.80. The SB, which is a cloudy basin with more than 70% of AOD values missing, still presents reliable predictive performance, with R^2 of ~0.75. On the YRD and TLB, the R^2 values fluctuate between 0.65 and 0.90 but still exceed 0.80 in most cases. On the NeCP and PRD, the predictive performance is also acceptable with R^2 values of approximately or over 0.70. The regional differences of R^2 in these five regions might be affected by three factors, including pollution levels, RH and the distribution of meteorological stations. Under low levels of pollution, $PM_{2.5}$ concentrations are not closely related to visibility that is the most important feature for our model. As pollution levels increase, visibility is increasingly affected by $PM_{2.5,}$ and the non-linear relationship between these two variables might be more easily built by our model. As a result, the NCP and the GZP, which experience the most severe aerosol pollution, exhibit the best performance of predictive power among all the five regions. In addition to $PM_{2.5}$, visibility is also affected by RH that can enhance aerosol hygroscopic growth [41]. RH exhibited significant differences between northern and southern China. In northern China, RH remains relatively low, and haze frequently occurs; while in southern China, RH remains relatively high, and mist or fog often occurs [42]. The non-linear relationship between visibility and $PM_{2.5}$ is more complicated in southern China and might be more difficult to build by our model. As a result, in southern China, including the PRD, the TLB, the YRD and the SB, the R^2 is slightly lower than that in northern China, including the NCP and the GZP. It is worth noting that the R^2 is also slightly lower in parts of northern China, including Inner Mongolia and the NeCP, which might result from

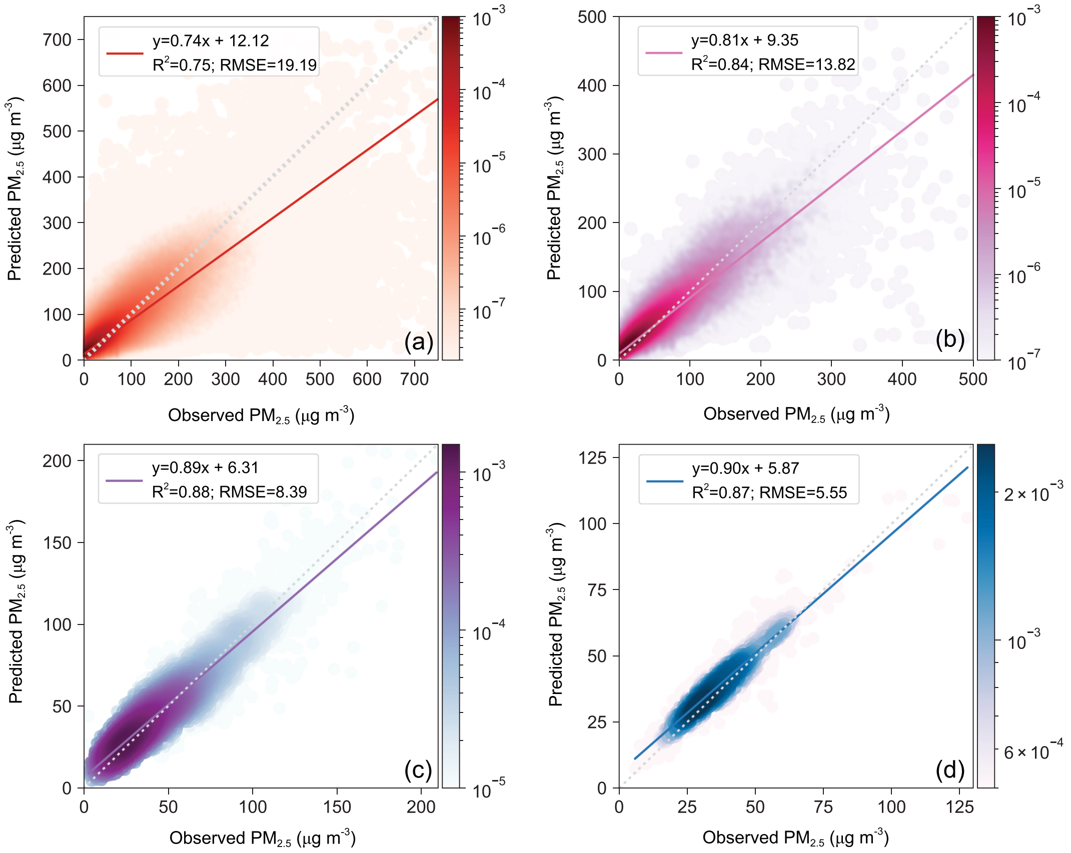

Figure 3. Density scatterplots of observed $PM_{2.5}$ and predicted $PM_{2.5}$ on (a) hourly (N = 10 522 939), (b) daily (N = 477 867), (c) monthly (N = 49 886) and (d) yearly (N = 1440) timescales in 2019 across China (colors are probability distribution densities).

