A Validation Approach Considering the Uneven Distribution of Ground Stations for Satellite-Based $PM_{2.5}$ Estimation

Tongwen Li, Huanfeng Shen, *Senior Member, IEEE*, Chao Zeng, and Qiangqiang Yuan, *Member, IEEE*

***Abstract*—Satellite remote sensing has been increasingly employed for the estimation of ground-level atmospheric $PM_{2.5}$. There have been several cross-validation (CV) approaches applied for the validation of satellite-based $PM_{2.5}$ estimation models. However, these validation approaches often lead to confusion, due to the unclear applicable conditions. For this, we fully analyze and assess the existing validation approaches, and provide suggestions on applicable conditions for them. Furthermore, the existing validation approaches still have limitations to disregard the uneven distribution of ground stations, and tend to overestimate the performance of the $PM_{2.5}$ estimation models. To this end, a CV-based validation approach considering the uneven spatial distribution of monitoring stations (denoted as SDCV) is proposed. SDCV introduces the spatial distance between validation station and modeling station into the CV process, and evaluates the spatial performance through a strategy of excluding modeling stations within a specific distance. Meanwhile, this approach has designed reasonable evaluation indices for the model validation. Taking China as a case study, the results indicate that SDCV can yield a more complete and effective evaluation for the popular $PM_{2.5}$ estimation models than the traditional validation approaches.**

***Index Terms*—Aerosol optical depth (AOD), ground station distribution, $PM_{2.5}$, satellite remote sensing, validation.**

I. Introduction

With the rapid development of the economy, air pollution has evolved into an increasingly serious problem in recent years. As reported in a study conducted by the World Health Organization [1], public health has been heavily influenced by air pollution during the 21st century. Therein, fine particulate matter ($PM_{2.5}$, particulate matter with an aerodynamic diameter of less than 2.5 μm) is one of the main air pollutants [2]–[8]. Towards the monitoring of $PM_{2.5}$ pollution, ground stations are considered the most reliable way to obtain high-accuracy $PM_{2.5}$ measurements. However, due to the high cost of ground stations, the $PM_{2.5}$ station network is sparsely and unevenly distributed in space.

Owing to the broad spatiotemporal coverage, satellite remote sensing exactly owns the capacity to expand the monitoring of $PM_{2.5}$ beyond ground stations [9]–[13]. To estimate ground atmospheric $PM_{2.5}$ from satellite observations, the popular approach is to establish a statistical relationship between the satellite observations (e.g., aerosol optical depth (AOD) [12], top-of-atmosphere reflectance [14]) and ground $PM_{2.5}$ measurements. There have been numerous satellite-based $PM_{2.5}$ estimation models developed for the estimation of $PM_{2.5}$, primarily including the early statistical models, such as multiple linear regression [15], semiempirical model [16], and so on; and the more advanced statistical models, for instance, the linear mixed effects model [17], geographically weighted regression [18], [19], and neural networks [20]– [22], etc. With the use of these models, high-resolution ground $PM_{2.5}$ data can be effectively generated from satellite observations.

To evaluate the estimation accuracy of the satellite-based $PM_{2.5}$ estimation models, the $PM_{2.5}$ model estimates are usually compared with $PM_{2.5}$ station measurements. A cross-validation (CV) technique [23], which in fact leaves out some station-based $PM_{2.5}$ observations for the model validation, is often adopted for the validation of satellite-based $PM_{2.5}$ estimation models. Based on the CV technique, several validation approaches have been developed, including sample-based CV [14], [21], site-based CV [17], [24], region-based CV [25], [26], and time-based CV [27], [28]. In addition, some studies have concentrated on the historical prediction of ground $PM_{2.5}$. As a result, historical validation [29], [30], which is not derived from a CV technique, has also been exploited. To evaluate the $PM_{2.5}$ estimation model performance, some studies have adopted only one of these validation approaches, while some studies have simultaneously used several ones.

However, the existing validation approaches often lead to confusion due to the unclear applicable conditions. The same $PM_{2.5}$ estimation model may report notably different validation results with different validation approaches [30], the applicable conditions for each validation approach still remain unclear. On the other hand, the ground stations are often unevenly distributed in space, and can be clustered in the urban areas of cities [31].

Manuscript received December 8, 2019; revised February 19, 2020; accepted February 26, 2020. Date of publication March 30, 2020; date of current version April 17, 2020. This work was supported in part by the National Key R&D Program of China under Grant 2018YFB2100500 and Grant 2016YFC0200900 and in part by Major Projects of Technological Innovation of Hubei Province under Grant 2019AAA046. *(Corresponding author: Huanfeng Shen.)*

Tongwen Li and Chao Zeng are with the School of Resource and Environmental Sciences, Wuhan University, Wuhan 430072, China (e-mail: litw@whu.edu.cn; zengchaozc@hotmail.com).

Huanfeng Shen is with the School of Resource and Environmental Sciences, Wuhan University, Wuhan 430072, China, with the Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China, and also with the Key Laboratory of Geographic Information System Ministry of Education, Wuhan University, Wuhan 430072, China (e-mail: shenhf@whu.edu.cn).

Qiangqiang Yuan is with the Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China,, and also with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430072, China (e-mail: yqiang86@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2020.2977668

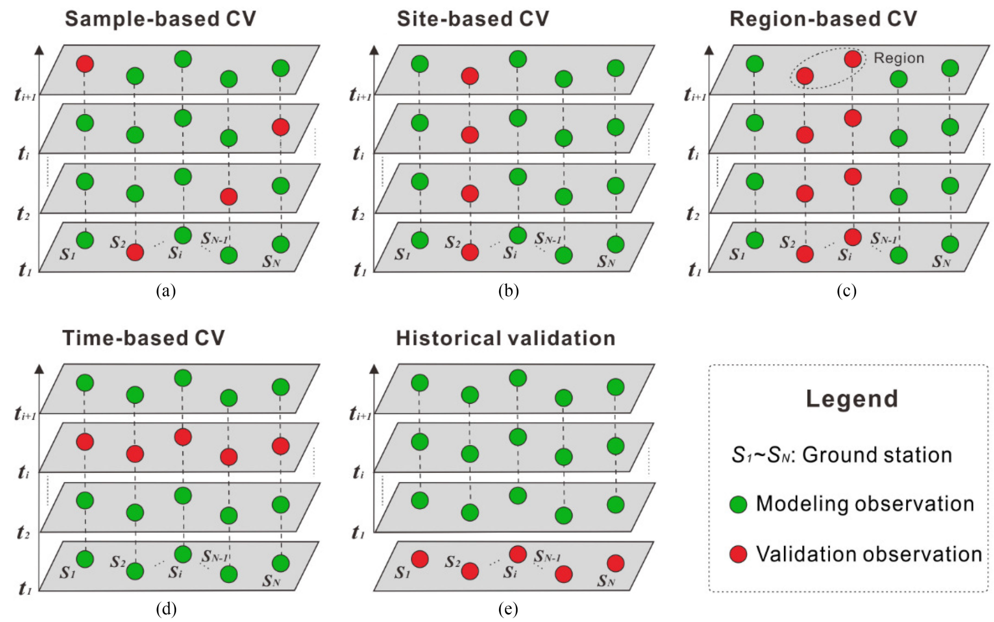


Fig. 1. Schematics of the various validation approaches. (a)–(d) Only represent one round of validation. (a) Sample-based CV. (b) Site-based CV. (c) Region-based CV. (d) Time-based CV. (e) Historical validation.

