Combining machine learning and numerical simulation for high-resolution $PM_{2.5}$ concentration forecast

Jianzhao Bi[1,**], K. Emma Knowland[2,3], Christoph A. Keller[2,3], Yang Liu[4,*]

[1]Department of Environmental & Occupational Health Sciences, University of Washington, Seattle, Washington 98195, United States

[2]NASA Goddard Space Flight Center, Greenbelt, Maryland 20771 , United States

[3]Universities Space Research Association, Columbia, Maryland 21046 , United States

[4]Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, Georgia 30322, United States

Corresponding Authors

[*]Mailing Address: Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA. E-mail: yang.liu@emory.edu.

[**]Mailing Address: Department of Environmental & Occupational Health Sciences, University of Washington, 4225 Roosevelt Way NE, Seattle, WA 98105, USA. E-mail: jbi6@uw.edu.

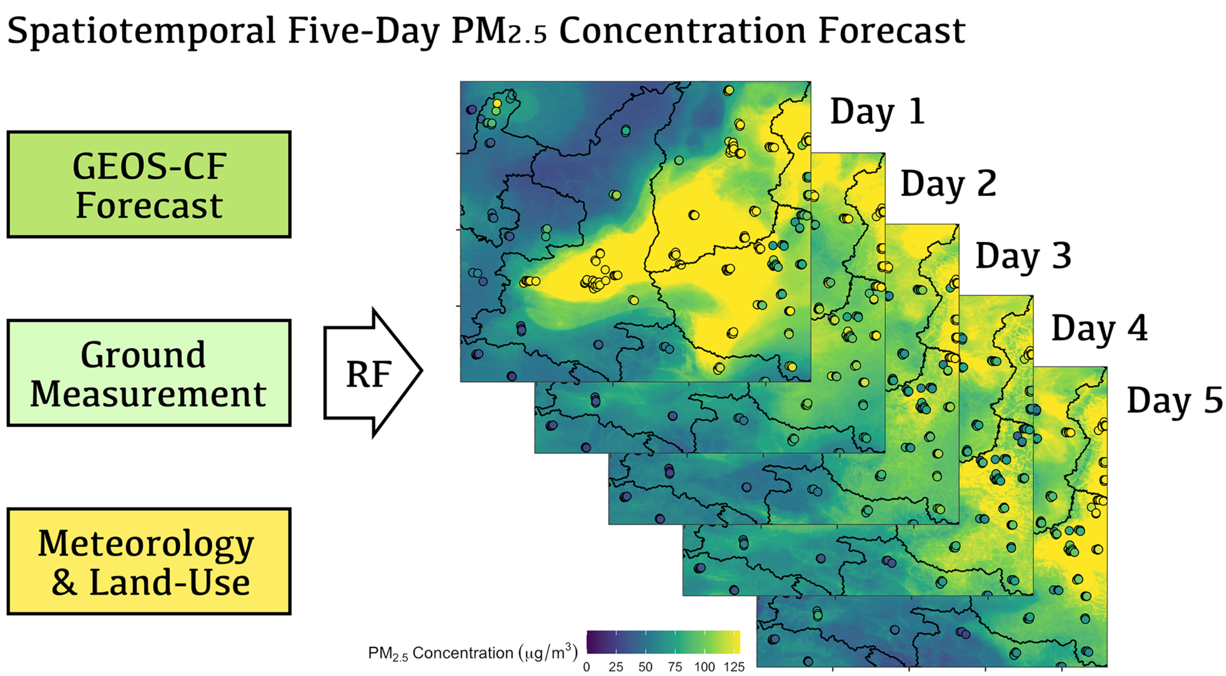
Abstract

Forecasting ambient $PM_{2.5}$ concentrations with spatiotemporal coverage is key to alerting decision-makers of pollution episodes and preventing detrimental public exposure, especially in regions with limited ground air monitoring stations. The existing methods either rely on chemical transport models (CTMs) to forecast spatial distribution of $PM_{2.5}$ with nontrivial uncertainty or statistical algorithms to forecast $PM_{2.5}$ concentration time-series at air monitoring locations without continuous spatial coverage. In this study, we developed a $PM_{2.5}$ forecast framework by combining the robust Random Forest algorithm with a publicly accessible global CTM forecast product - NASA's Goddard Earth Observing System "Composition Forecasting" (GEOS-CF), providing spatiotemporally continuous $PM_{2.5}$ concentration forecasts for the next five days at a 1-km spatial resolution. Our forecast experiment was conducted for a region in Central China including the populous and polluted Fenwei Plain. The forecast for the next two days had overall validation $R^2$ of 0.76 and 0.64, respectively; the $R^2$ was around 0.5 for the following three forecast days. Spatial cross-validation showed similar validation metrics. Our forecast model, with validation normalized mean bias close to zero, substantially reduced the large biases in GEOS-CF. The proposed framework requires minimal computational resources compared to running CTMs at urban scales, enabling near-real-time $PM_{2.5}$ forecast in resource-restricted environments.

Keywords: Air pollution forecast; Chemical transport model; Near-real-time; Near-term; Random Forest; XGBoost

Synopsis: A spatiotemporal high-resolution model for five-day $PM_{2.5}$ concentration forecast was developed by incorporating chemical transport simulations into a machine learning algorithm.