the sparsity of meteorological stations that cannot truly reflect real situations. In contrast to the model performance in these five regions, at dozens of stations on the Tibet Plateau and its surrounding areas the model shows poor performance in hourly $PM_{2.5}$ prediction, with R^2 less than 0.40. The 25 stations with the lowest R^2 values were extracted to explore the causes of the low R^2 values, while another 25 stations with the highest R^2 values were used as contrasts. We found that the poor predictive capability at these stations mainly results from extremely low $PM_{2.5}$ concentrations ($<20\ \mu g\ m^{-3}$) that cannot effectively be reflected by visibility and long-distance surrounding meteorological stations (Fig. S5). The nearest meteorological station is ∼80 km on average away from the $PM_{2.5}$ station, and the 20th nearest meteorological station is 300 km away (Fig. S5). Such a distance indicates that surrounding meteorological stations cannot truly reflect real situations around $PM_{2.5}$ stations. Nevertheless, this disadvantage will be overcome when we incorporate regional meteorological stations in the future.

Previous studies also revealed that training regional models for each region can improve model performance due to significant spatial heterogeneity in relationships between $PM_{2.5}$ and meteorological variables [21]. Therefore, we selected three representative regions and trained regional models for each region, respectively. Figure S6 shows performance differences between the national and regional models on the NCP, the YRD and the Tibet Plateau. Compared with regional models, the national model performed almost equally well on the NCP, slightly better on the YRD, and slightly worse on the Tibet Plateau (Fig. S6). This finding indicates that regional models for each region cannot provide more accurate results. The spatial heterogeneity might have been incorporated into our national model by taking advantage of spatial features.

Given our model's hourly predictions, we assessed its predictive capacity for diurnal variations in $PM_{2.5}$. Figure S7 shows a clear diurnal variation in observed $PM_{2.5}$ across China. $PM_{2.5}$ concentrations significantly increased after 20:00 (Beijing Time, BJT) and decreased after 12:00. This diurnal variation is well captured by our $PM_{2.5}$ predictions, which are almost the same as the $PM_{2.5}$ observations (Fig. S7). Due to the diurnal $PM_{2.5}$ variations, using satellite-based AOD as the daily average, which is