The monitoring stations are usually close to their neighbors, and the validation stations tend to have a close distance to the modeling stations. As a result, the previous validation approaches may only be able to evaluate the estimation accuracy for locations close to a monitoring station, and may fail to reflect the estimation accuracy for locations at a farther distance. Without consideration of the uneven station distribution, the previous validation approaches are likely to result in some bias for the evaluation of $PM_{2.5}$ estimation models.

Therefore, one of our main purposes is to comprehensively analyze and assess the existing validation approaches, and give suggestions on the applicable conditions for their use. Second, a CV-based validation approach that considers the uneven spatial distribution of monitoring stations is proposed. Taking China as an example, the proposed approach and the previous validation approaches are compared and assessed.

II. Previous Validation Approaches

Using ground station measurements to validate the estimates from satellite remote sensing is a common strategy. Hence, a common approach is to fit the satellite-based $PM_{2.5}$ estimation model using some of the station observations, and leave the other observations for the model validation. This solution is actually based on a CV technique [23]. For the k-fold CV, the samples (stations, regions, or time) are divided into k folds randomly and evenly. $k-1$ folds are then used for the model fitting, and the remaining one is used for the model validation. Finally, the abovementioned process is repeated k times to evaluate the model performance on each fold. When k is set to 10, this indicates the widely used 10-fold CV technique; when k is equivalent to the number of samples (stations, regions, or time), this is referred to as "leave-one-out CV". The 10-fold CV and leave-one-out CV techniques are the two most popular CV strategies. Meanwhile, the input data for the CV can be data samples, monitoring sites, stations in one region, or stations at one time, which are named sample-based CV, site-based CV, region-based CV, and time-based CV, respectively. Moreover, historical validation, which does not belong to a CV technique, has also been adopted for the validation of satellite-based $PM_{2.5}$ estimation models. The schematics of these validation approaches are illustrated in Fig. 1.

A. Sample-Based CV

Sample-based CV has been the most commonly adopted CV-based validation approach [14], [18], [21], [32], [33]. As presented in Fig. 1(a), we mix up the locations and times of the satellite-$PM_{2.5}$ matchup samples, and then randomly select some samples for the model validation. Hence, sample-based CV involves conducting the validation with integrated samples from both the spatial and temporal dimensions, and it is often employed to reflect the overall predictive ability of $PM_{2.5}$ estimation models. For a certain monitoring station, the samples from this station at some certain time can be utilized for the model fitting, and the samples from the other time are used for the model validation. This means that the modeling dataset and validation dataset may contain the same monitoring stations. Consequently, sample-based CV has limitations in that the same monitoring station may be simultaneously involved in the model fitting and model validation. This brings some bias when evaluating the predictive ability of a model for the satellite-based mapping of $PM_{2.5}$, because the locations with $PM_{2.5}$ values to be estimated have no ground stations in real life.

B. Site-Based CV

Unlike sample-based CV, the monitoring stations are randomly chosen for the model validation in site-based CV [see Fig. 1(b)]. For site-based CV, the validation stations are never included in the model fitting. Hence, site-based CV has the potential to evaluate the accuracy of $PM_{2.5}$ spatial prediction. For those $PM_{2.5}$ estimation models that include historical $PM_{2.5}$ data in the model fitting, the site-based CV approach may show a relatively large decrease in model performance, compared to the sample-based CV approach. This is because the historical $PM_{2.5}$ data of the validation stations are used in the sample-based CV approach, whereas not in site-based CV. In addition, this may also be related to the spatial representation of the site observations [34]. For another, as mentioned previously, the ground stations are often located in the centers of cities, so the validation stations are often very close to the modeling stations. Thus, the site-based CV has limitations in that it is prone to merely evaluating the $PM_{2.5}$ prediction accuracy on the locations with a close distance to the station. Finally, it is noteworthy that site-based CV in fact often refers to grid cell-based CV, because the $PM_{2.5}$ values from multiple stations within a grid cell are averaged in the model development [14], [35], [36].

C. Region-Based CV

To some degree, region-based CV has the potential to avoid the limitations of site-based CV. As can be observed in Fig. 1(c), certain regions (e.g., a province) are chosen for the model validation. The stations in the validation region are all used for the model validation, and it may, thus, be capable of evaluating the $PM_{2.5}$ spatial prediction accuracy for locations at a farther distance from the monitoring station. However, what is the optimal extent for the validation region? In view of the uneven distribution of monitoring stations, it is still a huge challenge to determine a reasonable extent for the model validation. For instance, province-based CV was conducted across China in our previous work [25]. Whether it is reasonable to leave one province of stations for the model validation to reflect the spatial prediction ability remains to be discussed.

D. Time-Based CV

The process of time-based CV is illustrated in Fig. 1(d). Unlike the abovementioned validation approaches that pay more attention to the evaluation of the spatial prediction, the time-based CV approach is aimed at evaluating the accuracy of the temporal prediction. Under some situations, satellite data may be available, whereas the $PM_{2.5}$ station observations are absent. How well the $PM_{2.5}$ estimation models perform in this situation without satellite-$PM_{2.5}$ matchup needs further evaluation. Hence, we randomly choose some times (e.g., some days during the study period) of observations for the model validation, and the remaining times of observations are utilized for the model fitting. Time-based CV can, thus, be expected to evaluate the prediction accuracy for those times without satellite-$PM_{2.5}$ matchups [27]. However, in a real situation, the times with satellite data but without $PM_{2.5}$ data are relatively rare. Therefore, due to the limited applicable conditions, the use of time-based CV is limited. In addition, some $PM_{2.5}$ estimation models are built using the station $PM_{2.5}$ data of the estimation time (e.g., daily geographically weighted regression [37]), they are bound not to be evaluated via time-based CV.

E. Historical Validation

With the wide temporal coverage of satellite observations, the $PM_{2.5}$ estimation models own the capacities to predict historical $PM_{2.5}$ concentrations. Accordingly, the historical validation approach was developed to evaluate the accuracy of historical prediction. As presented in Fig. 1(e), a long time series of historical $PM_{2.5}$ data are collected for the validation of the $PM_{2.5}$ estimation models. The major differences between historical validation and time-based CV lie in the fact that the historical validation uses a long time series of historical $PM_{2.5}$ data for the model validation, whereas the time-based CV technique randomly chooses some $PM_{2.5}$ observations from a particular time in the study period. It is noteworthy that some studies have also exploited future $PM_{2.5}$ data for historical validation; for instance, Ma *et al.* [29] established a $PM_{2.5}$ estimation model with samples from 2013, and validated it using $PM_{2.5}$ data from the first half of 2014. The model was subsequently employed to predict historical $PM_{2.5}$ values during 2004–2012. The main limitation of the historical validation approach is that it is often difficult to collect sufficient $PM_{2.5}$ data for validation.

Finally, through the abovementioned analysis, the applicable conditions for and limitations of these validation approaches are summarized in Table I.

III. Proposed Validation Approach

As explained before, $PM_{2.5}$ monitoring stations often have an uneven spatial distribution, e.g., they are often found mainly in the center of cities, and the monitoring stations are often close to their neighbors. As a result, site-based CV may only reflect the prediction accuracy of the locations near the monitoring stations. On the other hand, due to the uneven station distribution, the existing validation approaches face the risk of misjudging the superiorities of the satellite $PM_{2.5}$ estimation methods. For instance, the $PM_{2.5}$ estimation models that strongly depend on adjacent stations have an advantage due to the closeness of the validation stations and modeling stations; nevertheless, they may lose their superiority with increasing distance to the modeling station. It is, therefore, critical to develop a CV-based validation approach considering the uneven spatial distribution of monitoring stations (denoted as SDCV in this article), for a more complete spatial evaluation of the $PM_{2.5}$ estimation models.

