Earth Syst. Sci. Data, 16, 3781–3793, 2024
https://doi.org/10.5194/essd-16-3781-2024


Open Access Earth System Science Data

# Retrieving ground-level $PM_{2.5}$ concentrations in China (2013–2021) with a numerical-model-informed testbed to mitigate sample-imbalance-induced biases

**Siwei Li[1,3,4], Yu Ding[1], Jia Xing[2], and Joshua S. Fu[2]**

[1]Hubei Key Laboratory of Quantitative Remote Sensing of Land and Atmosphere, School of Remote Sensing and Information Engineering, Wuhan University, Hubei 430000, China
[2]Department of Civil and Environmental Engineering, the University of Tennessee, Knoxville, TN 37996, USA
[3]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[4]Hubei Luojia Laboratory, Wuhan University, Wuhan 430079, China

**Correspondence:** Siwei Li (siwei.li@whu.edu.cn) and Jia Xing (jxing3@utk.edu)

Received: 7 May 2024 – Discussion started: 15 May 2024
Revised: 5 July 2024 – Accepted: 11 July 2024 – Published: 27 August 2024

**Abstract.** Ground-level $PM_{2.5}$ data derived from satellites with machine learning are crucial for health and climate assessments. However, uncertainties persist due to the absence of spatially covered observations. To address this, we propose a novel testbed using nontraditional numerical simulations to evaluate $PM_{2.5}$ estimation across the entire spatial domain. The testbed emulates the general machine-learning approach by training the model with grids corresponding to ground monitoring sites and subsequently testing its predictive accuracy for other locations. Our approach enables comprehensive evaluation of various machine-learning methods' performance in estimating $PM_{2.5}$ across the spatial domain for the first time. Unexpected results are shown in the application in China, with larger absolute $PM_{2.5}$ biases found in densely populated regions with abundant ground observations across all benchmark models due to the higher baseline concentration, though the relative error (approximately 20 %) is smaller compared to that in rural areas (over 50 %). The imbalance in training samples, mostly from urban areas with high emissions, is the main reason, leading to significant overestimation due to the lack of monitors in downwind areas where $PM_{2.5}$ is transported from urban areas with varying vertical profiles. Our proposed testbed also provides an efficient strategy for optimizing model structure or training samples to enhance satellite-retrieval model performance. Integration of spatiotemporal features, especially with conventional neural network (CNN)-based deep-learning approaches like the residual neural network (ResNet) model, has successfully mitigated $PM_{2.5}$ overestimation (by 5–30 $\mu g\,m^{-3}$) and the corresponding exposure (by 3 million people $\cdot\,\mu g\,m^{-3}$) in the downwind area over 9 years (2013–2021) compared to the traditional approach. Furthermore, the incorporation of 600 strategically positioned ground monitoring sites identified through the testbed is essential for achieving a more balanced distribution of training samples, thereby ensuring precise $PM_{2.5}$ estimation and facilitating the assessment of the associated impacts in China. In addition to presenting the retrieved surface $PM_{2.5}$ concentrations in China from 2013 to 2021, this study provides a testbed dataset derived from physical modeling simulations which can serve to evaluate the performance of data-driven methodologies, such as machine learning, in estimating spatial $PM_{2.5}$ concentrations for the community (Li et al., 2024a; https://doi.org/10.5281/zenodo.11122294).

Published by Copernicus Publications.

## 1 Introduction

Accurate knowledge of $PM_{2.5}$ pollution is vital for understanding its impact on human health (Lelieveld et al., 2015; Geng et al., 2021) and the climate (Mitchell et al., 1995; Bellouin et al., 2005). Satellite products provide direct measurements of aerosol loading on broad spatial and temporal scales. While the aerosol optical depth (AOD) measured by satellites reflects the total column of particulate matter, this is challenged by the complex relationship between AOD and ground $PM_{2.5}$ influenced by various factors (Hoff and Christopher, 2009), including aerosol chemical composition and vertical profiles. Compared to traditional statistics, machine learning excels in addressing nonlinearities. Therefore, numerous recent studies leverage machine learning, such as in the random forest (RF) (Hu et al., 2017), XGBoost (Xiao et al., 2018), lightGBM (Zhong et al., 2021), and deep-learning (Li et al., 2020; Yan et al., 2020; Wang et al., 2022a, b; Wei et al., 2023) models, to establish correlations between AOD and $PM_{2.5}$, treating AOD and related factors, including meteorological variables, as features for predicting surface $PM_{2.5}$ based on ground measurements (Ma et al., 2022). However, a limitation arises as most ground measurements are concentrated in urban and polluted areas. Their main purpose is to monitor the high pollution levels to protect human health, leading to an uneven spatial distribution. It is expected that training models predominantly on urban sites will introduce an imbalance in ground-based measurements, resulting in significant uncertainties in spatially allocating surface $PM_{2.5}$ based on satellite AOD (Shin et al., 2020). This deficiency might be particularly notable in suburban areas experiencing downwind transport of $PM_{2.5}$ from urban areas (Bai et al., 2022). The discrepancy between urban and downwind sites largely lies in their vertical profiles of aerosol across the entire vertical layer. Urban sites, which have abundant emission sources such as residential areas, transportation, construction, and industries, exhibit a higher share of ground-level aerosol relative to the total AOD compared to downwind and rural areas, where pollution tends to be lifted to upper layers through atmospheric dynamics. Accurately representing the varying aerosol vertical profiles in source/urban and downwind/rural areas is crucial for retrieving ground-level $PM_{2.5}$ from the AOD. However, imbalanced training samples make the machine-learning model unable to adequately capture such variations. The traditional cross-validation methods based either on samples or sites (Dong et al., 2020), which still rely mostly on samples available in urban sites, fail to provide a comprehensive assessment of model performance across the entire prediction space. Consequently, uncertainties in $PM_{2.5}$ estimation for these areas remain unexplored, and solutions to reduce such uncertainties are yet to be developed.

To overcome these limitations, we introduce a novel testbed utilizing a numerical model, specifically a chemical transport model (CTM) such as the Community Multiscale Air Quality Modeling System developed by the U.S. Environmental Protection Agency (EPA) and its community (Appel et al., 2013), to establish ground-truth data beyond monitoring points. This allows for the evaluation of interpolation performance using various machine-learning models and provides solutions to mitigate the uncertainties stemming from the sample imbalance problem. More specifically, we emulate traditional machine-learning methods by using CTM-simulated $PM_{2.5}$ concentrations in grid cells corresponding to ground monitoring sites as labels for training machine-learning models. Subsequently, we validate the trained model's performance in predicting $PM_{2.5}$ concentrations in other grid cells. In addition to providing a "ground truth" for assessing performance across the entire space, the CTM-simulated data act as a testbed for efficiently seeking solutions to enhance satellite-retrieval model performance. This involves optimizing features, model structures, and training samples, as depicted in Fig. 1.

## 2 Methods

The proposed testbed is implemented in a Chinese domain, utilizing 1 whole year's simulation for 2017 with a 27 km by 27 km resolution. To ensure internal consistency during training, all the feature and label data are derived from the input and output of the commonly used Community Multiscale Air Quality (CMAQ) model. The meteorological variables include $U$ wind, $V$ wind, humidity, 2 m temperature, the convective velocity scale, shortwave radiation, 10 m wind speed, planetary boundary layer (PBL) height, leaf area index (LAI), cloud fraction, and precipitation and are simulated by the Weather Research and Forecasting (WRF) model (Skamarock et al., 2008). The simulated AOD is calculated based on simulated $PM_{2.5}$ chemical compositions and corresponding meteorological variables across all vertical layers (Liu et al., 2010). We also include the $NO_2$ column density as an important feature as it is highly correlated with emission sources and can be directly observed from satellites to better represent the emission information (Martin et al., 2003). Similarly, the simulated $NO_2$ column density is calculated based on simulated $NO_2$ concentrations and corresponding meteorological variables across all vertical layers.