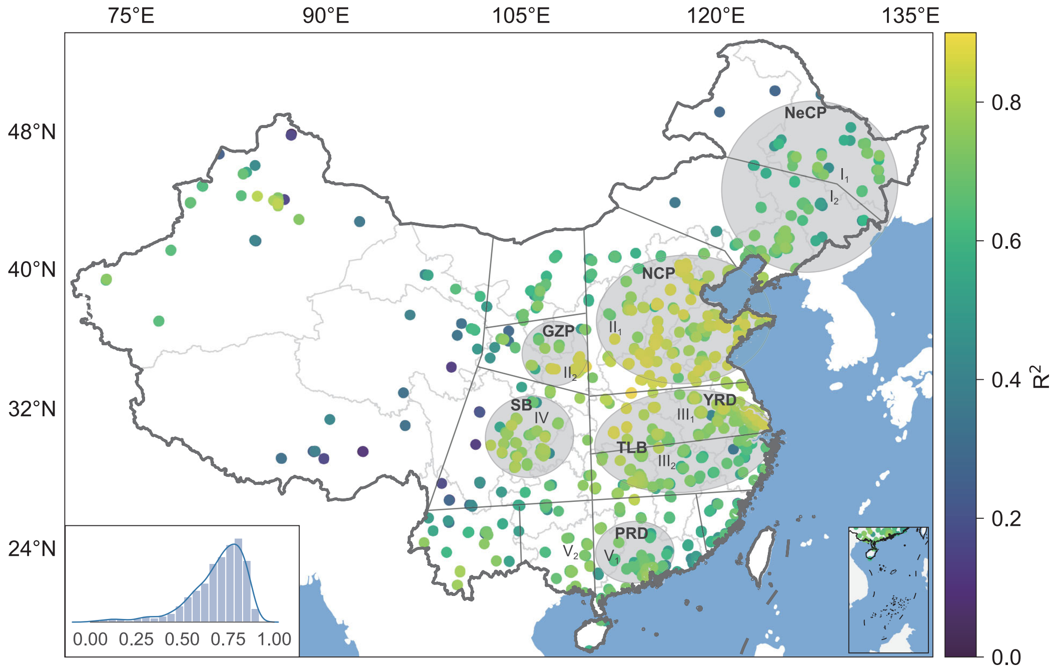

Figure 4. Spatial distribution of R^2 between observed $PM_{2.5}$ and predicted $PM_{2.5}$ on an hourly scale in 2019 across China. Eastern China is divided into similar visibility-changing regions with black lines as defined in Zhang *et al.* [39], and key polluted regions marked with shaded circles. Review drawing number: GS(2020)6868.

obtained twice a day (using MODIS as an example), will inevitably overestimate the actual daily conditions.

The robust prediction of our model can also be demonstrated by the hourly time series of observed and predicted $PM_{2.5}$ at several representative stations in Beijing, Shijiazhuang, Xi’an, Chengdu and Shanghai. As shown in Fig. S8, this model accurately predicts low values, high values and variation trends in both low-pollution and high-pollution areas.

Gridded $PM_{2.5}$ networks at high spatial-temporal resolutions

The distribution of $PM_{2.5}$ stations is uneven, with most stations located in urban areas in eastern China. In contrast, meteorological stations are distributed more evenly across China at a higher density. By taking advantage of spatial features from surrounding meteorological stations, our model can construct densely-gridded $PM_{2.5}$ networks at high temporal resolutions. For better visualization, we set the grid point precision to 0.25°, which can be further increased if required with consideration for numbers, distribution density and spacing distance of meteorological stations in the target area. Figure 5a and b shows the distribution of observed $PM_{2.5}$ stations and gridded $PM_{2.5}$ networks from our prediction. The gridded networks accurately correspond to observed $PM_{2.5}$ concentrations and provide more detailed information on spatial distributions. We found that low $PM_{2.5}$ concentrations occur in the northwestern mountain areas and that several pollution centers existed in the west and south, as well as in the hinterlands of the mountains. The region in southern Hebei and north-central Henan experience the widest and highest pollution among all the polluted centers. Compared with that region, the region in central Shandong undergoes less severe pollution. Another two centers that experience weak pollution at small ranges are located in eastern Hebei and the hinterlands in Shanxi. Apart from the assessment on a yearly scale, gridded networks from diurnal variations are illustrated in Fig. S9. Compared with observations in Fig. S10, the gridded networks present the diurnal variations in a better and clearer way.

The accuracy of the gridded $PM_{2.5}$ networks depends on whether the model can well predict $PM_{2.5}$ concentrations at locations outside the scope of the training stations. In Beijing, 23 regional $PM_{2.5}$ stations were untouched during the training process (Fig. 5c) and thus can be used to evaluate this kind of accuracy. Hourly $PM_{2.5}$ concentrations in these stations were predicted by the model and compared with $PM_{2.5}$ observations in 2019. As shown in Fig. 5d, the R^2 values exceed 0.75 at 22 of 23 regional stations and do not exhibit significant differences between national and regional stations.