A. Establishment of SDCV

Based on 10-fold CV, the monitoring stations ("station" is used for easier understanding, as it actually refers to the grid cell containing the station) are partitioned into 10 folds randomly, with approximately 10% of the total stations in each fold. For a given validation fold, supposing that m validation stations are collected, they then form a validation collection $\mathbf{S}_{\text{val}} = \{S_{v,1}, S_{v,2}, \ldots, S_{v,m}\}$, and $\mathbf{S}_{\text{fit}} = \{S_{f,1}, S_{f,2}, \ldots, S_{f,n}\}$ denotes the modeling collection with n stations $(n \approx 9 \cdot m)$, where S means the monitoring station (S_v and S_f denote the validation station and modeling station, respectively). To evaluate the performance considering the uneven distribution of stations, a

TABLE I
SUMMARY OF THE EXISTING VALIDATION APPROACHES

Approach	Applicable conditions	Limitations	Evaluation dimension
Sample-based CV	The overall prediction accuracy	Modeling dataset and validation dataset may contain the same stations	Overall
Site-based CV	The spatial prediction accuracy	Tends to only reflect the prediction accuracy of locations close to a station	Spatial
Region-based CV	The spatial (regional) prediction accuracy	Selection of the optimal extent for the validation region is uncertain	
Time-based CV	The temporal prediction accuracy	Times with satellite data but without $PM_{2.5}$ observations are relatively rare	Temporal
Historical validation	The temporal (historical) prediction accuracy	It is often difficult to collect sufficient historical $PM_{2.5}$ data	

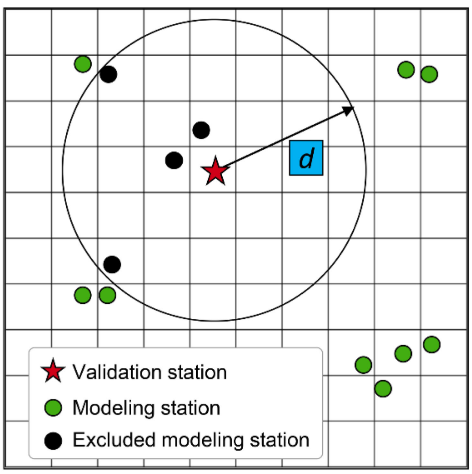


Fig. 2. Schematic for excluding modeling stations with the distance d.

distance of d (km) is set. For a given distance d, the modeling stations with a distance to any validation station of less than d are excluded from the modeling collection $\mathbf{S}_{fit}$, which is illustrated in Fig. 2. Thus, the distances of the validation stations to modeling stations are all no less than d. The $PM_{2.5}$ estimation model is established by considering the distance to the monitoring station, as shown in

$$\mathrm{PM}_{d\text{-fit}} = f_{(d)}\left(X_{d\text{-fit}}\right) \tag{1}$$

where $X_{d\text{-fit}}$ is the input variables (e.g., satellite data, meteorological data, etc.) of the modeling dataset updated by the distance d, and $f_{(d)}$ refers to the estimation function established by considering the distance d. Subsequently, based on the established $PM_{2.5}$ estimation model [see (1)], the $PM_{2.5}$ values can be predicted on the validation dataset, as shown in

$$\mathrm{PM}_{\mathrm{val}} = f_{(d)}\left(X_{\mathrm{val}}\right) \tag{2}$$

where X_{val} is the input variables of the validation dataset, which do not change with the varying distance d. By comparing the model estimates with the station-based $PM_{2.5}$ measurements, the model performance can be evaluated.

Fig. 3 illustrates the workflow of the proposed SDCV approach, which consists of four main steps. The details of the workflow are as follows.

Step 1: Based on 10-fold CV, the monitoring stations are divided into 10 folds randomly and averagely, where nine folds of stations are used as the modeling collection ($\mathbf{S}_{\mathrm{fit}}$), and the remaining fold forms the validation collection of stations ($\mathbf{S}_{\mathrm{val}}$).

Step 2: For a given distance d, the modeling collection is updated in terms of the distance from the validation station to modeling station (see Fig. 2). Using the updated modeling dataset, the $PM_{2.5}$ estimation model can be established, i.e., $f_{(d)}$.

Step 3: Based on the established $PM_{2.5}$ estimation model, the $PM_{2.5}$ values are predicted on the validation dataset. Thus, $PM_{2.5}$ values for all the stations can be estimated via a repeat process of the abovementioned steps.

Step 4: Statistical indices are designed to evaluate the $PM_{2.5}$ estimation models. First, with a given distance d, the corresponding statistics (e.g., coefficient of determination (R^2) and root-mean-square error (RMSE) with distance d: $\mathrm{R}^2(d)$ and $\mathrm{RMSE}(d)$, respectively) can be applied to reflect the model performance at the locations that have a distance of greater than d km to the closest monitoring station. Second, when a distance sequence (i.e., $d_1, d_2, \ldots, d_n$, where n is the number of distances) is utilized, it derives a performance curve, which is capable of obtaining a fuller evaluation of the $PM_{2.5}$ estimation model. Finally, for the quantitative evaluation of a specific region, an optimal distance (d_x) can be computed. The determination of the optimal distance d_x is described in Section III-B.

B. Determination of the Optimal Distance for a Specific Region

As shown in Fig. 4, each grid cell has a distance to the closest station (D_{grid}), and the average of D_{grid} is $\overline{D_{\mathrm{grid}}}$. Meanwhile, given a distance (d), each validation station has a distance to the closest modeling station ($D_{\mathrm{site}}, D_{\mathrm{site}} \geq d$), and after 10 rounds of validation, the average of D_{site} is $\overline{D_{\mathrm{site}}}$. In the model validation, the grid cell with a validation station is considered without the station and to be predicted (the prediction grid cell); and in the $PM_{2.5}$ spatial prediction, every grid cell is to be predicted, i.e., the prediction grid cell. When $\overline{D_{\mathrm{site}}} = \overline{D_{\mathrm{grid}}}$, the average of the distances from the prediction grid cells to their respective closest stations in the model validation is equal to that in the $PM_{2.5}$ spatial prediction for the region, which can

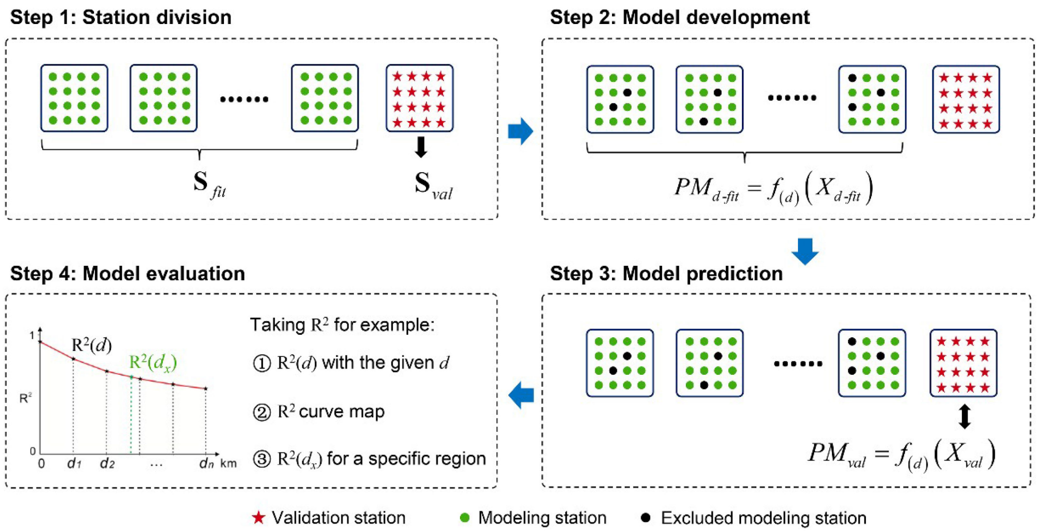


Fig. 3. Workflow of the proposed SDCV approach.