In addition, we conduct model training using 9-year observational data from 2013 to 2021 to evaluate potential biases under real-world conditions. This is quantified by measuring the difference in retrieved $PM_{2.5}$ between the traditional model and the improved retrieval method optimized with the proposed testbed. The dataset for this evaluation comprises MODerate-resolution Imaging Spectroradiometer (MODIS) (Remer et al., 2008) satellite observations for AOD, Ozone Monitoring Instrument (OMI) (Celarier et al., 2008) satellite data for $NO_2$ column density, and ground monitoring observations of $PM_{2.5}$ from the China National Environmental Monitoring Center (CNEMC, covering a total of more than

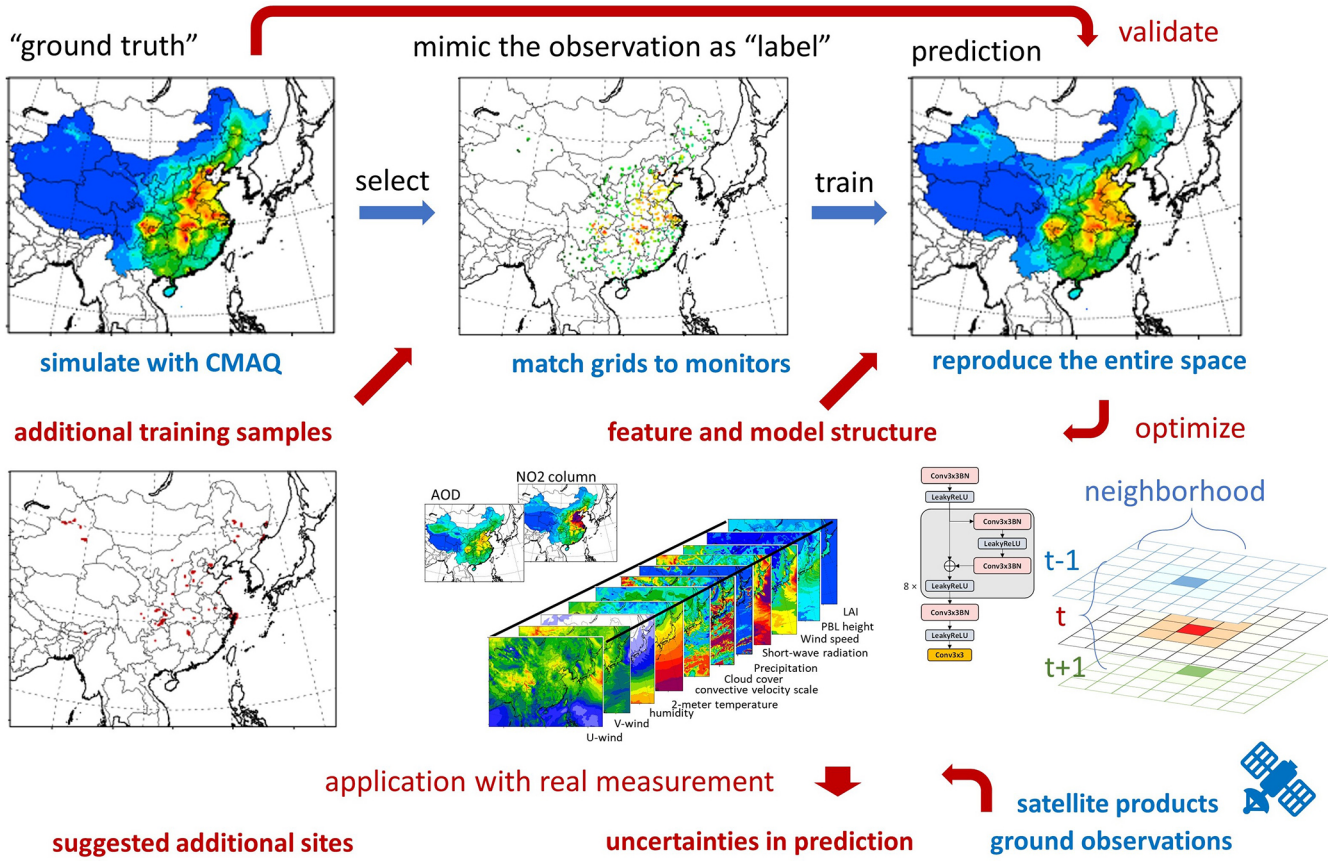


**Figure 1.** The proposed testbed for evaluation of satellite-retrieval surface $PM_{2.5}$ concentration.

600 grid cells of 27 km by 27 km) (Kong et al., 2021). Following the same data-filling method (He et al., 2020), we conduct data filling for the satellite measurement of $NO_2$ column density and AOD when applying our approach to real data. The generation of testbed data and the machine-learning methods are detailed as follows.

## 2.1 WRF and CMAQ numerical models

In this study, we utilized version 5.2 of the CMAQ model (Appel et al., 2018), incorporating the Carbon Bond 6 (Yarwood et al., 2010) gas-phase chemistry mechanism and the AERO6 particulate matter chemistry mechanism. CMAQ, a widely recognized CTM, is renowned for its accurate simulation of air pollutant concentrations, including $PM_{2.5}$, which is attributed to its comprehensive representation of particulate matter formations. Meteorological data were generated using the WRF model, version 3.8, configured in the same manner as our previous studies (Ding et al., 2019a, b). Emission data were obtained from the high-resolution emission inventory developed by Tsinghua University (ABaCAS-EI) (Zheng et al., 2019), characterized by a spatial resolution of 27 km by 27 km and a temporal resolution of 1 h to match the CMAQ model. Biogenic emissions were derived from the estimation of the Model for Emissions of Gases and Aerosols from Nature (MEGAN) (Guenther et al., 2012). We conducted a thorough assessment of the performance of WRF and CMAQ in simulating meteorological variables and air pollutant concentrations, employing extensive comparisons with observational data in our previous studies (Ding et al., 2019a, b).

The simulation domain spans a significant portion of East Asia and is depicted by a grid consisting of 182 rows and 232 columns, featuring a horizontal resolution of 27 km by 27 km. The entirety of the troposphere (from ground level to 100 mbar) is represented using 14 layers with sigma values, i.e., 1.00, 0.995, 0.99, 0.98, 0.96, 0.94, 0.91, 0.86, 0.8, 0.74, 0.65, 0.55, 0.4, 0.2, and 0.00. These sigma values correspond to altitudes of 19, 57, 114, 230, 386, 584, 910, 1375, 1908, 2618, 3598, 5061, 7620, and 11944 m above ground level, both in the domain and on an annually averaged basis.

We align the simulated $PM_{2.5}$ concentrations from CMAQ with the CNEMC based on their respective locations, treating them as the "label" for training the machine-learning model. Though previous studies provide some validation schemes to evaluate the model's extrapolation capacity (Dong et al., 2020), for the remaining grid cells that encompass surrounding $PM_{2.5}$ areas where observations are not available, the pre-

dicted concentrations with machine-learning methods cannot be directly compared and examined. This paper focuses on assessing the model's performance in predicting these points, accounting for over 90 % of the total number of grid cells. The simulation data serve as the ground truth for the evaluation of the output of the machine-learning model.

The WRF and CMAQ simulations were evaluated in our previous studies (Ding et al., 2019a, b), demonstrating acceptable agreement with CNEMC observations, albeit with limitations in areas where observations are available. In rural areas where no observations are available, direct comparison of CMAQ predictions with actual observations is not possible. However, the CMAQ data used in this study primarily serve to establish a testbed representing scenarios based on physical laws such as emission, diffusion, advection, and deposition. This approach contrasts with reanalysis or data fusion methods, which may deviate from these physical functions, even though they might exhibit better agreement with observations when available.

### 2.2 Decision-tree-based machine-learning method

This study employed three decision-tree-based machine-learning algorithms, i.e., random forest (Belgiu and Drăguţ, 2016), XGBoost (Chen and Guestrin, 2016), and LightGBM (Ke et al., 2017), to serve as benchmark cases given their widespread use in previous studies. Additionally, Deep Forest (denoted as DeepRF in this study) (Zhou and Feng, 2019), known for its superior performance (Wei et al., 2023), was included as an additional method to be evaluated in this study.

We incorporated similar features used in the machine-learning model, including observed meteorological variables (WRF output) and land use information. The reason is that we deliberately avoided using CTM simulation results for two key reasons, while some previous studies included CTM results as additional features in training machine-learning models. First, the CTM will be applied to the testbed to evaluate the model's performance, and introducing CTM results could leak information as these results are utilized as labels and therefore cannot be used as input thereafter. Second, we aimed to propose a comprehensive CTM-free method that relies exclusively on satellite products and meteorological variations obtained from observations. This choice is motivated by the low efficiency when using the CTM and the uncertainties that this introduces. Furthermore, the only additional information provided by the CTM is related to emissions, which still suffer from uncertainties. Therefore, instead of relying on the CTM or prior emission data, we introduce the $NO_2$ column density. This variable is highly correlated with emission sources and can be directly observed from satellites, offering a more accurate representation of emission information.

Given our objective of assessing grid cells outside the designated label, there is no overlap between the training and test datasets. To evaluate the model's performance on the labels, we employ temporal validation. Specifically, the model is trained using data from only the first 25 d of each month, and the remaining days are reserved for testing. This approach helps gauge the model's effectiveness in handling temporal variations and provides a robust assessment of its performance on the specified labels. We fixed the days for training rather than selecting them randomly to ensure that all the methods use exactly the same data for training and testing, enabling a fair comparison. Random selection would still require us to fix the randomly selected days for all the methods, similar to fixing all the days at the outset. Moreover, the purpose of this study is to investigate prediction errors for grid cells not included in the training dataset. Even when using the first 25 d of the training dataset, we consistently observe similar prediction errors in other sites (similar to out-of-sample validation), regardless of which days are selected for training or testing.