TOC Graphic:



## 1. Introduction

Fine particulate matter with an aerodynamic diameter of 2.5 μm or smaller ($PM_{2.5}$) can be inhaled and deposit in lung alveoli. Epidemiological research has shown that $PM_{2.5}$ is detrimental and casually associated with morbidity and mortality related to different body systems, especially cardiovascular and respiratory systems.[1, 2] Exposure to ambient $PM_{2.5}$ was estimated to contribute to 3 million deaths and 83 million disability-adjusted life years (DALYs) globally in 2017.[3] Countries in Asia, *e.g.*, China and India, are among the regions with the highest ambient $PM_{2.5}$ concentrations in the world.[4] While comprehensive control policies have been implemented and air quality has since been improved in China from the early 2010s, ambient $PM_{2.5}$ concentration levels in some polluted regions are still above China's air quality standards and the World Health Organization (WHO) air quality guidelines.[5]

Near-term forecast of ambient $PM_{2.5}$ concentrations is key to alerting decision-makers of potential pollution episodes and preventing detrimental public exposure. Chemical transport models (CTMs) have been widely used to numerically forecast spatiotemporal $PM_{2.5}$ concentrations in the near term - from next hours to days.[6, 7] CTMs forecast $PM_{2.5}$ concentrations based on estimated emissions and simulated atmospherically physical and chemical processes. Well-known CTM forecast products include those derived from global CTMs, *e.g.*, the Copernicus Atmosphere Monitoring Service (CAMS)[8] and the National Aeronautics and Space Administration (NASA) Goddard Earth Observing System "Composition Forecasting" (GEOS-CF),[9] and from regional CTMs, *e.g.*, the Community Multiscale Air Quality Modeling System (CMAQ).[10, 11] However, the CTM forecast products are subject to large biases due to uncertainties in emission inventories, parameterization of physical and chemical processes, and initial and/or boundary conditions.[12] Efforts have been made to improve CTM-based $PM_{2.5}$ forecast data.[12-19] For instance, the ensemble approach utilizes multiple inputs of emission inventories and meteorological fields or multiple models to reduce random errors in $PM_{2.5}$ simulations.[12, 19] Assimilation techniques, *e.g.*, the variational (VAR) method (3D- and 4D-VAR) and the Kalman filter, have also been used to incorporate ground truth (*i.e.*, $PM_{2.5}$ observations) to reduce systematic biases in $PM_{2.5}$ simulations.[13, 15, 17] However, the improved CTM forecast products are still obviously deviated from the ground truth and are usually not able to provide high-resolution forecast data at the scale of a kilometer.[9, 20, 21] More importantly, the CTM-based methods are computationally intensive and expensive, thus less practical for routine $PM_{2.5}$ forecast in resource-restricted environments.

Machine learning algorithms, as novel statistical methods, have been increasingly used to forecast near-term $PM_{2.5}$ concentrations. The majority of these algorithms are designed to forecast temporal variations (*i.e.*, time-series) of $PM_{2.5}$ at individual air monitoring sites. A typical example is the recurrent neural network (RNN) and its variant, the long short-term memory network (LSTM).[22-24] Unlike the regular neural network, the RNN allows connections between nodes to form a directed graph along a time sequence, therefore to process a time-series of inputs. Other parametric or machine learning algorithms have also been applied in $PM_{2.5}$ time-series forecast.[25-32] The advantages of machine learning algorithms over the CTM-based methods include higher forecast accuracy and substantially lower computational resources needed. However, $PM_{2.5}$ time-series forecast at air monitoring locations alone is less informative as the monitors, mostly regulatory agency monitors located in urban centers, cannot well represent pollution variations in suburban and rural areas.[33, 34] Few attempts have been made to forecast spatiotemporal variations of $PM_{2.5}$ based on statistical algorithms.[35, 36] Specifically, Ma et al [35] proposed a geo-layer of $PM_{2.5}$ concentrations (a spatially interpolated concentration surface) and integrated it into LSTM to forecast $PM_{2.5}$ with spatiotemporally complete coverage. Lu et al [36] incorporated $PM_{2.5}$ time-series forecast at monitoring sites from a LSTM model into a 3D-VAR model to spatially extrapolate $PM_{2.5}$ concentrations. However, the existing studies have several limitations. First, the studies tended to forecast spatial $PM_{2.5}$ variations based on inaccurate spatial information, *e.g.*, geographical interpolation or CTM simulations. Exposure modeling studies have shown that statistical prediction of $PM_{2.5}$ based on ground observations and meteorological/land-use predictors can generate more accurate $PM_{2.5}$ spatial distribution than spatial interpolation or chemical simulation.[20] Second, the multi-stage modeling process would lead to error propagation which in turn increases overall modeling uncertainty.[37] Third, the forecast model training and validation processes were not rigorously designed in these previous studies, in which future $PM_{2.5}$ observations tended to be used to train the forecast model, thus improperly inflating the validation performance. A rigorous validation set should not include ground observations on and after the day for which the forecast is made.

In this study, we developed a near-term $PM_{2.5}$ forecast framework - with limited computational resources needed - by combining a robust machine learning algorithm with a publicly accessible global CTM forecast product. We aimed to utilize the machine learning framework to improve the CTM forecast product by incorporating ground truth. Given the limitations of the existing methods for $PM_{2.5}$ forecast, we opted to use the Random Forest (RF) algorithm, a widely used machine learning method for spatiotemporal $PM_{2.5}$ prediction,[38-41] as our forecast model. Unlike time-series forecast algorithms such as LSTM, we showed that RF can forecast $PM_{2.5}$

concentrations in regions without ground monitors in a unified modeling framework. The proposed framework provided spatiotemporally continuous $PM_{2.5}$ forecast data for the next five days (daily averages) at a spatial resolution of 1 km. We also designed model training and validation processes that can mimic real-world $PM_{2.5}$ forecast to minimize validation biases. We chose a region in Central China with a large population and one of the most polluted city clusters in terms of $PM_{2.5}$, Fenwei Plain, as our study domain. Unlike other polluted regions in China, *e.g.*, the Beijing-Tianjin-Hebei region, our study domain was less influenced by emergency air pollution response and control actions, hence the proposed forecast framework could be reliably validated.