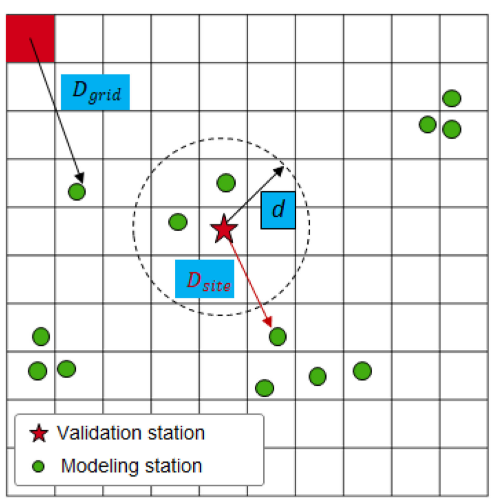


Fig. 4. Schematic for the related distance in a given region.

reflect the real estimation accuracy, to the greatest extent. Thus, the objective is to search for the optimal distance $d = d_x$, which ensures $\overline{D_{\text{site}}} = \overline{D_{\text{grid}}}$.

The searching process for d_x is as follows. First, we manually set an ordered sequence of distances $(d_1, d_2, \ldots, d_n$, e.g., 0–200 km with a step of 10 km), where n stands for the number of distances. For each distance $d_i (i = 1, 2, \ldots, n)$, the modeling station collection is updated, and we calculate $\overline{D_{\text{site}}}$ that can be written as $\overline{D^i_{\text{site}}}$. Meanwhile, $\overline{D^i_{\text{site}}} (i \geq 2)$ and $\overline{D_{\text{grid}}}$ are compared, when $\overline{D_{\text{grid}}} < \overline{D^i_{\text{site}}}$ and $\overline{D_{\text{grid}}} > \overline{D^{i-1}_{\text{site}}}$, the optimal distance d_x is interpolated within $[d_{i-1}, d_i]$.

IV. Validation of $PM_{2.5}$ Estimation Models: Case Study

A. Study Region and Data

To verify the validation approaches, a case study was conducted for the whole of China. The study region is shown in Fig. 5, where ∼1500 monitoring stations are located. As can be observed in Fig. 5, the monitoring stations are generally clustered in urban areas, and the station network exhibits a sparse distribution over a large range whereas a denser distribution over a small range. The study period in this article was the year of 2015. The annual averages of $PM_{2.5}$ values for each monitoring station were calculated.

The data used included four main parts, which are briefly described as follows.

1) Ground-level $PM_{2.5}$. We acquired hourly $PM_{2.5}$ data from the China National Environmental Monitoring Center (CNEMC) website.[1] The hourly $PM_{2.5}$ data were averaged to daily means.
2) Moderate Resolution Imaging Spectroradiometer (MODIS) AOD [38]–[40]. Terra and Aqua MODIS AOD products are widely used for ground $PM_{2.5}$ estimation, and were obtained from the Level 1 and Atmosphere Archive and Distribution System (LAADS).[2] The Collection 6 AOD product was used, which has a spatial resolution of 10 km.
3) Meteorological variables. Surface pressure (unit: Pa), wind speed at 10 m above ground (unit: m/s), relative humidity (unit: %), air temperature at a 2-m height (unit: K), and planetary boundary layer height (unit: m) were extracted from the second Modern-Era Retrospective Analysis for Research and Applications (MERRA-2) reanalysis data.[3]
4) The MODIS normalized difference vegetation index product (MOD13) was also obtained from the LAADS website. For full details, please refer to our previous study [25].

B. $PM_{2.5}$ Estimation Models

The $PM_{2.5}$ estimation models used in the analysis were spatial interpolation (inverse distance weighted interpolation), the linear mixed effect (LME) model [17], [24], daily geographically weighted regression (GWR) [18], [37], geographically

[1]http://106.37.208.233:20035/
[2]https://ladsweb.modaps.eosdis.nasa.gov/
[3]http://gmao.gsfc.nasa.gov/GMAO_products/

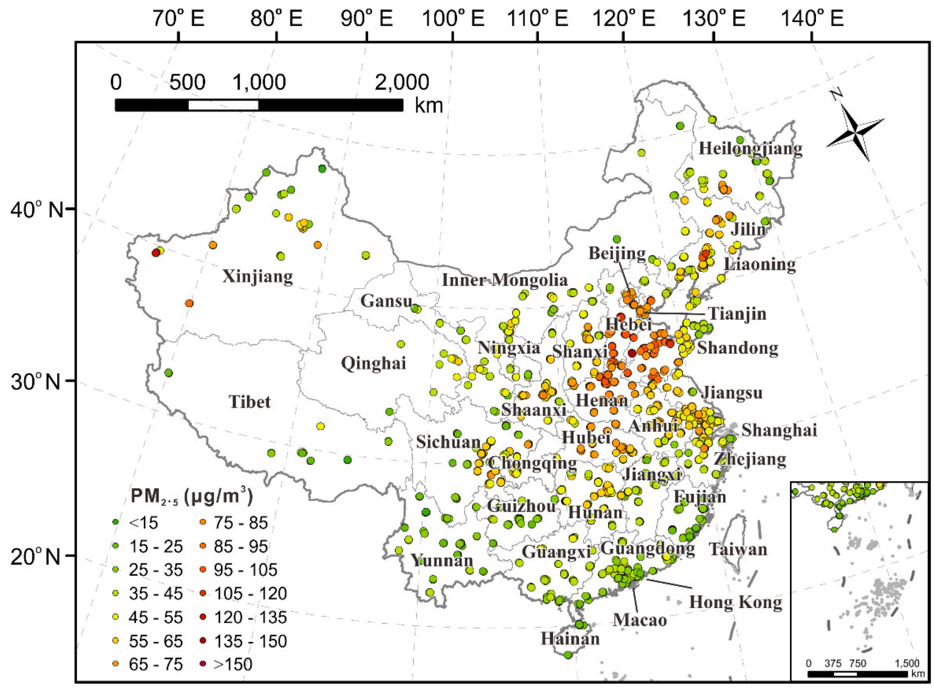


Fig. 5. Study domain and the annual mean spatial distribution of $PM_{2.5}$ monitoring stations in China.

TABLE II
RESULTS WITH THE PREVIOUS VALIDATION APPROACHES

Model	Sample-based CV		Site-based CV		Region-based CV	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
Interpolation	0.83	15.99	0.83	15.79	0.46	28.89
LME	0.55	25.76	0.55	25.80	0.49	27.35
GWR	0.72	20.46	0.72	20.54	0.50	27.54
GTWR	0.75	19.53	0.73	20.26	0.53	26.59
Geoi-DBN	0.88	13.03	0.84	15.39	0.54	26.10

and temporally weighted regression (GTWR) [41], [42], and the geo-intelligent deep belief network (Geoi-DBN) [25].