### 2.3 Residual neural network (ResNet) method

The incorporation of spatiotemporal-neighborhood features is crucial for enhancing the model's ability to discern the evolution of vertical profiles in both urban and downwind areas (Chen et al., 2023). Beyond simply including corresponding features from the surrounding neighborhood grid cells as additional predictors of $PM_{2.5}$ concentrations at target grid cells in decision-tree-based methods, we also employ a deep-learning method, i.e., ResNet (He et al., 2016). This choice is motivated by its demonstrated advantage in handling the nonlinearity inherent in atmospheric processes, as suggested in our previous study (Xing et al., 2020).

ResNet consists of an initial layer with 128 channels and incorporates eight residual blocks. The feature maps, encompassing meteorological variables, land use information, and AOD, are fed into the conventional neural network (CNN)-based structure with a 3-by-3 kernel size, as illustrated in Fig. S1 in the Supplement. Additionally, we incorporate the feature into the previous and next days to help capture the transport flow of pollutants in the model. The training loss will concentrate solely on points corresponding to the monitoring sites, generating predictions exclusively for these specific locations given the scattered nature of the labels. As a result, predictions for other points will be entirely out of sample, relying on data from the same locations as the monitoring sites.

One thing should be noted: all machine-learning methods use the same input features to ensure a fair comparison. The only difference is that the features for the new proposed methods (e.g., ResNet) include data from the neighborhood (nearby grid cells and the previous or next time steps) in addition to the local grid and time data.

Throughout the training phase, we employed the mean squared error (MSE) loss function, conducting a total of 1000 epochs, which demonstrated sufficient effectiveness in achieving robust performance during both training and test-

ing. The learning rate started at 0.0001 and underwent linear decay, reaching zero by the conclusion of the training process. Additionally, we utilized the Adam optimizer (Kingma and Ba, 2014) to enhance the convergence of the model.

## 3 Results

### 3.1 Imbalance in site distribution leads traditional methods to overestimate downwind $PM_{2.5}$

To explore uncertainties in traditional machine-learning methods, we initially adhere to their typical design, relying exclusively on local features within each grid cell. This approach involves utilizing only the feature data from the same location as the target grid cell. The model trained with the RF method successfully captures the spatial distribution of $PM_{2.5}$, showing elevated levels in eastern China and lower levels in the west (Fig. 2a–b). It exhibits acceptable performance for the label grid cells during validation (Fig. 2c–d: $R^2 = 0.98$ and RMSE $= 5.28\,\mu g\,m^{-3}$ in the training dataset, $R^2 = 0.81$ and RMSE $= 16.1\,\mu g\,m^{-3}$ in the test dataset).

However, considerable errors are observed across space, particularly in polluted regions with high baseline $PM_{2.5}$ concentrations (Fig. 2e). Positive biases (i.e., predictions greater than those of CMAQ) increase with the distances from the monitoring sites, even as $PM_{2.5}$ concentrations decrease. This suggests an overestimation in predictions for downwind areas away from the monitoring sites (Fig. 2f). This is mainly attributed to the traditional model's difficulty in discerning variations in vertical profiles between urban and suburban areas. Training is primarily focused on urban areas, where pollution is concentrated near the surface due to ground-level emission sources. In contrast, pollution in downwind areas is transported aloft. Therefore, the model, trained on urban sites where ground-level pollution from AOD is more prominent, failed to accurately capture the diverse aerosol vertical profiles in source/urban and downwind/rural areas. This discrepancy resulted in overestimations in downwind areas (as illustrated in Fig. 2g).

Contrary to traditional expectations, significant absolute errors mostly occur in eastern China rather than in the west due to the large baseline concentration, even though the relative error is smaller (about 20 %) (Fig. S2 in the Supplement). While the east has more densely located monitoring sites, these are primarily situated in urban centers. This imbalance in site distribution, combined with much higher concentrations, results in substantial biases in eastern China.

Similar phenomena are observed in three other benchmark models that have been applied in previous studies, i.e., XgBoost, LightGBM, and DeepRF. All of these models demonstrate robust performance in both training and testing at the monitoring sites ($R^2 > 0.8$ and RMSE $< 16.2\,\mu g\,m^{-3}$ in the test cases, as depicted in Fig. S3 in the Supplement). However, they display similar uncertainties in downwind $PM_{2.5}$, with significant errors occurring in the surrounding grid cells of the monitoring sites rather than in remote sites where concentrations are relatively low. Clearly, we can conclude that the uneven distribution of sites introduces considerable biases into $PM_{2.5}$ estimation when using traditional methods that rely on local features.

### 3.2 Inclusion of spatiotemporal-neighborhood features improves the surrounding $PM_{2.5}$ prediction over traditional approaches

As previously discussed, the ineffectiveness of a machine-learning model trained on imbalanced samples can be attributed primarily to insufficient information regarding the spatial variation of vertical profiles from the source to the downwind area. To enhance the integration of crucial information regarding vertical profiles, we introduce spatiotemporal-neighborhood features into the model. This addition aims to empower the model with the ability to distinguish between urban and downwind areas. Leveraging the CNN-based structure, known for its effectiveness in exploring nonlinear relationships between neighboring grid cells, we opt for the widely used deep-learning ResNet model. This choice facilitates the establishment of nonlinear relationships between predicted $PM_{2.5}$ concentrations and multiple spatiotemporal-neighborhood features. In contrast to traditional methods that solely focus on single time features, our approach incorporates both preceding and succeeding time features to enhance a model's capacity to discern differences between urban and downwind grid cells. This inclusion is motivated by the fact that plume transport is predominantly influenced by flow dynamics represented by the variation in the temporal neighborhood (before and after) features. Our previous studies also demonstrated the effectiveness of linking grid cells to time series information on $PM_{2.5}$ estimation, underscoring the rationale for this inclusive approach (Teng et al., 2023; Ding et al., 2024). Additionally, considering that AOD measured by satellites, such as MODIS, captures only a single time step while predictions are made for daily averages, the inclusion of extra time step information proves beneficial in capturing a broader temporal context compared to a single time snapshot (the model is called ResNet-time).

The results indicate that, while the spatial pattern of $PM_{2.5}$ predicted with ResNet closely resembles that of other models (Fig. 3a), it significantly enhances model performance in predicting $PM_{2.5}$ for both the training dataset (reducing RMSE from 4 to $2\,\mu g\,m^{-3}$ and increasing $R^2$ slightly) and the test dataset (reducing RMSE from 14 to $8\,\mu g\,m^{-3}$ and increasing $R^2$ from 0.8 to 0.9). This improvement can be attributed to the incorporation of both spatial and temporal features. The performance of the traditional RF model is enhanced by replacing it with the ResNet model, and this improvement is further amplified by including temporal features (previous and next time steps) in the ResNet-time model (see Fig. 3b). The incorporation of surrounding features in the ResNet-time model significantly mitigates both absolute and