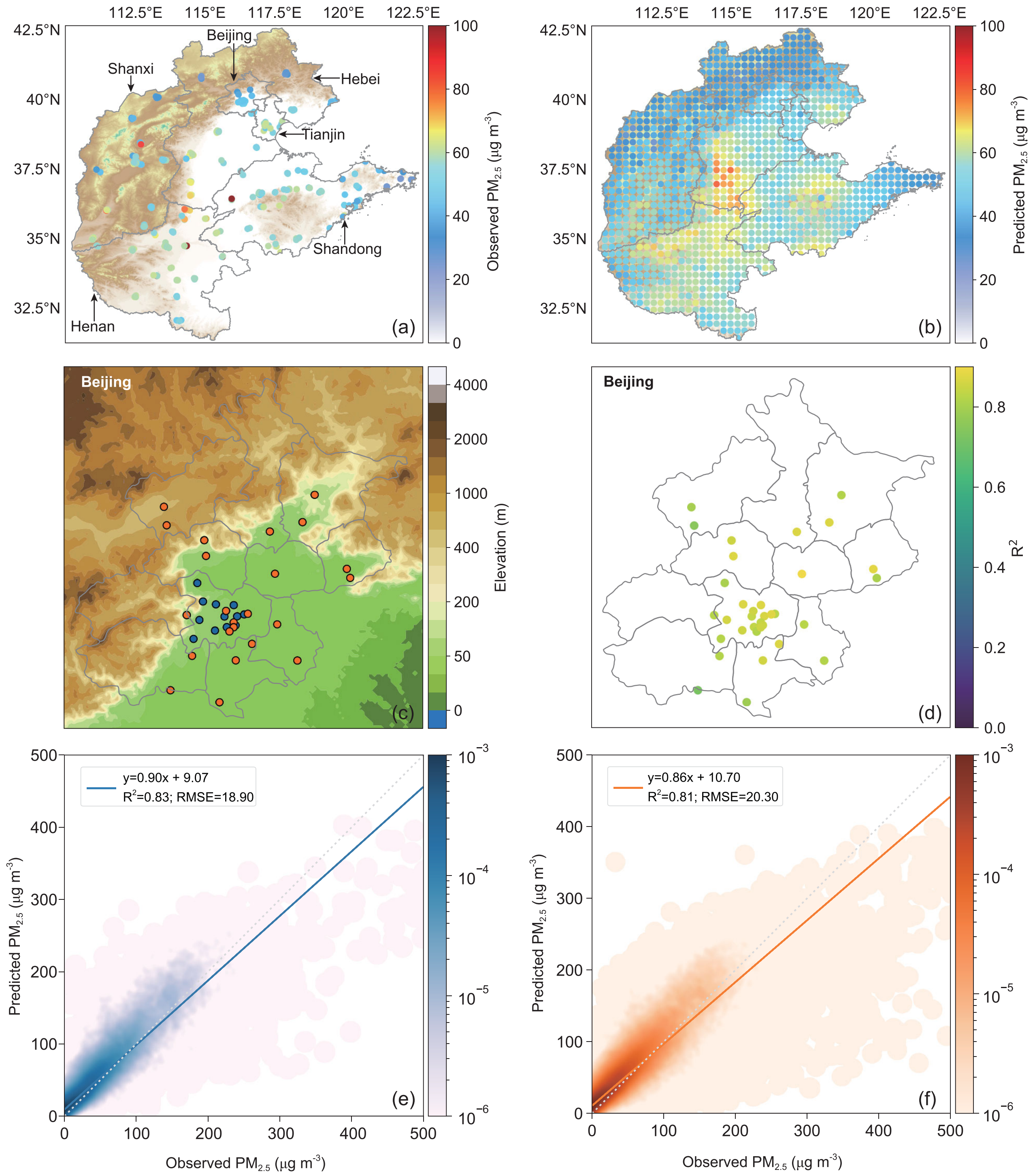

Figure 5. Spatial distribution of (a) observed $PM_{2.5}$ and (b) predicted gridded networks of $PM_{2.5}$ on a yearly scale on the North China Plain; (c) the distributions of 12 national stations (blue) that have been used during the training process and 23 regional stations (orange) that are untouched during the training process in Beijing; (d) the distribution of R^2 for both national and regional stations; (e) density scatterplots of observed $PM_{2.5}$ and predicted $PM_{2.5}$ for 12 national stations on an hourly scale and (f) density scatterplots of observed $PM_{2.5}$ and predicted $PM_{2.5}$ for 23 untouched regional stations on an hourly scale. Review drawing number: GS(2020)6868.

For hourly $PM_{2.5}$ concentrations at all 23 regional stations, the R^2 and RMSE are 0.81 and 20.30, respectively, which are just slightly weaker than those ($R^2 = 0.83$, RMSE = 18.90) at 12 national stations (Fig. 5e and f). These results indicate that our model is able to well predict $PM_{2.5}$ concentrations at locations both inside and outside the scope of the training stations.

Several polluted stations that almost coincide with other stations are not incorporated into our gridded networks. This phenomenon is mainly due to our grid precision settings and will be effectively resolved if sufficiently high precision is set. Furthermore, it might be difficult for us to construct gridded networks in western China where meteorological stations are scarce, but this will be significantly improved when we introduce regional meteorological stations in the future. The significant increase in the number of meteorological stations will enable us to build densely-gridded networks on an hourly scale.