The reasons for the selection of these models were as follows. First, spatial interpolation is one of the simplest methods without the use of remote sensing data, and it can be considered as a baseline for the comparison with the remote sensing methods. Second, LME, daily GWR, and GTWR are widely used for satellite-based estimation of ground $PM_{2.5}$. Wherein, LME often takes the temporal heterogeneity of the AOD-$PM_{2.5}$ relationship into account, whereas it is a global model in space. This means that the LME model may be less sensitive to the distance to the monitoring station. In contrast, GWR uses a spatially local regression technique, which is easily influenced by the spatial distance to the modeling station. The GTWR model is a further development of GWR, with the incorporation of temporal dependency. The comparison between the GWR and GTWR model validation results can manifest the sensitivity to temporal information for the validation approaches. Finally, the Geoi-DBN model considers the spatiotemporal autocorrelation of $PM_{2.5}$, and has been reported to achieve a state-of-the-art estimation performance. In summary, LME is a global spatial model, and the others are distance-dependent models.

C. Results and Analysis

First, the abovementioned $PM_{2.5}$ estimation models were evaluated by the previous validation approaches (as listed in Table II). Time-based CV and historical validation were not carried out, because: 1) the abovementioned $PM_{2.5}$ estimation models, which rely on the station-based $PM_{2.5}$ measurements of the estimation time for the model establishment are, in principle, unable to be evaluated by these two validation approaches; and 2) in this article, we pay more attention to the spatial mapping of $PM_{2.5}$.

As shown in Table II, from the sample-based CV to the site-based CV, the spatial interpolation, LME, and GWR models report a similar performance, indicating that these models have a comparable spatial prediction ability with overall prediction ability. The explanation for this is that the spatial interpolation and GWR models are separately established for individual days,

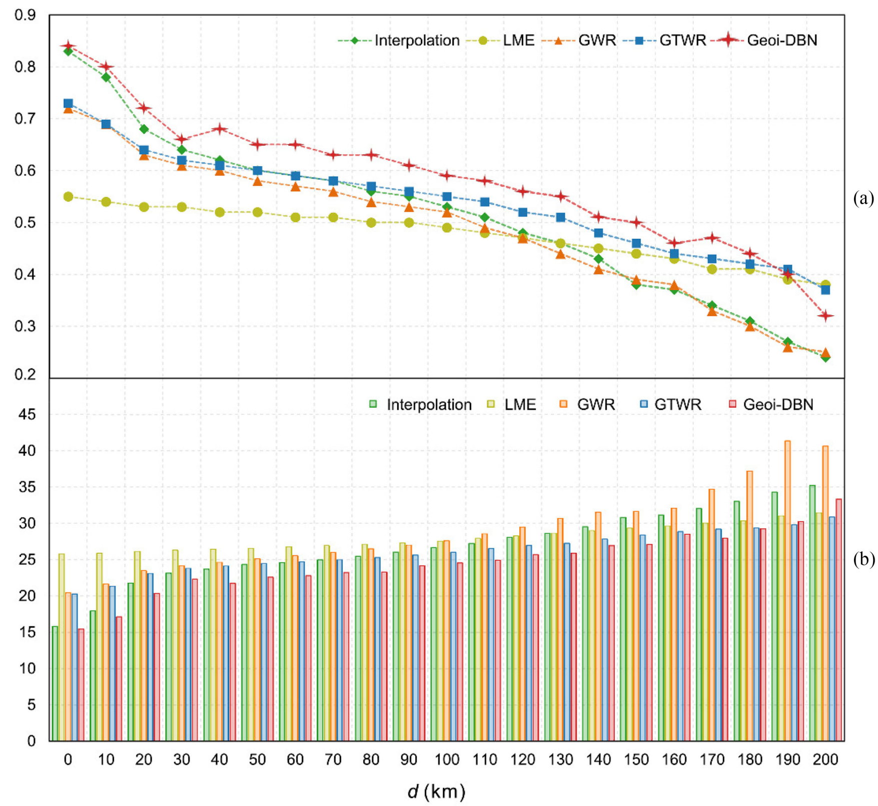
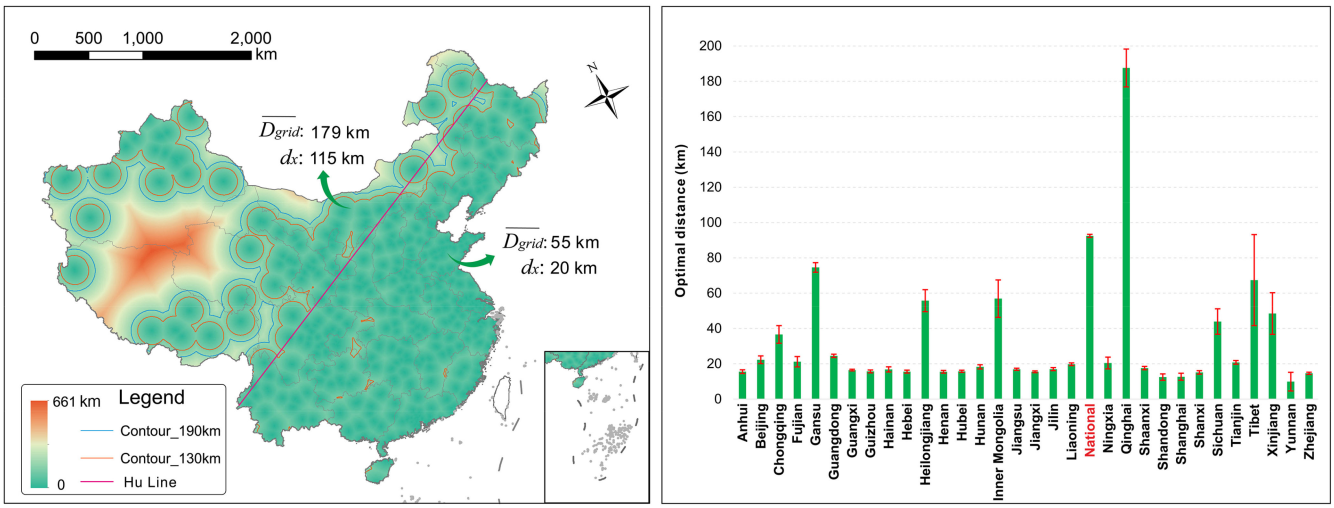


Fig. 6. Validation results with the proposed SDCV approach. (a) R^2. (b) RMSE.

and the temporal dependency (used in the sample-based CV but not in the site-based CV) is not incorporated. Furthermore, the temporal dependency may not have great benefits for the time-specific LME model. However, with the use of temporal information, the GTWR model shows a significant decrease from the sample-based CV ($R^2 = 0.75$) to the site-based CV ($R^2 = 0.73$), and the Geoi-DBN model reports a similar trend.