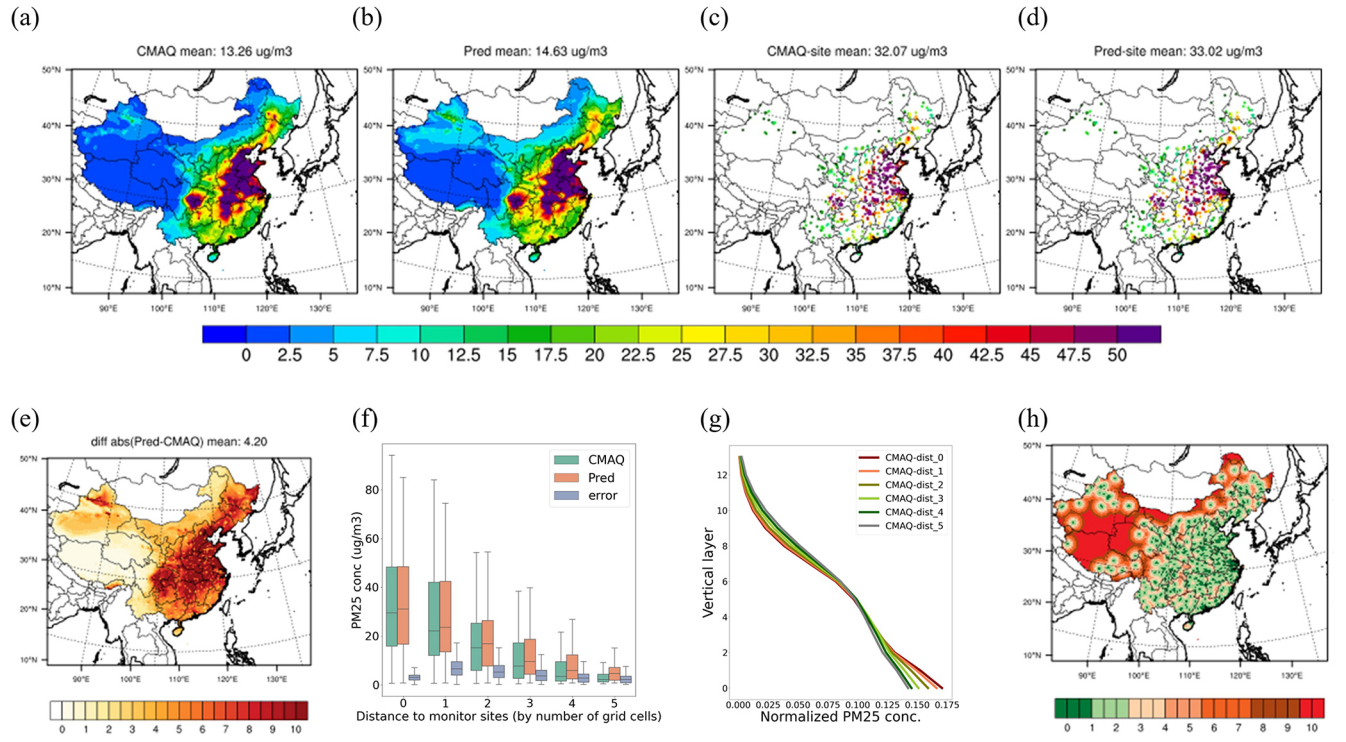


**Figure 2.** Performance of the monitor-located RF model in predicting surface $PM_{2.5}$ levels. The comparison includes the spatial distribution of surface $PM_{2.5}$ (**a**: ground truth; **b**: model prediction), $PM_{2.5}$ levels at monitoring sites (**c**: label values; **d**: predicted values), the error distribution across space **(e)** and distance **(f)**, and the vertical structure of $PM_{2.5}$ concentration **(g)** across the distances from the monitors, with their spatial distribution shown in panel **(h)**.

relative errors in eastern China across the spatial domain (see Figs. 3c and S2). However, some deterioration is observed in the west, which is primarily attributable to limited samples. The model, becoming more complex, lacks sufficient training samples in the west, leading to overfitting in that region.

The inclusion of spatiotemporal-neighborhood features also significantly improves the performance of traditional benchmark models by incorporating corresponding features of the surrounding eight neighborhood grid cells and the temporal (before and after) neighborhood information as additional predictors of $PM_{2.5}$ concentrations at the target grid cells. Improvements are observed in both the training and test datasets across all four benchmark models, as depicted in Fig. S4 in the Supplement. Notably, all the models demonstrate a reduction in RMSE after integrating spatiotemporal-neighborhood features, especially for the downwind area (within a distance of one to three grid cells) (see Fig. 3d). However, performance is barely improved or even worsens in faraway sites (distance of more than four grid cells) due to the limitations of the training samples. Even the ResNet-time model demonstrates better performance, primarily in eastern China, where the distance to the monitoring sites is within zero to two grid cells. The performance is slightly worse in the western region, where the distance to the monitoring sites exceeds four grid cells (Fig. 3d). The "new" method, applied to the original tree-based method, also shows superior performance compared to the original, although it performs slightly worse in western China. Clearly, enhancing the training sample is crucial for further improving the model predictions, as discussed in the following.

### 3.3 Balancing site distribution is crucial for improving the prediction for the entire space of $PM_{2.5}$

Utilizing the ResNet-time model, we explore the correlation between model errors and the distance to the monitoring sites, together with the concentrations at the nearest monitor. Notably, significant errors were observed in sites within a distance of two grid cells (refer to Fig. S5 in the Supplement) and those near monitors exhibiting high concentrations (refer to Fig. S6 in the Supplement). Consequently, two criteria, i.e., (1) the baseline concentration in nearby monitoring sites (referred to as "B-conc") and (2) the distance from monitoring sites (referred to as "D-site"), are established to select the potential samples to refine predictions across the spatial domain.

Three sample groups are delineated based on these criteria:

1. B-conc $>30\,\mu g\,m^{-3}$ and D-site of one to five grid cells;
2. B-conc within $20–30\,\mu g\,m^{-3}$ and D-site of two to five grid cells; and

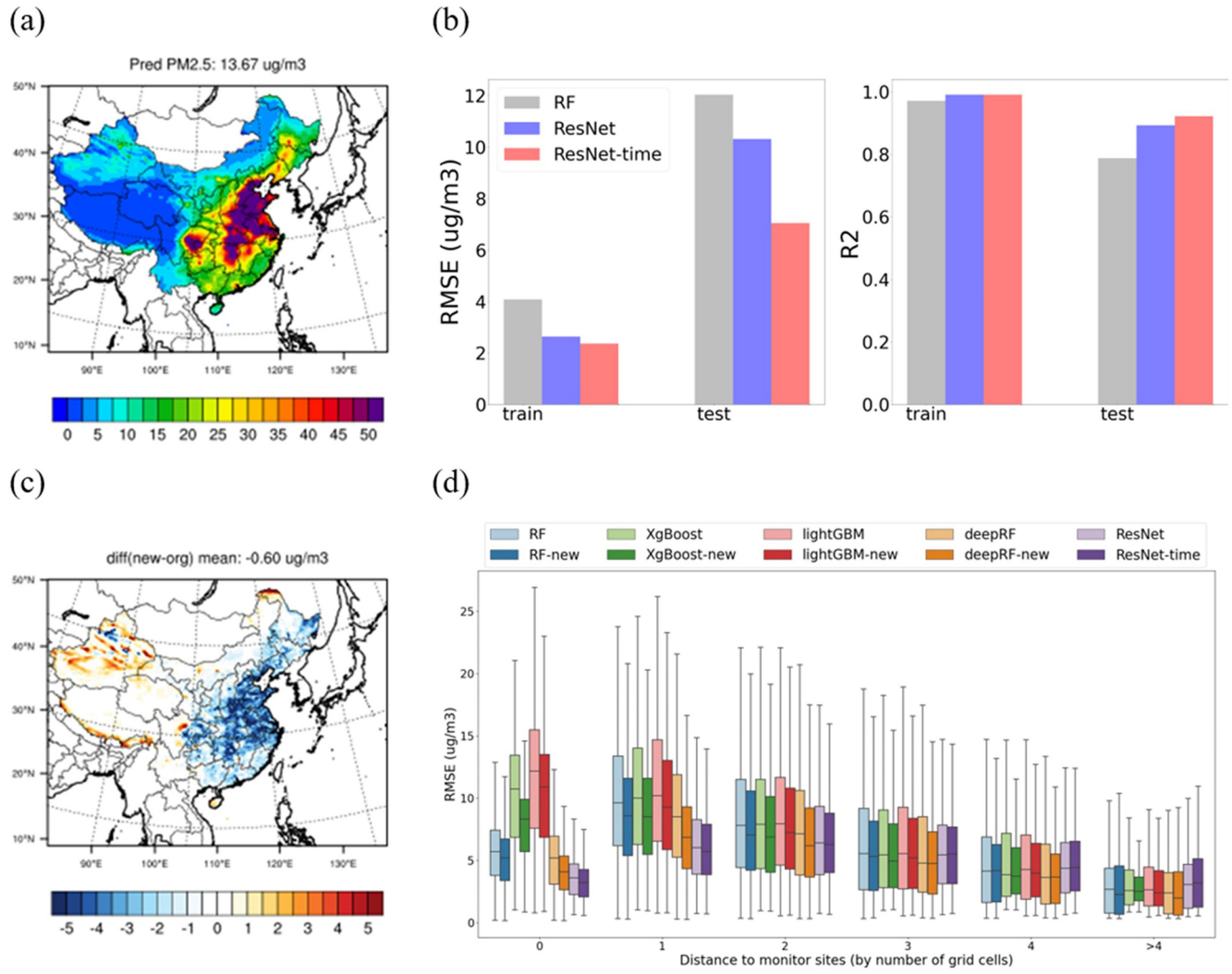


**Figure 3.** Improvement after implementing surrounding grid cell features. The panels show **(a)** prediction of surface $PM_{2.5}$ using ResNet, **(b)** comparison with the random forest (RF) model at the monitoring sites, **(c)** error comparison between ResNet and RF (where blue indicates better performance by ResNet and red indicates worse performance), and **(d)** error comparison across the distances to the monitors for all the models, including the new addition of surrounding features to each model's prediction.

3. B-conc within 10–20 $\mu g\ m^{-3}$ and D-site of three to five grid cells.

This design not only focused on the area suffering from large impacts of pollution but also allowed the selection of sites in remote regions with moderate baseline concentrations, as illustrated in Fig. S7 in the Supplement. To enhance the representativeness of the chosen sites, random selections are independently conducted in each of the three groups, encompassing 10 % (~ 300 sites, half of the existing sites), 20 % (~ 600 sites, equal to the existing sites), 30 % (~ 900 sites, 1.5 times the existing sites), 40 % (~ 1200 sites), 70 % (~ 2100 sites), and all of the samples (~ 3000 sites). The testbed developed in this study enables an efficient evaluation of the model's performance by training it with these additional sites.