## 2. Data and methods

### 2.1. Study domain and ground $PM_{2.5}$ observations

We collected daily $PM_{2.5}$ concentration measurements from regulatory air quality stations of the China National Environmental Monitoring Center (CNEMC, http://www.cnemc.cn). We pre-defined a 1-km modeling grid and calculated daily-level, 1-km $PM_{2.5}$ concentrations from the ground measurements by spatial aggregation. Figure 1(a) shows our study domain with the locations of the $PM_{2.5}$ monitoring sites (N of locations = 226). The study domain covered multiple central and western provinces of China, including (alphabetically) Gansu, Hebei, Henan, Hubei, Inner Mongolia, Ningxia, Shaanxi, Shanxi, and Sichuan. The Fenwei Plain was entirely covered. The population within the study domain was estimated to be 150 million in 2018 (https://landscan.ornl.gov/). There were 97038 daily $PM_{2.5}$ observations at 226 1-km grid cells from January 1st, 2019 to March 14th, 2020. The study domain had a mean $PM_{2.5}$ concentration of 50 $\mu g/m^3$ (standard deviation = 44 $\mu g/m^3$) in 2019.

### 2.2. CTM-based $PM_{2.5}$ forecast data

We acquired $PM_{2.5}$ forecast data from a publicly accessible global CTM database, GEOS-CF, as the baseline forecast data. GEOS-CF is a novel atmospheric composition and meteorology model, providing three-dimensional distributions of hourly-level, five-day forecast $PM_{2.5}$ concentrations at a spatial resolution of 25 km (https://gmao.gsfc.nasa.gov/).[9] While the current version of GEOS-CF is known to have nontrivial systematic bias in $PM_{2.5}$ forecast data due to model representation errors, inaccurate input data (meteorology and emission), and biases in chemical/physical processes, the spatial distribution of $PM_{2.5}$ is reasonably captured.[9] In this study, we used the surface-level (two-dimensional) GEOS-CF data and calculated daily mean $PM_{2.5}$ concentrations for the five forecast days based on the China Standard Time (CST) and interpolated the concentrations into the pre-defined 1-km grid by ordinary kriging. Due to the difference between CST and Coordinated Universal Time (UTC) based on which GEOS-CF reports the forecast, the first to fourth forecast days had complete 24-hour forecast data while the fifth day had 21-hour forecast data from 12 AM to 8 PM CST.

### 2.3. Forecast meteorological data

The surface-level meteorological parameters for the five forecast days were acquired from GEOS-CF as well, including total cloud area fraction (unitless), surface pressure (Pa), 10-m specific humidity (kg/kg), 10-m air temperature (K), total precipitation (kg/m$^2$/s), tropopause pressure based on blended estimate (Pa), surface skin temperature (K), 10-m eastward/northward wind (m/s), and planetary boundary layer height (m). We calculated daily averages of the meteorological parameters and interpolated them into the pre-defined 1-km grid by ordinary kriging. These meteorological parameters were used as spatiotemporally varying predictors in our forecast model.

Prior to the launch of the five-day forecast, GEOS-CF runs a historical segment for the previous 24 hours to have the best initial conditions for the forecast. These historical estimates of the recent global atmospheric composition and meteorology are constrained by meteorological observations.[9] In this analysis, we used the GEOS-CF historical data to build a "now-cast" model for model parameter tuning (see Section 2.5).

## 2.4. Land-use data

We used land-use parameters as two-dimensional, spatially varying predictors of our forecast model. The parameters included the LandScan ambient population in 2018 at a 900-m resolution (https://landscan.ornl.gov/), the Copernicus Climate Change Service (C3S) global land cover (LC) products in 2018 at a resolution of 0.002778° (approximately 300 m) (https://cds.climate.copernicus.eu/), and distances to the nearest primary and secondary roads extracted and computed from the OpenStreetMap (OSM) road network data (https://www.openstreetmap.org/). The original C3S LC types were reclassified and reprocessed as percentages (%) of vegetation cover, urban areas, bare areas, and water bodies. We under-sampled the parameters into the pre-defined 1-km grid to match with other variables.

## 2.5. Forecast model training and prediction

Figure 1(b) shows the workflow of our forecast modeling and validation processes. The forecast framework was based on the RF algorithm, a widely used algorithm providing satisfactory $PM_{2.5}$

predictions with little configuration.[38-41] The RF algorithm constructs multiple decision trees to recover the non-linear relationships between the $PM_{2.5}$ concentration and its predictors and returns the mean prediction of $PM_{2.5}$ from the individual trees as the final prediction result. We focused on two major hyperparameters of RF: (1) the number of decision trees ($n_{tree}$) and (2) the number of predictors randomly tried at each split ($m_{try}$). We built a current-day ("now-cast") $PM_{2.5}$ prediction model for hyperparameter tuning (we note that this was not a forecast model; this "now-cast" model was only used for hyperparameter tuning). In the current-day model, ground $PM_{2.5}$ observations were used as the dependent variable and the same-day GEOS-CF meteorological variables and temporally invariant land-use parameters were used as predictors. We determined the values of the hyperparameters capable of minimizing the out-of-bag (OOB) error of the current-day model. Specifically, $n_{tree}$ and $m_{try}$ were determined to be 500 and 4, respectively. Following with previous studies,[39, 40] we relied on RF variable importance for predictor selection. The RF variable importance measures explain the relative importance and contribution of predictors. In this study, we opted to use the permutation variable importance defined to be the decrease in model performance when a single predictor's values are randomly shuffled. We excluded predictors with importance values close to zero and substantially smaller than other predictors' values, including percentages of bare areas and water bodies. These two predictors were spatially homogeneous at the monitoring locations within our study domain, thus minimally contributing to model performance. Table S1 lists the final predictors used to build the RF-based forecast model. The RF algorithm was based on the R (Ver. 4.0.2) package "ranger" (Ver. 0.12.1).[42]