More interestingly, it is surprising to find that the spatial interpolation method achieves a much better result than the widely used LME, GWR, and GTWR models in the sample-based CV and the site-based CV. The reason for this could be that the validation stations are very close to the modeling stations. The findings indicate that due to the closeness of the validation stations and modeling stations, the simple spatial interpolation method tends to outperform the widely used LME, GWR, and GTWR remote sensing models. However, how well the $PM_{2.5}$ estimation models perform at locations with a greater distance to the modeling station needs a more complete evaluation. For the region-based CV (which is province-based CV here), all the models, except for the LME model, report a great decrease in performance compared to the sample-based CV and the site-based CV, indicating that the performance of the distance-dependent estimation models is liable to be overestimated by the widely used sample-based CV and site-based CV techniques. Moreover, the spatial interpolation method performs worse than the LME, GWR, and GTWR models, which is contrary to the results for the sample-based CV and the site-based CV. The results show that the remote sensing methods show some advantages than the spatial interpolation when one province of stations is left for model validation. Nevertheless, it is unreasonable to leave one province of stations for the model validation, because it is too strict compared with the real situation.

Second, to evaluate the abovementioned $PM_{2.5}$ estimation models via the proposed SDCV approach, we set distance d within the bounds of 0–200 km and a step size of 10 km. As can be observed in Fig. 6, generally speaking, all the models report a downward trend in performance with the increasing distance. First, for the spatial interpolation method, it obtains a superior result when the distance is 0 km (i.e., site-based CV). The performance of the spatial interpolation method then decreases dramatically from $d = 0$ km to $d = 30$ km, with the R^2 values falling from 0.83 to 0.64. The reason for this could be that the spatial interpolation method is strongly dependent on the nearby monitoring stations. This is, then, followed by the LME model, which exhibits a tardy decreasing trend in model performance, with R^2 values of 0.55 for 0 km and 0.38 for 200 km. The possible reason for this is that LME is a global spatial model, and is much less sensitive to the spatial distance. Subsequently, a notable decrease can be observed in the GWR model performance, especially between 0–30 km. This can be attributed to the fact that the GWR model establishes the AOD-$PM_{2.5}$ relationship through a spatially local regression technique, and is greatly influenced by the distance to the modeling station. As with the GWR model, the GTWR model reports a consistent decreasing trend. Finally, the Geoi-DBN model shows a similar decreasing pattern to the spatial interpolation method, for the reason that Geoi-DBN incorporates the spatial autocorrelation of $PM_{2.5}$.



(a) (b)

Fig. 7. Distance analysis for China. (a) Distribution of distance to the closest station in China. (b) Optimal distance for the various provinces. Hong Kong, Macau, and Taiwan are excluded from the analysis due to the lack of stations. Red error bar refers to the standard deviation of the multiple experiments.

Furthermore, interesting findings about the comparisons between the various models are revealed in Fig. 6. When the distance is set as 0 km (i.e., site-based CV), among the above-mentioned models, Geoi-DBN yields the best performance. The spatial interpolation method is a close second, with an R^2 value being 0.83 and RMSE being 15.79 μg/m^3, respectively, and it notably outperforms the LME, GWR, and GTWR models. The results indicate that the spatial interpolation method performs better than the satellite-based estimation methods (LME, GWR, and GTWR) based on conventional site-based CV. However, as the distance increases, the LME/GTWR models surpass the spatial interpolation method when the distance is $\sim$130/$\sim$70 km. As for the Geoi-DBN model, it incorporates the spatial autocorrelation of $PM_{2.5}$, and consequently performs better than the spatial interpolation method at all the distances. Nevertheless, it should be noted that the LME and GTWR models show some superiorities over Geoi-DBN when the distance is $\sim$190 km. All these results indicate that the proposed SDCV approach is able to obtain a more complete evaluation for the $PM_{2.5}$ estimation models.

D. Discussion

Compared with the previous validation approaches, the proposed SDCV technique shows an improvement by considering the unevenness of the station distribution. Although SDCV does not allow for spatial heterogeneity, and extending the validation law obtained from the monitoring stations to the whole space still has limitations, we attempted to give the optimal distance within China for consideration. Fig. 7(a) shows the distribution of the distance to the closest station in China, which has a mean value of 134 km and a range of 0 to 661 km. In consideration of the randomness in the first step of SDCV, 10 repetitive experiments were conducted to determinate the optimal distance, and the mean result with $d_x = 92$ km was used to evaluate the $PM_{2.5}$ estimation models for the whole of China. For the model performance with this optimal distance, we refer to the results at $d = 90$ km in Fig. 6, where the Geoi-DBN model achieves the best performance ($R^2 = 0.61$), and LME performs the worst. There is also a great variation of mean distance (179 and 55 km, respectively) on the two sides of the Hu line (a.k.a. the Heihe–Tengchong line) [43], [44], which indicates the notably different distance being suggested for the two sides (115 km and 20 km, respectively). In addition, the optimal distances (d_x) for the various provinces in China were calculated and are shown in Fig. 7(b). Because of the sparse and uneven distribution of monitoring stations, Qinghai, Tibet, Xinjiang, Gansu, Inner Mongolia, and Heilongjiang report larger optimal distances, and they also have relatively higher standard deviations, which indicates higher instability in the determination of the optimal distance.

On the other hand, we also sought to compare the above-mentioned $PM_{2.5}$ estimation models in terms of the distance distribution in Fig. 7(a). For instance, middle China and eastern China almost have a distance to the closest station of less than 130 km [Contour_130 km in Fig. 7(a)], indicating that the spatial interpolation information may be more effective for $PM_{2.5}$ estimation compared with LME in these regions. Meanwhile, the Geoi-DBN model appears better suited to address the $PM_{2.5}$ estimation in most other parts of China (distance < 190 km, see Contour_190 km), but it is likely to encounter more challenges in the northwest of China, compared with LME and GTWR.

Due to the high cost of $PM_{2.5}$ ground stations, the station network exhibits a sparse and uneven spatial distribution, which brings challenges to the validation of the satellite-based $PM_{2.5}$ estimation models. With the development of $PM_{2.5}$ monitors, the intensive observation network [6], [34], [45], [46] may provide new solutions for the validation of $PM_{2.5}$ estimation models. First, portable low-cost devices and observation vehicles are likely to offer more validation approaches for $PM_{2.5}$ estimation models. For instance, the $PM_{2.5}$ estimation model is established by station observations, and the accuracy of the model estimates can be evaluated using portable devices and/or observation vehicles. Second, if the station network is to be continuously updated, the intensive station network will make the validation of $PM_{2.5}$ estimation models easier. In short, the intensive observations and abundance of observations will provide new perspectives for the validation of $PM_{2.5}$ estimation models.

Another issue that needs to be considered is the scale effect. The satellite-based AOD data used have a spatial resolution of 0.1°, indicating that a grid cell (also known as a pixel) represents an $\sim$10 km observation of AOD data. Meanwhile, the station-based $PM_{2.5}$ data are point-based measurements. Therefore, the spatial scale of the satellite AOD does not match that of the station-based $PM_{2.5}$ measurements. As a previous study [34] indicated that the measurements at a ground surface site observation are often representative of an area around 0.5–16 km^2. Accordingly, scale variation exists in the AOD-based $PM_{2.5}$ estimates and station-based $PM_{2.5}$ measurements. In the AOD-$PM_{2.5}$ study field, station $PM_{2.5}$ measurements are often used to represent a region (e.g., $\sim$10 $\times$10 km here) of $PM_{2.5}$. As a result, the point-based $PM_{2.5}$ measured from one (or multiple) stations are adopted to evaluate the accuracy of the $PM_{2.5}$ estimates in one grid cell. Whether it is reasonable to evaluate the grid cell estimates using station measurements deserves further study.