The results indicate that an increase in the number of training samples effectively enhances a model's performance in $PM_{2.5}$ estimation, with RMSEs continuously decreasing as the number of samples increases (see Fig. 4a). This improvement is primarily observed in the downwind area, while the performance at the monitoring sites deteriorates due to the original model being overfitted to these specific sites (Fig. 4b). The rate of improvement diminishes after the inclusion of 20 % of the samples, implying that just doubling the current ground monitors wisely can effectively balance the training samples to ensure the accuracy of $PM_{2.5}$ estimations (RMSE reduced by 20 %–30 %). As illustrated in the example with 20 % sample inclusion in Fig. 4c–e, it is recommended that more than half of the new sites be set up in eastern China, where $PM_{2.5}$ concentration is high. Addition-

ally, it is suggested that 10 % of the sites be set up in remote areas that are influenced by transport from heavy pollution regions but lack nearby ground measurements. The inclusion of additional sites proves effective in significantly reducing prediction errors across the entire spatial domain, leading to much closer agreement with the ground truth (Fig. 2a) in the $PM_{2.5}$ spatial pattern.

It is important to acknowledge that errors may be influenced by factors beyond site distribution problems, such as systematic errors arising from insufficient features. Baseline errors are referenced to those trained with all points using ResNet, amounting to within 1.7 $\mu g\,m^{-3}$ (Fig. S8 in the Supplement). Similarly, training with all points may increase errors in monitoring sites, as the original model might be overfitted to these sites rather than representing the overall situation (Fig. 4b).

### 3.4 Potential biases and optimized site selections under real-world conditions

While ground measurements are unavailable for the entire space, we conducted the evaluation using both the traditional RF method and the ResNet-time model developed previously with satellite data. Both models were trained using real-world satellite data and ground monitoring $PM_{2.5}$ observations during 2013–2021, and their differences can be considered part of the potential biases associated with the influence of incorporating spatiotemporal features for enhancing the model's ability to identify vertical structures.

The results suggest that both models effectively replicate the time series of monthly mean $PM_{2.5}$ concentrations across monitoring sites from 2013 to 2021 (see Fig. 5a–b). However, considerable disparities emerge in their predictions for other areas (Fig. 5c). The new predictions using the ResNet-time model generally exhibit lower $PM_{2.5}$ concentrations, particularly in the northern and western regions (Fig. 5d–f), with a more significant impact observed as the distance to the ground monitoring sites increases (Fig. 5j). A notable discrepancy in the population-weighted $PM_{2.5}$ concentration is observed in eastern China, which has a large population, implying that the errors also applied to human health assessment (Fig. 5g). Since the ResNet-time model demonstrates superior performance compared to the traditional RF model in both training (reducing RMSE from 6 to 2.5 $\mu g\,m^{-3}$ and increasing $R^2$ from 0.9 to 1.0) and test data (reducing RMSE from 20 to 15 $\mu g\,m^{-3}$ and increasing $R^2$ from 0.4 to 0.6), it appears that traditional methods might significantly overestimate $PM_{2.5}$ concentrations (by 5–30 $\mu g\ m^{-3}$) and $PM_{2.5}$ exposure in suburban and rural areas by 3 million people $\cdot$ $\mu g\,m^{-3}$ (Fig. 5k) due to the sample imbalance problem throughout 2013–2021. Similar results are also suggested in the other three benchmark models (Figs. S9–S10 in the Supplement). The actual errors might be even larger, as the inclusion of spatiotemporal-neighborhood features in the ResNet-time model can only mitigate a portion of the errors.

Incorporating a significant number of additional sites is necessary to balance the training samples and further reduce the uncertainties. Following the two previously defined criteria (i.e., B-conc and D-site), three groups of samples are selected. Similarly, 20 % of the samples (631 in total, close to the number of existing sites) in each group are proposed as potential additional sites in the future, as presented in Fig. 5l–m. Group 1 (30 % of the total number of add-on sites) is primarily situated in polluted regions (B-conc $>60\,\mu g\,m^{-3}$, D-site one to five grids), encompassing areas such as the Beijing–Tianjin–Hebei region and the desert region in the west. Group 2 (40 % of the total number of add-on sites) represents sites with a moderate distance to existing monitoring sites with a heavier pollution level, compared to Group 1 (B-conc within 40–60 $\mu g\,m^{-3}$, D-site two to five grid cells). Lastly, Group 3 (30 % of the total number of add-on sites) represents sites located far away from existing monitoring sites with a low pollution level (B-conc $<40\,\mu g\,m^{-3}$, D-site four to five grid cells), situated in remote areas with limited influence from transport originating in polluted regions. As indicated by the previous testbed analysis, including these additional sites has the potential to reduce errors by at least 20 %, leading to more accurate machine-learning estimation of $PM_{2.5}$ concentrations with a more balanced training sample set.

## 4 Data availability

The numerical-model-informed testbed and the corresponding estimated $PM_{2.5}$ concentrations spanning the 9 years (2013–2021) can be found at https://doi.org/10.5281/zenodo.11122294 (Li et al., 2024a), and an updated version with files in NetCDF format can be found at https://doi.org/10.5281/zenodo.12636976 (Li et al., 2024b). In addition to the long-term $PM_{2.5}$ dataset created using our new method, which can be used for health assessments and studying air pollution influences, we provide testbed data crucial for evaluating machine-learning-based retrieval methods, especially in scenarios where no ground-truth data are available.

The testbed dataset includes all inputs and outputs following the physical model simulation, which naturally correlates with physical laws such as emission, diffusion, advection, and deposition, representing typical conditions that any prediction method should meet. These data can be used to evaluate and compare methods using the same dataset, allowing for continuous improvement. Besides traditional cross-validation, our proposed testbed validation is highly recommended for examining a method's predictive ability. We will continue updating the testbed data for other pollutants and with different resolutions and regions in future studies.