CONCLUSION

For retrieving historical $PM_{2.5}$ datasets, satellite-based AOD has some inherent limitations that are difficult to overcome, i.e. massive missing values due to cloud cover, low sampling frequency and weak predictive capability for data outside the training period. Here, hourly meteorological observations with over 40 000 000 samples were employed to overcome the disadvantages of satellite-based retrieval. Developing a novel feature engineering approach to extract spatial features of surrounding stations, we built a LightGBM model that outperformed previous models regardless of validation performance or predictive capability. The R^2 and RMSE of the 10-fold CV of our model are 0.80 and 19.80 $\mu g\ m^{-3}$ on an hourly scale and 0.89 and 12.78 $\mu g\ m^{-3}$ on a daily scale, respectively. This model can even achieve unprecedented hour-by-hour $PM_{2.5}$ predictions with high and stable accuracy. For hourly $PM_{2.5}$ prediction, the overall R^2 and RMSE are 0.75 and 19.19 $\mu g\ m^{-3}$, respectively. For daily, monthly and yearly $PM_{2.5}$ predictions, the R^2 values are 0.84, 0.88 and 0.87, respectively, and the RMSE values are 13.82 $\mu g\ m^{-3}$, 8.39 $\mu g\ m^{-3}$ and 5.55 $\mu g\ m^{-3}$, respectively. By taking advantage of spatial features, our model can also construct hourly gridded networks of $PM_{2.5}$ at high spatial resolutions that provide more detailed information on spatial distribution. Our results show that this model has great potential in reconstructing historical $PM_{2.5}$ datasets at high spatial-temporal resolutions and retrieving real-time gridded $PM_{2.5}$ networks across China. However, this model still has some weaknesses, with the main weakness being the poor performance in predicting hourly $PM_{2.5}$ in dozens of stations in western China where meteorological stations are sparse. This disadvantage will be effectively overcome when regional meteorological stations are incorporated. In the future, we will employ this model to hindcast two sets of historical $PM_{2.5}$ datasets from the 1950s. One dataset will be for existing $PM_{2.5}$ stations and the other for gridded networks of $PM_{2.5}$. Then, we will incorporate regional meteorological stations to improve our model's precision and then use the model to retrieve gridded networks of $PM_{2.5}$ at high spatial-temporal resolutions. This will serve to overcome the disadvantages of existing $PM_{2.5}$ stations that are unevenly distributed and far fewer than meteorological stations in number. In addition, the retrieved historical $PM_{2.5}$ datasets will be assimilated into models to produce long-term re-analysis that incorporates interactions between aerosols and the physical processes of the climate system. This re-analysis will facilitate investigating aerosols' impacts on society, epidemiology and climate change.

MATERIALS AND METHODS

Observational data

This study used ground-based $PM_{2.5}$ observations at ∼1600 national stations across China from 2016 to 2019 (Fig. 1a). The hourly $PM_{2.5}$ data were archived at the China National Environmental Monitoring Center (CNEMC, http://www.cnemc.cn, 11 October 2020). We conducted a series of quality controls to produce high-quality data. A total of 1440 stations met the quality criterion of having at least 60% of valid data and were retained for this study. Severe outliers that were abnormally higher than surrounding data were effectively removed using a method that compared the differences between hourly $PM_{2.5}$ and a five-point moving average. After several tests, the threshold of 150 $\mu g\ m^{-3}$ was effective in eliminating outliers (Fig. S11). With this threshold, ∼3% of the hourly $PM_{2.5}$ was removed. A total of 42 386 717 samples remained for model development and application. There are also 23 regional $PM_{2.5}$ stations in Beijing in addition to 12 national stations (https://quotsoft.net, 11 October 2020). Hourly $PM_{2.5}$ data from these 23 stations were not used for model training but for evaluating predictive capability.

National ground-based surface meteorological observations from 2016 to 2019 were archived at the National Meteorological Information Center of the China Meteorological Administration. Similarly, we only use the stations with valid values over 60%. There were 2450 stations remaining in total (Fig. 1). Three meteorological variables, including visibility, RH and temperature, were selected from the surface observations as the main predictors to develop the LightGBM model.

Feature engineering approach

The longitude, latitude and time variables (year, month, day and hour) of $PM_{2.5}$ stations were selected as features at first (Fig. 1b). Then we matched each $PM_{2.5}$ station with its nearest meteorological station and added the visibility, RH, and temperature from that meteorological station as features (Fig. 1b). The distance between these two stations was also added as a feature. Previous studies have shown that the pollution levels at one station are largely affected by the surrounding environment, i.e. pollution transport is the primary cause of early pollution formation in Beijing [43,44]. Therefore, spatial effects need to be considered for better accuracy. Here, we developed a novel feature engineering approach that extracts spatial features to incorporate the surrounding environment's effects. Specifically, the nearest 20 meteorological stations for each $PM_{2.5}$ station were matched, and the nearest station that had already been used was excluded. We extracted five variables from the remaining 19 meteorological stations, including latitude, longitude, visibility, temperature and RH. For each variable, we calculated the mean values, maximum values, minimum values, standard deviation and skewness values in order and added them as features. After all the features were obtained, feature selection was performed to reduce training time and improve accuracy. We employed a relatively small sample dataset (1/3 of training data) to train the model and obtain each feature's importance in the dataset. According to the highest to lowest importance, we selected the top 30 features for the following model fitting, validation and evaluation. These features included visibility, distance and other spatial features.