V. Conclusion

To sum up, several different validation approaches are used for satellite-based $PM_{2.5}$ estimation models, but their applicable conditions remain unclear. Hence, one important contribution of this study is that we fully analyzed and assessed the existing validation approaches, and gave some suggestions as to their applicable conditions. Among the existing validation approaches, sample-based CV can be used to reflect the overall predictive ability; site-based CV and region-based CV have the potential to evaluate spatial prediction performance; and time-based CV and historical validation are more suitable to evaluate temporal prediction accuracy. Furthermore, the existing validation approaches do not consider the uneven distribution of monitoring stations, which may bring some bias for the evaluation of $PM_{2.5}$ estimation models. A CV-based validation approach considering the uneven SDCV was proposed. The results indicated that SDCV can obtain a more complete and effective evaluation for the popular $PM_{2.5}$ estimation models than the traditional validation approaches. In summary, this study will provide application implications and new perspectives for the validation of satellite $PM_{2.5}$ estimation models.

Acknowledgment

The authors would like to thank the CNEMC, LAADS, and MERRA-2 team for providing ground $PM_{2.5}$ data, satellite products, and meteorological data, respectively, and also like to thank the editors and anonymous reviewers for their valuable suggestions.

References

[1] WHO, *Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide*: World Health Organization, Geneva, Switzerland, 2006.

[2] J. A. Engel-Cox, N. T. Kim Oanh, A. van Donkelaar, R. V. Martin, and E. Zell, “Toward the next generation of air quality monitoring: Particulate matter,” *Atmospheric Environ.*, vol. 80, pp. 584–590, 2013.

[3] J. Peng, S. Chen, H. Lü, Y. Liu, and J. Wu, “Spatiotemporal patterns of remotely sensed $PM_{2.5}$ concentration in China from 1999 to 2011,” *Remote Sens. Environ.*, vol. 174, pp. 109–121, 2016.

[4] Y. Liu, B. A. Schichtel, and P. Koutrakis, “Estimating particle sulfate concentrations using MISR retrieved aerosol properties,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 2, no. 3, pp. 176–184, Sep. 2009.

[5] H. Fan, C. Zhao, and Y. Yang, “A comprehensive analysis of the spatio-temporal variation of urban air pollution in China during 2014–2018,” *Atmospheric Environ.*, vol. 220, 2020, Art. no. 117066.

[6] C. Zhao *et al.*, “Estimating the contribution of local primary emissions to particulate pollution using high-density station observations,” *J. Geophysical Res., Atmospheres*, vol. 124, pp. 1648–1661, 2019.

[7] Y. Yang *et al.*, “$PM_{2.5}$ pollution modulates wintertime urban heat island intensity in the Beijing-Tianjin-Hebei megalopolis, China,” *Geophysical Res. Lett.*, vol. 47, 2020, Art. no. e2019GL084288.

[8] H. Wang *et al.*, “High-spatial-resolution population exposure to $PM_{2.5}$ pollution based on multi-satellite retrievals: A case study of seasonal variation in the yangtze river delta, China in 2013,” *Remote Sens.*, vol. 11, p. 2724, 2019.

[9] X. Hu *et al.*, “Estimating $PM_{2.5}$ concentrations in the conterminous united states using the random forest approach,” *Environ. Sci. Technol.*, vol. 51, pp. 6936–6944, 2017.

[10] D. J. Lary *et al.*, “Estimating the global abundance of ground level presence of particulate matter ($PM_{2.5}$),” *Geospatial health*, vol. 8, pp. S611–30, 2014.

[11] R. M. Hoff and S. A. Christopher, “Remote sensing of particulate pollution from space: Have we reached the promised land?” *J. Air Waste Manage. Assoc.*, vol. 59, pp. 645–675, 2009.

[12] R. V. Martin, “Satellite remote sensing of surface air quality,” *Atmospheric Environ.*, vol. 42, pp. 7823–7843, 2008.

[13] C. Zheng *et al.*, “Analysis of influential factors for the relationship between $PM_{2.5}$ and AOD in Beijing,” *Atmos. Chem. Phys.*, vol. 17, pp. 13473–13489, 2017.

[14] H. Shen, T. Li, Q. Yuan, and L. Zhang, “Estimating regional ground-level $PM_{2.5}$ directly from satellite top-of-atmosphere reflectance using deep belief networks,” *J. Geophysical Res., Atmospheres*, vol. 123, pp. 13,875–13,886, 2018.

[15] P. Gupta and S. A. Christopher, “Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach,” *J. Geophysical Res., Atmospheres*, vol. 114, 2009, Art. no. D14205.

[16] J. Tian and D. Chen, “A semi-empirical model for predicting hourly ground-level fine particulate matter ($PM_{2.5}$) concentration in southern Ontario from satellite remote sensing and ground-based meteorological measurements,” *Remote Sens. Environ.*, vol. 114, pp. 221–229, 2010.

[17] H. Lee, Y. Liu, B. Coull, J. Schwartz, and P. Koutrakis, “A novel calibration approach of MODIS AOD data to predict $PM_{2.5}$ concentrations,” *Atmospheric Chemistry Phys.*, vol. 11, pp. 7991–8002, 2011.

[18] X. Hu *et al.*, “Estimating ground-level $PM_{2.5}$ concentrations in the southeastern U.S. using geographically weighted regression,” *Environ. Res.*, vol. 121, pp. 1–10, 2013.

[19] B. Zou, Q. Pu, M. Bilal, Q. Weng, L. Zhai, and J. E. Nichol, “High-resolution satellite mapping of fine particulates based on geographically weighted regression,” *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 4, pp. 495–499, Apr. 2016.

[20] P. Gupta and S. A. Christopher, “Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach,” *J. Geophysical Res., Atmospheres*, vol. 114, 2009, Art. no. D20205.

[21] T. Li, H. Shen, C. Zeng, Q. Yuan, and L. Zhang, “Point-surface fusion of station measurements and satellite observations for mapping $PM_{2.5}$ distribution in China: Methods and assessment,” *Atmospheric Environ.*, vol. 152, pp. 477–489, 2017.

[22] Y. Sun, Q. Zeng, B. Geng, X. Lin, B. Sude, and L. Chen, “Deep learning architecture for estimating hourly ground-level $PM_{2.5}$ using satellite remote sensing,” *IEEE Geosci. Remote Sens. Lett.*, vo. 16, no. 9, pp. 1343–1347, Sep. 2019.

[23] J. D. Rodriguez, A. Perez, and J. A. Lozano, "Sensitivity analysis of k-fold cross validation in prediction error estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 569–575, Mar. 2010.

[24] Y. Xie, Y. Wang, K. Zhang, W. Dong, B. Lv, and Y. Bai, "Daily estimation of ground-level $PM_{2.5}$ concentrations over Beijing using 3 km resolution MODIS AOD," *Environ. Sci. Technol.*, vol. 49, pp. 12280–12288, 2015.

[25] T. Li, H. Shen, Q. Yuan, X. Zhang, and L. Zhang, "Estimating ground-level $PM_{2.5}$ by fusing satellite and station observations: A geo-intelligent deep learning approach," *Geophysical Res. Lett.*, vol. 44, pp. 11985–11993, 2017.

[26] M. D. Friberg, R. A. Kahn, J. A. Limbacher, K. W. Appel, and J. A. Mulholland, "Constraining chemical transport $PM_{2.5}$ modeling outputs using surface monitor measurements and satellite retrievals: Application over the San Joaquin Valley," *Atmospheric Chemistry Phys.*, vol. 18, pp. 12891–12913, 2018.