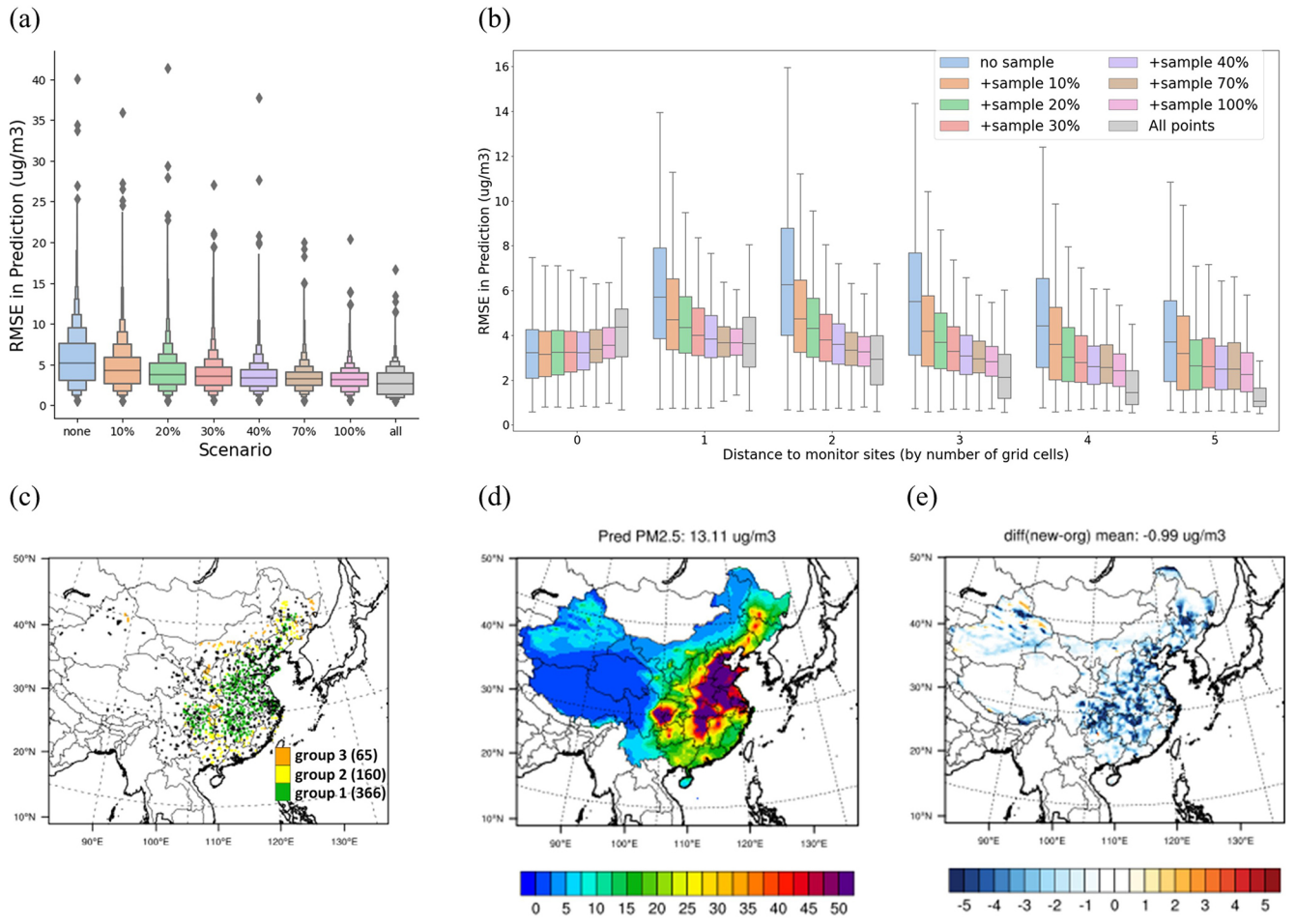


**Figure 4.** Improved performance through the integration of additional sites using the ResNet model. The comparison includes model performance when adding points across the overall domain **(a)** and across distances to the monitoring sites **(b)**, the scenario of adding 20 % more sample details to the locations of specific sites (black dots represent the 619 original monitoring sites) **(c)**, the prediction of the surface $PM_{2.5}$ concentration in this scenario **(d)**, and the corresponding reduction in errors **(e)**.

## 5 Code availability

The ResNet-time model developed in this study can be downloaded at https://doi.org/10.5281/zenodo.11122294 (Li et al., 2024a).

## 6 Discussion and conclusions

Amidst the advancements in satellite products and machine-learning techniques, ground-level $PM_{2.5}$ data have found extensive applications in health assessments and related fields. However, their uncertainties have remained unexplored due to the lack of ground-truth data covering the whole space. This study designed a physically informed testbed by leveraging CTM simulations to evaluate $PM_{2.5}$ estimation across the entire spatial domain and quantified the associated uncertainties in the $PM_{2.5}$ mapping across the whole space. Traditionally, it was believed that errors would be significant in remote areas with few or no ground-based measurements, while observation-dense regions, following the spatial interpolation principle, were expected to exhibit better accuracy. Contrary to our expectations, our findings reveal that the largest absolute biases occur differently. One reason is the heavier baseline $PM_{2.5}$ concentration, and another significant factor is the sample training imbalance problem. Ground-based measurements, designed for monitoring heavy pollution, are predominantly located in urban or industrial areas. Using these measurements as training samples misleads the machine-learning model into assuming uniform similarity to urban sites, especially in vertical structures. In reality, the vertical profile varies significantly with the flow after the pollutant is emitted from the source. This sample imbalance issue causes the machine-learning model to fail to provide accurate predictions for $PM_{2.5}$ across the entire spatial domain.

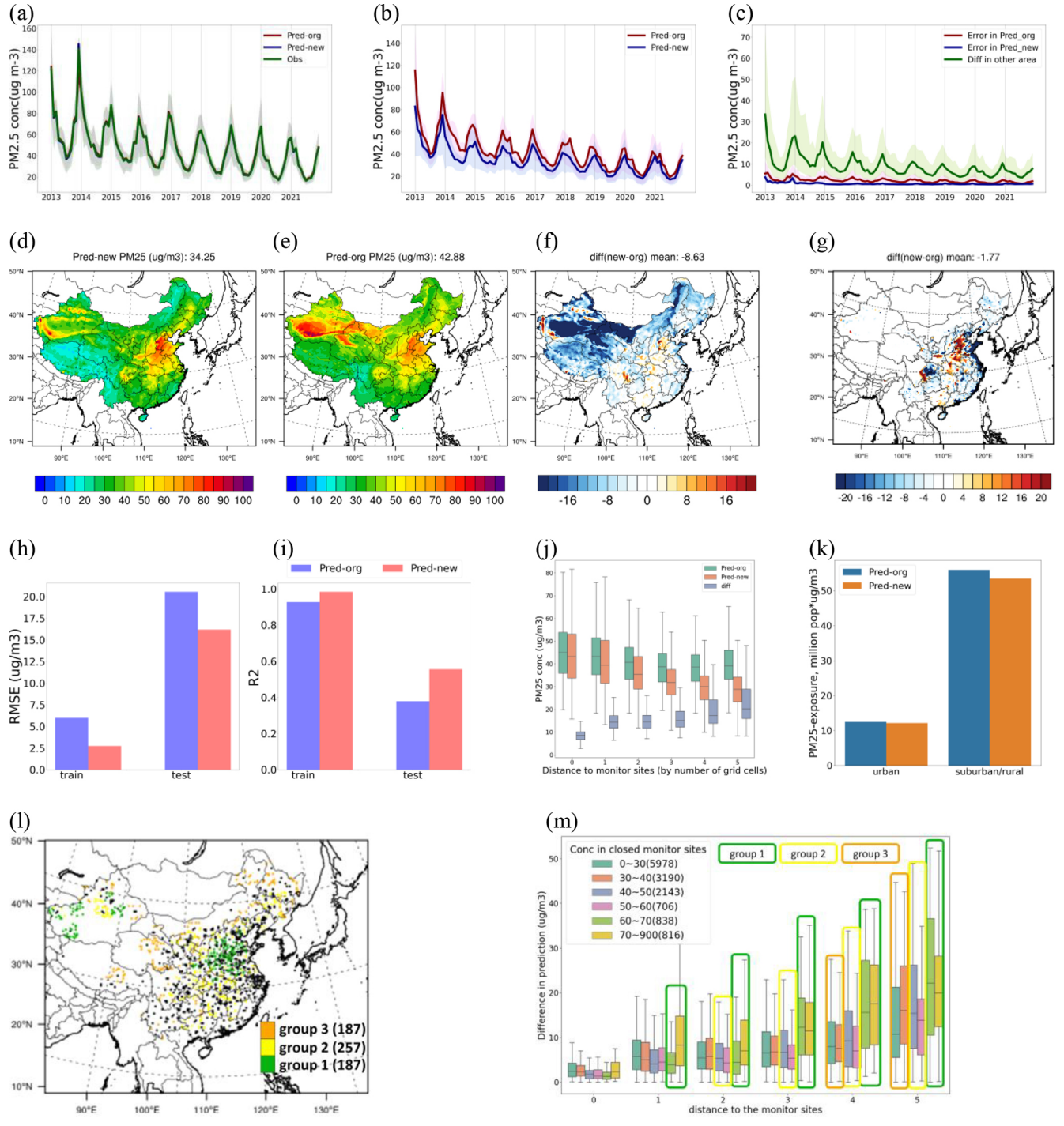


**Figure 5.** Improved estimation of $PM_{2.5}$ and the related exposure across China using satellite products and ground observations from 2013 to 2021. The comparison of predicting the time series of monthly mean $PM_{2.5}$ concentrations from 2013 to 2021 is conducted using the original RF model and the new ResNet-time model based on monitoring sites **(a)** and other grid cells **(b)**, with the difference between the new and original predictions shown in panel **(c)**. The prediction of the 9-year average $PM_{2.5}$ concentration over 2013–2021 is compared using the new ResNet-time model **(d)** and the original RF model **(e)**, with their differences being in absolute concentrations **(f)** and population-weighted concentrations **(g)**. The influence of considering spatiotemporal features in the new model on the performance at the monitoring sites is shown for RMSE **(h)**, $R^2$ **(i)**, $PM_{2.5}$ prediction across the distances to the monitors **(j)**, and $PM_{2.5}$ exposure (**k**, where suburban/rural represents areas within a distance of five grid cells from the monitoring sites). Potential grid cells as additional sites to improve the accuracy of downwind $PM_{2.5}$ prediction are shown with their spatial locations **(l)** and the selection of each group **(m)**.

The newly developed testbed also enables us to seek the best solutions, such as optimizing model structures or enhancing training samples, to improve satellite-retrieval model performance. Our results underscore the importance of incorporating spatiotemporal features to enhance the machine-learning model's ability to identify differences between urban and downwind conditions that are not explored in the recent literature. However, fully addressing the sample imbalance problem necessitates the addition of more ground monitoring sites to achieve a more balanced distribution of training samples for machine learning in China. In recent years, the Chinese government has expanded monitoring sites towards suburban areas, increasing the total number of monitoring sites by about 400 (from about 1600 in 2017 to about 2020 in 2021). While these additional samples have effectively improved $PM_{2.5}$ predictions (as presented in Fig. S11 in the Supplement), they only account for about 100 grid cells in the 27 km-by-27 km domain. According to the estimations in this study, approximately 600 grid cells are needed to locate monitoring sites in the future. Some studies incorporate CTM simulation data as an additional feature for predicting $PM_{2.5}$, while the uncertainties of CTM hinder performance enhancement. To demonstrate this, we conducted RF predictions with CMAQ data as an additional feature, but the improvement compared to the original RF model was minimal, especially when compared to using the additional neighborhood information proposed in this study (Fig. S12 in the Supplement). Besides, compared to CTM simulations, $NO_2$ column density better represents emission information and can significantly enhance model performance. As illustrated in Fig. S13 in the Supplement, excluding $NO_2$ column data from the features used in the machine-learning model reduces its performance in predicting surface $PM_{2.5}$, leading to even more errors due to the sample imbalance problem.