The LightGBM model and its development

LightGBM is a state-of-the-art gradient-boosting framework that uses tree-based learning algorithms [45]. It is designed with faster training speed, lower memory usage, better accuracy and capability of handling large-scale data [45]. Until now, it could achieve slightly higher accuracy with much faster speed compared to XGBoost. Therefore, it was more appropriate to use LightGBM in our study, which included more than 40 million samples. Two metrics, R^2 and RMSE, were employed to quantify the quality of predictions. Then we trained the LightGBM model with 30 features and $PM_{2.5}$ labels from 2016 to 2018. To achieve a better performance, we performed hyperparameter tuning with a randomized search CV (RSCV) optimized by a randomly cross-validated search on parameter settings. In contrast to grid search CV (GRCV) that tries out all parameter values, RSCV only uses a fixed number of parameter settings to obtain a local optimal solution. This solution, which saves much computation time, was more realistic for our study. Based on RSCV and our tuning experiences, we finally selected the following hyperparameters: max_depth = 16, num_leaves = 127, min_data_in_leaf = 10, learning_rat = 0.05, feature_fraction = 0.80, bagging_fraction = 0.80, bagging_freq = 5, max_bin = 255, lambda_l1 = 0.5, lambda_l2 = 0.5 and num_boost_round = 1000. To evaluate the model performance, we performed 10-fold CV on the training data. Thirty features and $PM_{2.5}$ labels from 2016 to 2018 were randomly divided into 10 sets. For each of the 10 folds, a model was developed using the other nine folds as training data, and subsequently, the resulting model was validated with the data of this fold. The R^2 and RMSE reported by 10-fold CV reflect the averages of the values calculated in the loop. After the model was built, 30 meteorological features in 2019 were input into this model to generate $PM_{2.5}$ predictions for further evaluating our model's predictive capability on hourly, daily, monthly and annual scales. The flow of building the LightGBM model is illustrated as a conceptual figure (Fig. 1b).

Construction of gridded networks

The gridded meteorological input was generated to construct gridded $PM_{2.5}$ networks. The detailed process was demonstrated using the NCP as an example. First, we define an area with latitude from 30°N to 45°N and longitude from 110°E to 125°E, which can cover the whole NCP. This area was then gridded at 0.25° intervals, and 3600 grid points were generated with latitude and longitude as features. The nearest meteorological station was matched for each point, and four variables from this station are added to the point as features, including visibility, RH, temperature and distance. Spatial variables from the surrounding 19 stations are also added to the point as features. After that, we generate 3600 grid points with geological information, meteorological variables and spatial features. As these grid points are input into our model, gridded $PM_{2.5}$ networks are constructed.

Since 23 regional $PM_{2.5}$ stations in Beijing are excluded during the training process, hourly $PM_{2.5}$ concentrations in 2019 at these stations can be used to evaluate our model's performance. For each station, four variables of its nearest meteorological station and spatial variables from surrounding stations in 2019 are added to the station as input features. The generated input datasets at 23 regional stations are input into our model to produce predicted $PM_{2.5}$ that were further compared with observed $PM_{2.5}$.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

SUPPLEMENTARY DATA

Supplementary data are available at *NSR* online.

FUNDING

This work was supported by the National Key Research and Development Program of China (2016YFC0203306).

AUTHOR CONTRIBUTIONS

X.Z., Y.W. and H.C. designed the research and led the overall scientific questions. J.Z. and K.G. carried out data processing and analysis. X.S., J.S., L.Z., W.Z. and Y.Z. collected $PM_{2.5}$ data and meteorological observations. J.Z. wrote the first draft of the manuscript, and X.Z. revised the manuscript. All authors read and approved the final version.

Conflict of interest statement. None declared.