The forecast model training process should mimic the real-world forecast scenario without future data included as the training sample. Therefore, we built the forecast model for each day individually on a rolling basis (as opposed to merging all training data together in a single model). There are two major forecast model features: (1) the forecast day, *i.e.*, for which day the forecast $PM_{2.5}$ concentrations are generated (from the first to fifth following days), and (2) the rolling period, *i.e.*, how many previous days' training data are included (we tested 10-, 30-, 60-, and 90-day rolling periods). For example, on the current day (Day 0), we aimed to forecast the next day's (Day 1) $PM_{2.5}$ concentrations when the rolling period was set to be 10 days. In this case, for model training, we matched the $PM_{2.5}$ observations on Day 0 with the GEOS-CF $PM_{2.5}$ and meteorological forecast data generated on the previous day (Day -1) for Day 0 and repeated this matching process for the 10-day rolling period from Day -9 to Day 0 (using GEOS-CF $PM_{2.5}$ and meteorological forecast data generated on Day -10 to Day -1); for model prediction, we then used the GEOS-CF $PM_{2.5}$ and meteorological forecast data generated on Day 0 for Day 1 to calculate $PM_{2.5}$ concentrations on Day 1 as the forecast results. The model building process is summarized in Table 1. We determined the rolling period to be 60 days for our forecast model as it allowed the model to have substantially higher forecast performance than those with shorter rolling periods, while the improvement in forecast performance was minimal for a longer rolling period (Tables S2 and S3).

We spatially interpolated the $PM_{2.5}$ observations on the current day by ordinary kriging to create a $PM_{2.5}$ convolutional layer and treated it as an additional spatiotemporal predictor. The $PM_{2.5}$ convolutional layer is a commonly used predictor of $PM_{2.5}$ exposure in previous modeling studies. [39, 43] It reflects the interpolated $PM_{2.5}$ concentrations generated with nearby observations, allowing the prediction model to account for spatial autocorrelation of $PM_{2.5}$. It is worth clarifying that the "convolutional layer" here is different from a similar term in deep convolutional neural networks (CNNs). Instead of a neural-network structure of CNN, our $PM_{2.5}$ convolutional layer is a two-dimensional $PM_{2.5}$ concentration surface generated before the modeling stage and was used as a model predictor. By using the $PM_{2.5}$ convolutional layer, we hypothesized that spatial variations in $PM_{2.5}$ on the current day were correlated with the variations on the forecast day, thus contributing to improved forecast performance.

Given that $PM_{2.5}$ prediction models based on statistical methods are not designed to predict extreme pollution events originated outside the study domain, *e.g.*, dust storms from northwest China in our case, we removed *a priori* the training and prediction data potentially associated with these extreme events. We adopted the Ultraviolet Aerosol Index (UVAI) from the TROPOspheric Monitoring Instrument (TROPOMI) onboard the Sentinel-5 Precursor satellite (http://www.tropomi.eu/) to identify extreme dust events within our study domain which typically occurred in spring. The TROPOMI UVAI is calculated based on wavelength dependent changes in Rayleigh scattering in the UV spectral range where ozone absorption is limited, which is able to track episodic aerosol plumes from dust outbreaks, volcanic ash, and biomass burning.[44] After checking the UVAI distributions on days with potential dust events, we determined an empirical UVAI threshold level of 0.5 (unitless) and removed the training and prediction data with UVAI values above the threshold (less than 1.2% of the data were removed). A sensitivity analysis for the UVAI threshold (different values around 0.5) showed that the identified training and prediction data associated with extreme dust events were robust (data not shown). Using UVAI to fully identify dust events associated with increased ground

$PM_{2.5}$ concentrations is challenging due to two reasons: (1) as TROPOMI only provides a single snapshot of UVAI each day, it is not able to reflect the evolution of dust plumes within a day; (2) UVAI captures aerosol plumes over the entire atmospheric column, which may sometimes be less correlated with ground $PM_{2.5}$. We identified a substantial dust storm that occurred within our study domain during the week of May 12$^{th}$, 2019, which was not fully captured by TROPOMI UVAI but significantly affected the performance of our forecast model. Therefore, we removed all training and prediction data on that week (from May 12$^{th}$ to 18$^{th}$, 2019) from our forecast process.

We deployed our forecast framework on a personal computing platform with 8 virtual central processing unit (CPU) cores (Intel® Xeon® CPU @ 2.00 GHz). Conducting one-day forecast with a 60-day rolling period in our study domain took approximately 5 seconds, which was negligible compared to generating CTM-based $PM_{2.5}$ forecast.

## 2.6. Forecast model validation

We validated our forecast model for each day by comparing the forecast predictions with ground $PM_{2.5}$ observations. The validation was out-of-sample because the $PM_{2.5}$ observations on the forecast days were not included in the training process.