Although this study is conducted at a relatively coarse resolution of 27 km over China due to the computational burden of running a CTM at a fine resolution in a large-scale domain, the testbed method proposed here can also be applied with higher-resolution retrievals when the simulation data are available. A similar testbed study conducted in the Continental United States (CONUS) domain at 12 km resolution revealed the same imbalance problem (Zhang et al., in preparation), indicating that this issue persists at finer-resolution scales, especially in urban and industrial areas, due to spatial heterogeneity in emissions (Li and Xing, 2024) and the complexity of spatial gradients of particulate matter pollution observed at high resolution through AOD (Lin et al., 2021). At a fine resolution (e.g., 1 km), while the number of observation sites may increase slightly (eliminating the need for grouping to one 27 km grid cell like in this study), the number of grid cells to be predicted increases significantly. Therefore, it is essential to conduct similar testbed studies using 1 km CMAQ results (Tao et al., 2020) to evaluate the performance of machine-learning methods. This might be more feasible with a more comprehensive CMAQ dataset using nesting rather than specific subdomains.

This study also successfully demonstrates leveraging of the CTM to generate abundant data to test machine-learning methods, overcoming limitations associated with data availability. While derived from a numerical-model-based testbed, it is important to acknowledge that the numerical model itself may encounter uncertainties related to emissions and chemical mechanisms, potentially leading to discrepancies with real observations. Nevertheless, the testbed serves as a specific scenario for evaluating satellite-retrieval methods, with the expectation that these methods should perform effectively in various scenarios, including those generated from CTM simulations. Therefore, the errors observed in the CTM-based testbed also imply their existence when applied to real data. Additionally, although this study primarily focuses on the analysis of $PM_{2.5}$ in China, the identified errors may extend to other pollutants and countries as a whole, particularly when facing similar sample imbalance problems (i.e., lacking suburban or rural representative sites). Leveraging the testbed developed in this study can be immensely helpful in examining uncertainties in other pollutants and countries or in other geoscience applications facing similar sample imbalance challenges.

**Supplement.** The supplement related to this article is available online at: https://doi.org/10.5194/essd-16-3781-2024-supplement.

**Author contributions.** SL and JX designed the experiments and carried them out. YD helped with the data processing. JSF helped with the manuscript review. SL prepared the manuscript with contributions from all the co-authors.

**Competing interests.** The contact author has declared that none of the authors has any competing interests.

**Disclaimer.** Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

**Acknowledgements.** The authors are grateful for the free use of the MODIS-AOD, OMI-$NO_2$ column density, and ground monitoring observations provided by the China National Environmental Monitoring Center.

**Financial support.** This research has been supported by the National Natural Science Foundation of China (grant no. 42375131),

the Fengyun Application Pioneering Project (grant no. FY-APP-2022.0503), and the Microsoft Climate Research Initiative program.

**Review statement.** This paper was edited by Yuqiang Zhang and reviewed by two anonymous referees.

## References

Appel, K. W., Pouliot, G. A., Simon, H., Sarwar, G., Pye, H. O. T., Napelenok, S. L., Akhtar, F., and Roselle, S. J.: Evaluation of dust and trace metal estimates from the Community Multiscale Air Quality (CMAQ) model version 5.0, Geosci. Model Dev., 6, 883–899, https://doi.org/10.5194/gmd-6-883-2013, 2013.

Appel, K. W., Napelenok, S., Hogrefe, C., Pouliot, G., Foley, K. M., Roselle, S. J., Pleim, J., Bash, J., Pye, H. O. T., Heath, N., Murphy, B., and Mathur, R.: Overview and evaluation of the community multiscale air quality (CMAQ) modeling system version 5.2, in: Air Pollution Modeling and its Application XXV 35, Springer International Publishing, 69–73, https://doi.org/10.1007/978-3-319-57645-9_11, 2018.

Bai, K., Li, K., Guo, J., and Chang, N. B.: Multiscale and multisource data fusion for full-coverage $PM_{2.5}$ concentration mapping: Can spatial pattern recognition come with modeling accuracy? ISPRS J. Photogramm., 184, 31–44, 2022.

Belgiu, M. and Drăguţ, L.: Random forest in remote sensing: A review of applications and future directions, ISPRS J. Photogramm., 114, 24–31, 2016.

Bellouin, N., Boucher, O., Haywood, J., and Reddy, M. S.: Global estimate of aerosol direct radiative forcing from satellite measurements. Nature, 438, 1138–1141, 2005.

Celarier, E. A., Brinksma, E. J., Gleason, J. F., Veefkind, J. P., Cede, A., Herman, J. R., Ionov, D., Goutail, F., Pommereau, J.-P., Lambert, J.-C., van Roozendael, M., Pinardi, G., Wittrock, F., Schönhardt, A., Richter, A., Ibrahim, O.W., Wagner, T., Bojkov, B., Mount, G., Spinei, E., Chen, C. M., Pongetti, T. J., Sander, S. P., Bucsela, E. J., Wenig, M. O., Swart, D. P. J., Volten, H., Kroon, M., and Levelt, P. F.: Validation of Ozone Monitoring Instrument nitrogen dioxide columns, J. Geophys. Res.-Atmos., 113, D15S15, https://doi.org/10.1029/2007JD008908, 2008.

Chen, D., Guo, H., Gu, X., Cheng, T., Yang, J., Zhan, Y., and Wei, X.: A spatial-neighborhood deep neural network model for $PM_{2.5}$ estimation across China, IEEE T. Geosci. Remote, 61, 4105815, https://doi.org/10.1109/TGRS.2023.3317905, 2023.

Chen, T. and Guestrin, C.: XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, San Francisco, CA, USA, 13–17 August 2016, 785–794, https://doi.org/10.1145/2939672.2939785, 2016.

Ding, D., Xing, J., Wang, S., Chang, X., and Hao, J.: Impacts of emissions and meteorological changes on China's ozone pollution in the warm seasons of 2013 and 2017, Front. Environ. Sci. Eng., 13, 76, https://doi.org/10.1007/s11783-019-1160-1, 2019a.

Ding, D., Xing, J., Wang, S., Liu, K., and Hao, J.: Estimated Contributions of Emissions Controls, Meteorological Factors, Population Growth, and Changes in Baseline Mortality to Reductions in Ambient $PM_{2.5}$ and $PM_{2.5}$-Related Mortality in China, 2013–2017, Environ. Health Persp., 127, 67009, https://doi.org/10.1289/EHP4157, 2019b.

Ding, Y., Li, S., Xing, J., Li, X., Ma, X., Song, G., Teng, M., Yang, J., Dong, J., and Meng, S.: Retrieving hourly seamless $PM_{2.5}$ concentration across China with physically informed spatiotemporal connection. Remote Sens. Environ., 301, 113901, https://doi.org/10.1016/j.rse.2023.113901, 2024.

Dong, L., Li, S., Yang, J., Shi, W., and Zhang, L.: Investigating the performance of satellite-based models in estimating the surface $PM_{2.5}$ over China, Chemosphere, 256, 127051, https://doi.org/10.1016/j.chemosphere.2020.127051, 2020.

Geng, G., Zheng, Y., Zhang, Q., Xue, T., Zhao, H., Tong, D., Zheng, B., Li, M., Liu, F., Hong, C., He, K., and Davis, S. J. Drivers of $PM_{2.5}$ air pollution deaths in China 2002–2017, Nat. Geosci., 14, 645–650, 2021.

Guenther, A. B., Jiang, X., Heald, C. L., Sakulyanontvittaya, T., Duhl, T., Emmons, L. K., and Wang, X.: The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2.1): an extended and updated framework for modeling biogenic emissions, Geosci. Model Dev., 5, 1471–1492, https://doi.org/10.5194/gmd-5-1471-2012, 2012.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, Nevada, USA, 26 June–1 July 2016, 770–778, https://doi.org/10.48550/arXiv.1512.03385, 2016.

He, Q., Qin, K., Cohen, J. B., Loyola, D., Li, D., Shi, J., and Xue, Y.: Spatially and temporally coherent reconstruction of tropospheric $NO_2$ over China combining OMI and GOME-2B measurements, Environ. Res. Lett., 15, 125011, https://doi.org/10.1088/1748-9326/abc7df, 2020.

Hoff, R. M. and Christopher, S. A.: Remote sensing of particulate pollution from space: have we reached the promised land?, J. Air Waste Manage., 59, 645–675, 2009.