REFERENCES

1. Zhang X, Sun J and Wang Y *et al.* Factors contributing to haze and fog in China. *Chin Sci Bull* 2013; **58**: 1178–87.
2. Dang R and Liao H. Radiative forcing and health impact of aerosols and ozone in China as the consequence of clean air actions over 2012–2017. *Geophys Res Lett* 2019; **46**: 12511–9.
3. Ding YH and Liu YJ. Analysis of long-term variations of fog and haze in China in recent 50 years and their relations with atmospheric humidity. *Sci China Earth Sci* 2014; **57**: 36–46.
4. Chen H and Wang H. Haze days in North China and the associated atmospheric circulations based on daily visibility data from 1960 to 2012. *J Geophys Res Atmos* 2015; **120**: 5895–909.
5. Pope CA, Burnett RT and Thun MJ *et al.* Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *J Am Med Assoc* 2002; **287**: 1132–41.
6. Beelen R, Hoek G and Den Brandt PAV *et al.* Long-term effects of traffic-related air pollution on mortality in a Dutch cohort (NLCS-AIR study). *Environ Health Perspect* 2007; **116**: 196–202.
7. Smith JD, Mitsakou C and Kitwiroon N *et al.* London Hybrid Exposure Model (LHEM): improving human exposure estimates to NO_2 and $PM_{2.5}$ in an urban setting. *Environ Sci Technol* 2016; **50**: 11760–8.
8. Wang Z, Akimoto H and Uno I. Neutralization of soil aerosol and its impact on the distribution of acid rain over east Asia: observations and model results. *J Geophys Res Atmos* 2002; **107**: D194389.
9. Mahowald N. Aerosol indirect effect on biogeochemical cycles and climate. *Science* 2011; **334**: 794–6.
10. Bai N, Khazaei M and van Eeden SF *et al.* The pharmacology of particulate matter air pollution-induced cardiovascular dysfunction. *Pharmacol Ther* 2007; **113**: 16–29.
11. Samoli E, Peng RD and Ramsay T *et al.* Acute effects of ambient particulate matter on mortality in Europe and North America: results from the APHENA study. *Environ Health Perspect* 2008; **116**: 1480–6.
12. Chen X, Zhang LW and Huang JJ *et al.* Long-term exposure to urban air pollution and lung cancer mortality: a 12-year cohort study in Northern China. *Sci Total Environ* 2016; **571**: 855–61.
13. Wang H, Zhang X and Gong S *et al.* Radiative feedback of dust aerosols on the East Asian dust storms. *J Geophys Res Atmos* 2010; **115**: D23214.
14. Wei P, Cheng SY and Li JB *et al.* Impact of boundary-layer anticyclonic weather system on regional air quality. *Atmos Environ* 2011; **45**: 2453–63.
15. Boucher O, Randall D and Artaxo P *et al.* Clouds and aerosols. In: *Climate Change 2013: the Physical Science Basis Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press, 2013, 571–657.
16. Zhang RY, Wang GH and Guo S *et al.* Formation of urban fine particulate matter. *Chem Rev* 2015; **115**: 3803–55.
17. Gui K, Che H and Wang Y *et al.* Satellite-derived $PM_{2.5}$ concentration trends over Eastern China from 1998 to 2016: relationships to emissions and meteorological parameters. *Environ Pollut* 2019; **247**: 1125–33.
18. Xiao Q, Wang Y and Chang HH *et al.* Full-coverage high-resolution daily $PM_{2.5}$ estimation using MAIAC AOD in the Yangtze River Delta of China. *Remote Sens Environ* 2017; **199**: 437–46.
19. Geng G, Murray NL and Chang HH *et al.* The sensitivity of satellite-based $PM_{2.5}$ estimates to its inputs: implications to model development in data-poor regions. *Environ Int* 2018; **121**: 550–60.
20. Huang K, Xiao Q and Meng X *et al.* Predicting monthly high-resolution $PM_{2.5}$ concentrations with random forest model in the North China Plain. *Environ Pollut* 2018; **242**: 675–83.
21. Xiao Q, Chang HH and Geng G *et al.* An ensemble machine-learning model to predict historical $PM_{2.5}$ concentrations in China from satellite data. *Environ Sci Technol* 2018; **52**: 13260–9.
22. Wei J, Huang W and Li Z *et al.* Estimating 1-km-resolution $PM_{2.5}$ concentrations across China using the space-time random forest approach. *Remote Sens Environ* 2019; **231**: 111221.
23. Geng G, Meng X and He K *et al.* Random forest models for $PM_{2.5}$ speciation concentrations using MISR fractional AODs. *Environ Res Lett* 2020; **15**: 034056.
24. Shin M, Kang Y and Park S *et al.* Estimating ground-level particulate matter concentrations using satellite-based data: a review. *GIScience Remote Sens* 2020; **57**: 174–89.