With the out-of-sample validation dataset, we designed three validation schemes: (1) an *overall* validation with all validation sample over the entire modeling period to reflect the overall forecast performance, (2) a *site-specific* validation to summarize forecast performance for individual monitoring sites (at the 1-km grid cells), and (3) a *day-specific* validation to summarize forecast model performance for individual days over the modeling period.

We used the out-of-sample coefficient of determination ($R^2$), root-mean-square error (RMSE), mean absolute percentage error (MAPE), and normalized mean bias (NMB) as validation metrics. Eq. S1 to S4 show the formulae of these metrics. $R^2$ and RMSE are commonly used metrics in $PM_{2.5}$ exposure prediction; reporting them facilitates the comparison of our model performance with other work. MAPE, with standardized values, can improve the comparability of forecast performance among sites and days with different $PM_{2.5}$ concentration levels. NMB can reflect the direction of the forecast bias.

Additionally, we performed 10-fold spatial cross-validation (CV) to evaluate the forecast performance in regions without ground air monitors. The spatial CV randomly split the ground monitors into 10 approximately equal-sized groups; one group was treated as the test set in which the $PM_{2.5}$ measurements were withheld from the forecast modeling process as well as the calculation of $PM_{2.5}$ convolutional layers, while the other nine groups were treated as the training set. This procedure was repeated 10 times (*i.e.*, for each group). We used the same validation metrics for spatial CV.

### 2.7. Auxiliary analyses

In addition to forecasting $PM_{2.5}$ concentrations as numerical values, we examined our model's ability to forecast $PM_{2.5}$ pollution categories. Based on China's air quality standards,[45] we classified the $PM_{2.5}$ pollution categories as clean (24-hour average < 75 μg/m$^3$), moderate pollution (75-150 μg/m$^3$), and heavy pollution (> 150 μg/m$^3$). The categorical forecast was performed by RF with the same set of predictors. We reported the accuracy of the categorical forecast with two metrics, positive predictive value (PPV), *i.e.*, the probability that following a positive forecast result (clean, moderate pollution, or heavy pollution), that day will truly have that specific pollution level, and negative predictive value (NPV), *i.e.*, the probability that following a negative forecast result, that day will truly not have that specific pollution level. PPV and NPV are more intuitive than sensitivity and specificity for the public to understand the categorical forecast accuracy. Eq. S5 and S6 show the formulae of the two metrics.

Furthermore, we assessed how the spatial resolution of predictors affected our forecast model performance by aggregating the 1-km predictor values to 25-km means (*i.e.*, at the original GEOS-CF resolution) centering around the ground monitoring locations (for model training) and the centers of a 25-km grid we created (for model prediction). We compared the overall validation performance and the forecast predictions of the model with 25-km predictors to those with 1-km predictors. The 1-km and 25-km models shared the same ground $PM_{2.5}$ measurements as the dependent variable and validation set.

We also examined another tree-based machine learning algorithm, eXtreme Gradient Boosting (XGBoost), as a reference algorithm. XGBoost has been used in high-resolution $PM_{2.5}$ exposure prediction with satisfactory prediction accuracy.[46] We used the same set of predictors for XGBoost. We tuned three major hyperparameters of XGBoost based on cross-validation to obtain an optimal model, including the number of trees, maximum depth of a tree, and learning rate ($\eta$). The learning rate is related to a technique to slow down the learning in the boosting process to prevent overfitting, by applying a weighting factor for the residual error corrections by new trees when added to the model. The number of trees, leaning rate, and maximum depth of a tree were determined to be 500, 0.1, and 2, respectively. The algorithm comparison was conducted for November 2019 with the highest forecast performance for both the RF and XGBoost models. The XGBoost algorithm was based on the R package "xgboost" (Ver. 1.4.1.1).

# 3. Results

## 3.1. Overall validation and spatial CV performance

Table 2 shows the overall model performance for five-day $PM_{2.5}$ forecast over a one-year validation period from March 11th, 2019 to March 10th, 2020. We chose this validation period as it was the only period with a whole calendar year's data allowing a fair comparison among the five forecast days (*i.e.*, after ruling out the data over the first 60-day rolling period; otherwise, the validation period would be less than a year, which was not representative of annual variations of $PM_{2.5}$). Table 2 also compares the performance of our RF-based forecast with the original GEOS-CF forecast. In general, our RF-based forecast model outperformed the GEOS-CF model for all five forecast days, in which the first two days had substantially better performance with a validation $R^2$ of 0.76 (over 0.56 of GEOS-CF) on the first day and 0.64 (over 0.56) on the second day. Also, even though the original GEOS-CF forecast data had large biases with large validation RMSE, MAPE, and NMB, our RF-based forecast model well corrected the biases with considerably smaller values of the validation metrics. Moreover, as expected, our forecast model performance decreased with smaller $R^2$ and larger RMSE, MAPE, and NMB when forecasting the $PM_{2.5}$ concentrations over a longer term.

Table 3 shows the forecast performance of the 10-fold spatial CV for five-day $PM_{2.5}$ forecast over a one-year validation period from March 11th, 2019 to March 10th, 2020. The spatial CV was based on the same validation dataset as the overall validation. Even though the ground monitoring locations where the validation was performed were withheld, the spatial CV could reach a comparable performance to the overall validation (Table 2) with slightly lower $R^2$, slightly higher RMSE and MAPE, and similar NMB (close to zero). Meanwhile, the spatial CV performance was better than the performance of the original GEOS-CF model for all five forecast days, especially for the first two days.