[27] Z. Ma, Y. Liu, Q. Zhao, M. Liu, Y. Zhou, and J. Bi, "Satellite-derived high resolution $PM_{2.5}$ concentrations in Yangtze River Delta Region of China using improved linear mixed effects model," *Atmospheric Environ.*, vol. 133, pp. 156–164, 2016.

[28] T. Zhang *et al.*, "Estimation of ultrahigh resolution $PM_{2.5}$ concentrations in urban areas using 160 m Gaofen-1 AOD retrievals," *Remote Sens. Environ.*, vol. 216, pp. 91–104, 2018.

[29] Z. Ma *et al.*, "Satellite-based spatiotemporal trends in $PM_{2.5}$ concentrations: China, 2004-2013," *Environ Health Perspect*, vol. 124, pp. 184–92, 2016.

[30] K. Huang *et al.*, "Predicting monthly high-resolution $PM_{2.5}$ concentrations with random forest model in the North China plain," *Environ. Pollut.*, vol. 242, pp. 675–683, 2018.

[31] Z. Ma, X. Hu, L. Huang, J. Bi, and Y. Liu, "Estimating ground-level $PM_{2.5}$ in China using satellite remote sensing," *Environ. Sci. Technol.*, vol. 48, pp. 7436–7444, 2014.

[32] W. You, Z. Zang, L. Zhang, Z. Li, D. Chen, and G. Zhang, "Estimating ground-level PM10 concentration in northwestern China using geographically weighted regression based on satellite AOD combined with CALIPSO and MODIS fire count," *Remote Sens. Environ.*, vol. 168, pp. 276–285, 2015.

[33] X. Hu *et al.*, "Estimating ground-level $PM_{2.5}$ concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model," *Remote Sens. Environ.*, vol. 140, pp. 220–232, 2014.

[34] X. Shi, C. Zhao, J. H. Jiang, C. Wang, X. Yang, and Y. L. Yung, "Spatial representativeness of $PM_{2.5}$ concentrations obtained using observations from network stations," *J. Geophysical Res.: Atmospheres*, vol. 123, pp. 3145–3158, 2018.

[35] Y. Zhan *et al.*, "Spatiotemporal prediction of continuous daily $PM_{2.5}$ concentrations across China using a spatially explicit machine learning algorithm," *Atmospheric Environ.*, vol. 155, pp. 129–139, 2017.

[36] X. Fang, B. Zou, X. Liu, T. Sternberg, and L. Zhai, "Satellite-based ground $PM_{2.5}$ estimation using timely structure adaptive modeling," *Remote Sens. Environ.*, vol. 186, pp. 152–163, 2016.

[37] W. Song, H. Jia, J. Huang, and Y. Zhang, "A satellite-based geographically weighted regression model for regional $PM_{2.5}$ estimation over the pearl river delta region in China," *Remote Sens. Environ.*, vol. 154, pp. 1–7, 2014.

[38] Y. Wang, Q. Yuan, T. Li, H. Shen, L. Zheng, and L. Zhang, "Evaluation and comparison of MODIS Collection 6.1 aerosol optical depth against AERONET over regions in China with multifarious underlying surfaces," *Atmospheric Environ.*, vol. 200, pp. 280–301, 2019.

[39] Y. Wang, Q. Yuan, T. Li, H. Shen, L. Zheng, and L. Zhang, "Large-scale MODIS AOD products recovery: Spatial-temporal hybrid fusion considering aerosol variation mitigation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 157, pp. 1–12, 2019.

[40] N. Liu, B. Zou, H. Feng, W. Wang, Y. Tang, and Y. Liang, "Evaluation and comparison of multiangle implementation of the atmospheric correction algorithm, dark target, and deep blue aerosol products over China," *Atmospheric Chemistry Phys.*, vol. 19, pp. 8243–8268, 2019.

[41] Q. He and B. Huang, "Satellite-based high-resolution $PM_{2.5}$ estimation over the Beijing-Tianjin-Hebei region of China using an improved geographically and temporally weighted regression model," *Environ. Pollut.*, vol. 236, pp. 1027–1037, 2018.

[42] Y. Bai, L. Wu, K. Qin, Y. Zhang, Y. Shen, and Y. Zhou, "A geographically and temporally weighted regression model for ground-level $PM_{2.5}$ estimation from satellite-derived 500 m resolution AOD," *Remote Sens.*, vol. 8, p. 262, 2016.

[43] Z. Feng, F. Li, Y. Yang, and P. Li, "The past, present, and future of population geography in China: Progress, challenges and opportunities," *J. Geographical Sci.*, vol. 27, pp. 925–942, 2017.

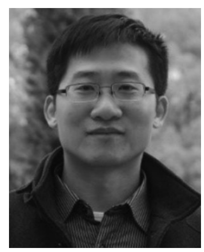
[44] H. Hu, "The distribution, regionalization and prospect of China's population," *Acta Geographica Sinica*, vol. 2, pp. 139–145, 1990.

[45] S. Xu, B. Zou, Y. Lin, X. Zhao, S. Li, and C. Hu, "Strategies of method selection for fine scale $PM_{2.5}$ mapping in intra-urban area under crowdsourcing monitoring," *Atmospheric Meas. Techn.*, pp. 1–21, 2019.

[46] J. Bi *et al.*, "Contribution of low-cost sensor measurements to the prediction of $PM_{2.5}$ levels: A case study in imperial County, California, USA," *Environ. Res.*, vol. 180, 2019, Art. no. 108810.

Tongwen Li received the B.S degree in geographical information system (GIS) from the School of Resource and Environmental Sciences, Wuhan University, Wuhan, China, in 2011, where he is currently working toward the Ph.D. degree in cartography and geographic information engineering.

His research interests include atmospheric $PM_{2.5}$ remote sensing and machine learning in environmental remote sensing.



Huanfeng Shen (Senior Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively.

In 2007, he was with the School of Resource and Environmental Sciences, Wuhan University, where he is currently a Luojia Distinguished Professor. He has been supported by several talent programs, such as the Youth Talent Support Program of China in 2015, the China National Science Fund for Excellent Young Scholars in 2014, and the New Century Excellent Talents by the Ministry of Education of China in 2011. He has authored and coauthored more than 100 research papers. His research interests include image quality improvement, remote sensing mapping and application, data fusion and assimilation, and regional and global environmental changes.

Dr. Shen is currently a member of the Editorial Board of the *Journal of Applied Remote Sensing*.

Chao Zeng received the B.S. degree in resources-environment and urban-rural planning management, the M.S. degree in surveying and mapping engineering, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2009, 2011, and 2014, respectively.

He was a Postdoctoral Researcher with the Department of Hydraulic Engineering, Tsinghua University, Beijing, China. He is currently with the School of Resource and Environmental Sciences, Wuhan University. His current research interests include remote sensing image processing and hydrological remote sensing applications.

Qiangqiang Yuan (Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively.

In 2012, he was with the School of Geodesy and Geomatics, Wuhan University, where he is currently a Full Professor. He has authored and coauthored more than 50 research papers, including more than 30 peer-reviewed articles in international journals such as the IEEE TRANSACTIONS IMAGE PROCESSING and the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. His current research interests include image reconstruction, remote sensing image processing and application, and data fusion.

Dr. Yuan was the recipient of the Top-Ten Academic Star of Wuhan University in 2011. In 2014, he received the Hong Kong Scholar Award from the Society of Hong Kong Scholars and the China National Postdoctoral Council. He has frequently served as a Referee for more than 20 international journals for remote sensing and image processing.