25. Xiao Q, Geng G and Liang F *et al.* Changes in spatial patterns of $PM_{2.5}$ pollution in China 2000–2018: impact of clean air policies. *Environ Int* 2020; **141**: 105776.
26. Bi J, Belle JH and Wang Y *et al.* Impacts of snow and cloud covers on satellite-derived $PM_{2.5}$ levels. *Remote Sens Environ* 2019; **221**: 665–74.
27. Chen ZY, Zhang TH and Zhang R *et al.* Extreme gradient boosting model to estimate $PM_{2.5}$ concentrations with missing-filled satellite data in China. *Atmos Environ* 2019; **202**: 180–9.
28. Gui K, Che H and Zeng Z *et al.* 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* 2020; **141**: 105801.
29. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019; **1**: 206–15.
30. Ma Z, Hu X and Huang L *et al.* Estimating ground-level $PM_{2.5}$ in China using satellite remote sensing. *Environ Sci Technol* 2014; **48**: 7436–44.
31. Fang X, Zou B and Liu X *et al.* Satellite-based ground $PM_{2.5}$ estimation using timely structure adaptive modeling. *Remote Sens Environ* 2016; **186**: 152–63.
32. Ma Z, Hu X and Sayer AM *et al.* Satellite-based spatiotemporal trends in $PM_{2.5}$ concentrations: China, 2004–2013. *Environ Health Perspect* 2016; **124**: 184–92.
33. You W, Zang Z and Zhang L *et al.* National-scale estimates of ground-level $PM_{2.5}$ concentration in China using geographically weighted regression based on 3 km resolution MODIS AOD. *Remote Sens* 2016; **8**: 184.
34. Li T, Shen H and Zeng C *et al.* Point-surface fusion of station measurements and satellite observations for mapping $PM_{2.5}$ distribution in China: methods and assessment. *Atmos Environ* 2017; **152**: 477–89.
35. Liu M, Bi J and Ma Z. Visibility-based $PM_{2.5}$ concentrations in China: 1957–1964 and 1973–2014. *Environ Sci Technol* 2017; **51**: 13161–9.
36. Yu W, Liu Y and Ma Z *et al.* Improving satellite-based $PM_{2.5}$ estimates in China using Gaussian processes modeling in a Bayesian hierarchical setting. *Sci Rep* 2017; **7**: 7048.
37. He Q and Huang B. Satellite-based mapping of daily high-resolution ground $PM_{2.5}$ in China via space-time regression modeling. *Remote Sens Environ* 2018; **206**: 72–83.
38. Shen Z, Cao J and Zhang L *et al.* Retrieving historical ambient $PM_{2.5}$ concentrations using existing visibility measurements in Xi’an, Northwest China. *Atmos Environ* 2016; **126**: 15–20.
39. Zhang XY, Wang YQ and Niu T *et al.* Atmospheric aerosol compositions in China: spatial/temporal variability, chemical signature, regional haze distribution and comparisons with global aerosols. *Atmos Chem Phys* 2012; **12**: 779–99.
40. Zhong J, Zhang X and Wang Y *et al.* The two-way feedback mechanism between unfavorable meteorological conditions and cumulative aerosol pollution in various haze regions of China. *Atmos Chem Phys* 2019; **19**: 3287–306.
41. Zhang L, Sun JY and Shen XJ *et al.* Observations of relative humidity effects on aerosol light scattering in the Yangtze River Delta of China. *Atmos Chem Phys* 2015; **15**: 2853–904.
42. Guo B, Wang Y and Zhang X *et al.* Temporal and spatial variations of haze and fog and the characteristics of $PM_{2.5}$ during heavy pollution episodes in China from 2013 to 2018. *Atmos Pollut Res* 2020; **11**: 1847–56.
43. Zhong J, Zhang X and Wang Y *et al.* Relative contributions of boundary-layer meteorological factors to the explosive growth of $PM_{2.5}$ during the red-alert heavy pollution episodes in Beijing in December 2016. *J Meteorolog Res* 2017; **31**: 809–19.
44. Zhong J, Zhang X and Dong Y *et al.* Feedback effects of boundary-layer meteorological factors on cumulative explosive growth of $PM_{2.5}$ during winter heavy pollution episodes in Beijing from 2013 to 2016. *Atmos Chem Phys* 2018; **18**: 247–58.
45. Ke G, Meng Q and Finley T *et al.* LightGBM: a highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems*. 2017, 3146–54.