Figure 2 summarizes the variable importance values of our forecast models for the five forecast days. The current-day $PM_{2.5}$ convolutional layer and GEOS-CF $PM_{2.5}$ forecast data were the top-two important variables for the first and second forecast days. On the following days, while the GEOS-CF $PM_{2.5}$ was still among the top, the importance of the convolutional layer decreased. The decrease in importance is expected as the current-day $PM_{2.5}$ concentrations tended to have weaker correlations with the concentrations on the following days. We also found that the forecast meteorological variables had higher importance values than the land-use variables, showing the larger contributions of these spatiotemporal variables.

Figure 3 exemplifies the $PM_{2.5}$ spatial distributions generated by our forecast model. Figure 3(a) shows the $PM_{2.5}$ concentrations from January 25th to 29th forecasted on January 24th, 2020. January 25th, 2020 was the start of the holiday week of the Lunar New Year in China. This date was also right after the lockdown of Wuhan (outside the study domain) due to the outbreak of novel coronavirus "COVID-19". The first day of the Lunar New Year holiday week, January

$25^{th}$, appeared to have higher $PM_{2.5}$ concentrations possibly associated with increased human activities and fireworks. The concentration levels then decreased over the following days. As expected, the high-level concentrations tended to be in and around the populous Fenwei Plain. Compared to the ground observations, our forecast data were shown to well capture the spatiotemporal variations of $PM_{2.5}$ over the period. Figure 3(b) shows the $PM_{2.5}$ concentrations from May $12^{th}$ to $16^{th}$ forecasted on May $11^{th}$, 2019. This is a negative example as the forecast data were not able to capture the strong dust storm event that occurred in the western part of our study domain (Gansu, Ningxia, and Shaanxi) during the period, when the ground observations appeared to be high. This example illustrates our forecast model's limitation to capture sudden extreme events that originated outside the domain.

### 3.2. Site-specific and day-specific validation performance

The site- and day-specific validation performance for the five forecast days, with the comparison to the GEOS-CF forecast performance, is shown in Figure 4. Tables S4 and S5 summarize the site- and day-specific validation metrics, respectively. Our forecast model was shown to substantially improve the forecast accuracy and precision of the original GEOS-CF forecast data.

For the site-specific validation (Figure 4(a)), our forecast model had higher $R^2$ for the first two forecast days (with a median > 0.7 for the first day and > 0.6 for the second day) over the GEOS-CF forecast model (with medians around 0.5); the validation $R^2$ values of the two models were comparable for the following days (with medians around or below 0.5). The interquartile ranges (IQR) of $R^2$ of our model were narrower for all five forecast days, indicating the robustness of the model. Our forecast model, with considerably lower RMSE (with medians < 30 $\mu g/m^3$), MAPE (with medians < 50%), and NMB (with medians around zero), corrected the large biases in the GEOS-CF data.

For the day-specific validation (Figure 4(b)), our forecast model had higher $R^2$ than the GEOS-CF forecast model for all five forecast days. The day-specific validation had wider IQRs of $R^2$ than the site-specific validation, indicating a greater challenge of our model to forecast spatial variability of $PM_{2.5}$ than its temporal variability, aligning with previous RF-based "now-cast" models.[39, 40] As in the site-specific validation, our model, with substantially lower RMSE (with medians < 20 μg/m$^3$), MAPE (with medians < 50%), and NMB (with medians close to zero), corrected the large biases in the GEOS-CF data.

Figure 5(a) shows the day-specific validation MAPE values with daily-mean $PM_{2.5}$ concentrations (using the first forecast day as an example). The daily MAPE variation displayed a pattern: MAPE tended to increase right after a sudden decrease in $PM_{2.5}$ concentrations. Figures 5(b) and (c) show the GEOS-CF wind speeds and directions (wind roses) within the study domain on days with validation $R^2$ above its 95$^{th}$ percentile (*i.e.*, good forecast performance) and below its 5$^{th}$ percentile (*i.e.*, poor forecast performance), respectively. The wind roses indicate that when the forecast models had an unsatisfactory performance, the dominant wind direction tended to be northeast with higher wind speeds. In comparison, there was not an obvious dominant wind direction when the models had a good performance. The association between sudden decreases in $PM_{2.5}$ and gusts of high-speed, northeast winds indicates that the northeast winds might bring relatively clean air to the study domain, therefore rapidly and temporarily eliminating $PM_{2.5}$ pollution. This result reflects a reduced forecast ability of our framework for sudden decreases in $PM_{2.5}$ resulting from wind elimination originated outside the domain.

### 3.3. Categorical forecast performance

Table 4 shows the forecast performance for categorical pollution levels (clean, moderate pollution, and heavy pollution) as well as the comparison between the original GEOS-CF and our RF-based forecast models. The original GEOS-CF model could not forecast well both the moderate and heavy pollution categories due to their large biases (with extremely low PPVs). In comparison, our RF-based data substantially improved the forecast accuracy for both categories with higher PPVs. The corresponding NPVs for the clean category increased as well. The clean category had the largest number of training sample (N = ~66000; the number varied for different forecast days) with high PPVs (~90%) and NPVs (~70 - 80%) for all five forecast days. With considerably fewer training sample, the moderate- (N = ~9800) and heavy-pollution (N = ~2300) categories had lower PPVs with decreased performance for longer forecast days. The NPVs for these two pollution categories were above 90% for all five forecast days.