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

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. Y.: Lightgbm: A highly efficient gradient boosting decision tree, in: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, 2017, 12.4–12.9, USA3149 – 3157, https://dl.acm.org/doi/10.5555/3294996.3295074 (last access: 24 August 2024), 2017.

Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, arXiv [preprint], https://doi.org/10.48550/arXiv.1412.6980, 2014.

Kong, L., Tang, X., Zhu, J., Wang, Z., Li, J., Wu, H., Wu, Q., Chen, H., Zhu, L., Wang, W., Liu, B., Wang, Q., Chen, D., Pan, Y., Song, T., Li, F., Zheng, H., Jia, G., Lu, M., Wu, L., and Carmichael, G. R.: A 6-year-long (2013–2018) high-resolution air quality reanalysis dataset in China based on the assimilation of surface observations from CNEMC, Earth Syst. Sci. Data, 13, 529–570, https://doi.org/10.5194/essd-13-529-2021, 2021.

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., and Pozzer, A.: The contribution of outdoor air pollution sources to premature mortality on a global scale, Nature, 525, 367–371, 2015.

Li, S. and Xing, J.: DeepSAT4D: Deep learning empowers four-dimensional atmospheric chemical concentration and emission

retrieval from satellite, The Innovation Geoscience, 2, 100061-1, https://doi.org/10.59717/j.xinn-geo.2024.100061, 2024.

Li, S., Ding, Y., Xing, J., and Fu, J.: Numerical model-informed testbed for surface PM2.5 concentration over China and its estimates during 2013–2021, Zenodo [code and data set], https://doi.org/10.5281/zenodo.11122294, 2024a.

Li, S., Ding, Y., Xing, J., and Fu, J.: Numerical model-informed testbed for surface PM2.5 concentration over China and its estimates during 2013–2021 Zenodo [data set], https://doi.org/10.5281/zenodo.12636976, 2024b.

Li, T., Shen, H., Yuan, Q., and Zhang, L.: Geographically and temporally weighted neural networks for satellite-based mapping of ground-level $PM_{2.5}$. ISPRS J. Photogramm., 167, 178–188, 2020.

Lin, H., Li, S., Xing, J., He, T., Yang, J., and Wang, Q.: High resolution aerosol optical depth retrieval over urban areas from Landsat-8 OLI images, Atmos. Environ., 261, 118591, https://doi.org/10.1016/j.atmosenv.2021.118591, 2021.

Liu, X. H., Zhang, Y., Cheng, S. H., Xing, J., Zhang, Q., Streets, D. G., Jang, C., Wang W., and Hao, J. M.: Understanding of regional air pollution over China using CMAQ, part I performance evaluation and seasonal variation, Atmos. Environ., 44, 2415–2426, 2010.

Ma, Z., Dey, S., Christopher, S., Liu, R., Bi, J., Balyan, P., and Liu, Y.: A review of statistical methods used for developing large-scale and long-term $PM_{2.5}$ models from satellite data, Remote Sens. Environ., 269, 112827, https://doi.org/10.1016/j.rse.2021.112827, 2022.

Martin, R. V., Jacob, D. J., Chance, K., Kurosu, T. P., Palmer, P. I., and Evans, M. J.: Global inventory of nitrogen oxide emissions constrained by space-based observations of $NO_2$ columns, J. Geophys. Res.-Atmos., 108, 4537, https://doi.org/10.1029/2003JD003453, 2003.

Mitchell, J. F., Johns, T. C., Gregory, J. M., and Tett, S. F. B. Climate response to increasing levels of greenhouse gases and sulphate aerosols, Nature, 376, 501–504, 1995.

Remer, L. A., Kleidman, R. G., Levy, R. C., Kaufman, Y. J., Tanré, D., Mattoo, S., Martins, J. V., Ichoku, C., Koren, I., Yu, H., and Holben, B. N.: Global aerosol climatology from the MODIS satellite sensors, J. Geophys. Res.-Atmos., 113, D14S07, https://doi.org/10.1029/2007JD009661, 2008.

Shin, M., Kang, Y., Park, S., Im, J., Yoo, C., and Quackenbush, L. J.: Estimating ground-level particulate matter concentrations using satellite-based data: A review, GIsci. Remote Sens., 57, 174–189, 2020.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., Huang, X.-Y., Wang, W., and Powers, J.G.: A Description of the Advanced Research WRF Version 3, NCAR Tech. Note, NCAR/TN-475+STR, 113 pp., https://doi.org/10.5065/D68S4MVH, 2008.

Tao, H., Xing, J., Zhou, H., Pleim, J., Ran, L., Chang, X., Wang, S., Chen, F., Zheng, H., and Li, J.: Impacts of improved modeling resolution on the simulation of meteorology, air quality, and human exposure to $PM_{2.5}$, $O_3$ in Beijing, China, J. Clean. Prod., 243, 118574, https://doi.org/10.1016/j.jclepro.2019.118574, 2020.

Teng, M., Li, S., Xing, J., Fan, C., Yang, J., Wang, S., Song, G., Ding. Y., Dong, J., and Wang, S.: 72-hour real-time forecasting of ambient $PM_{2.5}$ by hybrid graph deep neural network with aggregated neighborhood spatiotemporal information, Environ. Int., 176, 107971, https://doi.org/10.1016/j.envint.2023.107971, 2023.

Wang, Z., Hu, B., Huang, B., Ma, Z., Biswas, A., Jiang, Y., and Shi, Z.: Predicting annual $PM_{2.5}$ in mainland China from 2014 to 2020 using multi temporal satellite product: An improved deep learning approach with spatial generalization ability, ISPRS. J. Photogramm., 187, 141–158, 2022a.

Wang, Z., Li, R., Chen, Z., Yao, Q., Gao, B., Xu, M., Yang, L., Li, M., and Zhou, C.: The estimation of hourly $PM_{2.5}$ concentrations across China based on a Spatial and Temporal Weighted Continuous Deep Neural Network (STWC-DNN), ISPRS. J. Photogramm., 190, 38–55, 2022b.

Wei, J., Li, Z., Chen, X., Li, C., Sun, Y., Wang, J., Lyapustin, A.,Brasseur, G., Jiang, M., Sun, L., Wang, T., Jung, C., Qiu, B., Fang, Liu, X., Hao, J., Wang, Y., Zhan, M., Song, X., and Liu, Y.: Separating Daily 1 km $PM_{2.5}$ Inorganic Chemical Composition in China since 2000 via Deep Learning Integrating Ground, Satellite, and Model Data, Environ. Sci. Technol., 57, 18282–18295, https://doi.org/10.1021/acs.est.3c00272, 2023.

Xiao, Q., Chang, H. H., Geng, G., and Liu, Y.: An ensemble machine-learning model to predict historical $PM_{2.5}$ concentrations in China from satellite data, Environ. Sci. Technol., 52, 13260–13269, 2018.

Xing, J., Zheng, S., Ding, D., Kelly, J. T., Wang, S., Li, S., Qin, T., Ma, M., Dong, Z., Jang, C., Zhu, Y., Zheng, H., Ren, L., Liu, T.-Y., and Hao, J.: Deep learning for prediction of the air quality response to emission changes, Environ. Sci. Technol., 54, 8589–8600, 2020.

Yan, X., Zang, Z., Luo, N., Jiang, Y., and Li, Z.: New interpretable deep learning model to monitor real-time $PM_{2.5}$ concentrations from satellite data, Environ. Int., 144, 106060, https://doi.org/10.1016/j.envint.2020.106060, 2020.

Yarwood, G., Jung, J., Whitten, G. Z., Heo, G., Mellberg, J., and Estes, M.: Updates to the Carbon Bond mechanism for version 6 (CB6), in: 9th Annual CMAS Conference, Chapel Hill, NC, USA, 11–13 October 2010, https://cmascenter.org/conference/2010/abstracts/emery_updates_carbon_2010.pdf (last access: 19 August 2024), 2010.

Zheng, H., Zhao, B., Wang, S., Wang, T., Ding, D., Chang, X., Liu, K., Xing, J., Dong, Z., Aunan, K., Liu, T., Wu, X., Zhang, S., and Wu, Y.: Transition in source contributions of $PM_{2.5}$ exposure and associated premature mortality in China during 2005–2015, Environ. Int., 132, 105111, https://doi.org/10.1016/j.envint.2019.105111, 2019.

Zhou, Z. H. and Feng, J.: Deep forest, Natl. Sci. Rev., 6, 74–86, 2019.

Zhong, J., Zhang, X., Gui, K., Wang, Y., Che, H., Shen, X., Zhang, L., Zhang, Y., Sun, J., and Zhang, W.: Robust prediction of hourly $PM_{2.5}$ from meteorological data using LightGBM, Natl. Sci. Rev., 8, nwaa307, https://doi.org/10.1093/nsr/nwaa307, 2021.