### 3.4. Spatial resolution of predictors

Table S6 shows the overall model performance for five-day $PM_{2.5}$ forecast with 25-km predictors over a one-year validation period from March 11th, 2019 to March 10th, 2020. Figure S1 shows an example of spatial forecast concentrations derived with 1-km and 25-km predictors (the next-day $PM_{2.5}$ concentrations forecasted on January 24th, 2020). The overall validation performance was not substantially affected by the coarser 25-km resolution, with slightly lower $R^2$ and higher RMSE, MAPE, and NMB than the 1-km metrics. This comparison indicates that the original resolution of the spatiotemporal GEOS-CF variables might limit our model performance even after we interpolated them to 1-km. However, the forecast concentration surfaces exhibited different spatial patterns, where the 1-km concentration surface reflected substantially finer details of $PM_{2.5}$ distribution (associated with elevation, traffic, *etc*.) because the model took much greater advantage of high-resolution land-use information.

### 3.5. Comparison with XGBoost

Table S7 compares the forecast performance of the RF and XGBoost models in November 2019. Both models had similar validation $R^2$ and MAPE. Although RF slightly outperformed XGBoost, the differences between the two algorithms were not meaningful. Therefore, we expect these two tree-based algorithms can be interchangeable for our proposed forecast framework. We opted to use the RF algorithm due to its easy configuration with fewer major hyperparameters and its ability to provide robust predictions without much tuning effort.

## 4. Discussion

In this study, we proposed a RF-based framework for the near-term (next five days), daily-mean $PM_{2.5}$ concentration forecast at a 1-km spatial resolution. We also designed model training and validation processes that can mimic the real-world forecast scenario to minimize validation biases. All input data of our forecast framework are publicly accessible, including ground $PM_{2.5}$ observations, GEOS-CF $PM_{2.5}$ and meteorological forecast data, and land-use parameters. The forecast framework requires minimal computational resources and can be deployed in personal computing platforms. While the framework was evaluated in China with a satisfactory number of regulatory air monitoring stations in this study, we expect that it can also be deployed in resource-restricted environments in conjunction with a growing number of ground measurements from low-cost air quality monitors (when rigorously calibrated).[34] We note that our framework provides near-real-time rather than real-time forecast as the forecast product is generated at the end of each day when ground observations are collected and reported. However, as our framework provides rapid five-day forecast (at a scale of seconds for our study domain), the potential influence on heavy pollution awareness and response due to this level of delay is minimal.

The GEOS-CF $PM_{2.5}$ forecast data had large systematic biases as shown in this study (Table 2) and a previous evaluation study for a number of reasons, including model representation errors, uncertainties in the meteorology, and biases arising from errors in the treatment of emissions, deposition, or atmospheric chemistry.[9] The overall, site-specific, and day-specific validations showed that our statistical framework substantially improved the GEOS-CF data and generated acceptable $PM_{2.5}$ forecast concentrations, especially for the first two forecast days (Table 2 and Figure 4). While the third to fifth forecast days had comparable validation $R^2$ with the original GEOS-CF model, the large biases in the GEOS-CF data were well corrected. The spatial CV showed similar validation metrics to the overall validation (Table 3), indicating that our forecast framework was able to provide reliable forecast results in regions without ground monitors. This is a unique advantage of our RF-based framework over the widely adopted time-series forecast methods such as RNN and LSTM, which perform air pollution forecast only at ground monitoring locations based on their historical measurements. Intuitively, our modeling framework can be seen as a statistical "calibration" for the GEOS-CF forecast product by building a statistical model with "gold-standard" $PM_{2.5}$ observations as the dependent variable and the uncertain GEOS-CF forecast data as an independent variable with additional meteorological and land-use parameters as covariates, the concept of which is similar to low-cost air monitor calibration.[47]

The variable importance rankings suggested that the GEOS-CF $PM_{2.5}$ forecast data were always among the top important predictors (Figure 2), indicating that this product, although biased, was the key input of our forecast model as it provided meaningful information regarding $PM_{2.5}$ spatial distribution. We opted to use the forecast data from the GEOS-CF model with a relatively coarse spatial resolution because it was openly accessible with global coverage. We showed that the forecast concentrations greatly benefited from the interpolated GEOS-CF predictors at a higher spatial resolution, where detailed spatial patterns of $PM_{2.5}$ could be reflected more clearly (Figure S1). Meanwhile, we also expect that a regional CTM model at a higher spatial resolution, with proper boundary conditions, may further improve our forecast performance. According to the importance rankings, the current-day $PM_{2.5}$ convolutional layer contributed to an improved forecast performance, especially for the first two forecast days. This finding proves that the $PM_{2.5}$ convolutional layer is not only informative for the same-day prediction as shown in the previous studies,[39, 43] but for the near-term forecast as well (due to the correlations between the current-day and future $PM_{2.5}$ concentrations).

Categorical pollution levels are more intuitive than continuous concentrations for public awareness and emergency response to air pollution. Although the original GEOS-CF product had large biases in forecasting categorical $PM_{2.5}$ levels (clean, moderate pollution, and heavy pollution), our forecast model was able to substantially improving the forecast, especially for the first two to three forecast days (Table 4). The clean-day forecast had the highest accuracy for all five forecast days (with PPVs close or greater than 90%) possibly because the majority of the training data were in this category. Similarly, the moderate- and heavy-pollution forecast had higher NPVs than PPVs because of the fewer training data in these categories. While the heavy-pollution forecast had a PPV of ~70% on the first forecast day, the forecast accuracy decreased quickly on the fourth and fifth days. This pattern indicates a greater challenge of our framework to forecast high-level pollution over a longer term, which is worth further improvements.

The rolling period, *i.e.*, the number of previous days on which the $PM_{2.5}$ measurements are included as the training sample, was a key forecast model feature. We found that although a longer rolling period was associated with an increased forecast performance (Table S2), the increase was marginal when the rolling period was greater than 60 days (Table S3). Hence, we suggest that the two-month rolling period was optimal for our study domain and period, offering satisfactory forecast performance while minimizing the number of training data included. When applying the framework to other regions and periods, the optimal value of the rolling period should be re-evaluated according to forecast accuracy.

CTMs, though with higher uncertainties resulting from inaccurate emission inventories, atmospherically physical and chemical processes, and initial and boundary conditions, have been the dominant tool to forecast near-term $PM_{2.5}$ concentrations with spatiotemporal complete coverage.[6, 7] Few statistical efforts have been made to build more accurate spatiotemporal forecast models based on the ground truth (i.*e.*, $PM_{2.5}$ observations). Recently, Ma et al [35] and Lu et al [36] proposed statistical/empirical methods to forecast spatiotemporally complete $PM_{2.5}$

concentrations. The advantages of our proposed forecast framework over these studies are two-fold. First, our machine learning framework can generate more reliable spatial distributions of $PM_{2.5}$ than the distributions generated by spatial interpolation (*e.g.*, the geo-layer in Ma et al [35]) and the spatial information provided by CTM (*e.g.*, the simulation method used in Lu et al [36]). This advantage has been proven in previous $PM_{2.5}$ prediction studies using statistical algorithms.[20] Second, we proposed a more rigorous model training process by building daily forecast models on a rolling basis. This strategy guaranteed that no future $PM_{2.5}$ observations (*i.e.*, observations beyond the current day when the forecast is conducted) were included as the training sample. In contrast, if the observations across the entire period are randomly separated into a training set and a test set, the training set is very likely to include some same-day observations from the test set. In that case, the validation performance may be improperly inflated as the same-day observations are likely to be informative of the forecast on this day (even if the same-day training and validation samples are not at the same monitoring locations, the training locations can still be informative if they are geographically proximate to the validation locations).

The major limitation of our forecast framework is the limited ability to forecast $PM_{2.5}$ associated with out-of-domain factors, *e.g.*, extreme dust storms from the desert regions north/northwest to the domain and the sudden pollution elimination process associated with strong northeast winds. Without proper indicators of these out-of-domain factors, statistical models alone can hardly capture the associated pollution variations.[48] While CTM is supposed to have the ability to forecast these physical processes, the global GEOS-CF model was shown to unsatisfactorily simulate these processes in this study. We expect the forecast data from regional CTMs at a finer spatial resolution with better emission information and more accurate physical simulation processes, *e.g.*, CMAQ, may help our framework better capture and forecast these sudden events. It is also worth exploring the use of outputs from trajectory models, *e.g.*, the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model,[49, 50] in improving the forecast of sudden events with our framework. Additionally, the spatial interpolation of GEOS-CF $PM_{2.5}$ and meteorological forecast parameters based on ordinary kriging (to oversample them to the 1-km resolution) may not accurately reflect small-scale, terrain-related variations in the parameters, especially in mountainous areas. However, the potential interpolation uncertainty should have a limited influence on the $PM_{2.5}$ forecast as the uncertainty is likely to be substantially smaller than the CTM-related uncertainty in these parameters. Additional effort is needed to further reduce the potential interpolation uncertainty.

In summary, this study is among the first to generate high-resolution (1-km), near-term (next five days), and near-real-time $PM_{2.5}$ forecast data based on a robust machine learning framework. While we showcased the forecast ability of our framework in a populated region of Central China with high-level $PM_{2.5}$ pollution, we expect that the framework can be generalized to other regions and for other air pollutants, *e.g.*, ozone and nitrogen dioxide, based on the same input data sources.[51] Our proposed framework with near-real-time forecast products holds promise for improved public awareness, policy development, and emergency response regarding detrimental air pollution exposure.

Supporting Information

Six equations, seven tables, and a figure, providing additional information regarding $PM_{2.5}$ forecast model evaluation methods and results.

Acknowledgments

This research was supported by the National Aeronautics and Space Administration (NASA) Modeling, Analysis and Prediction (MAP) Program (16-MAP16-0025). The contents of this publication do not necessarily reflect the views of NASA.

References


1. Bourdrel, T.; Bind, M.-A.; Béjot, Y.; Morel, O.; Argacha, J.-F., Cardiovascular effects of air pollution. *Archives of Cardiovascular Diseases* **2017,** *110*, (11), 634-642.
2. Lu, F.; Xu, D.; Cheng, Y.; Dong, S.; Guo, C.; Jiang, X.; Zheng, X., Systematic review and meta-analysis of the adverse health effects of ambient PM2.5 and PM10 pollution in the Chinese population. *Environmental Research* **2015,** *136*, 196-204.
3. Bu, X.; Xie, Z.; Liu, J.; Wei, L.; Wang, X.; Chen, M.; Ren, H., Global PM2.5-attributable health burden from 1990 to 2017: Estimates from the Global Burden of disease study 2017. *Environmental Research* **2021,** *197*, 111123.
4. Lim, C.-H.; Ryu, J.; Choi, Y.; Jeon, S. W.; Lee, W.-K., Understanding global PM2.5 concentrations and their drivers in recent decades (1998–2016). *Environment International* **2020,** *144*, 106011.
5. Liang, F.; Xiao, Q.; Huang, K.; Yang, X.; Liu, F.; Li, J.; Lu, X.; Liu, Y.; Gu, D., The 17-y spatiotemporal trend of PM2.5 and its mortality burden in China. *Proceedings of the National Academy of Sciences* **2020,** *117*, (41), 25601-25608.
6. Cheng, X. H.; Liu, Y. L.; Xu, X. D.; You, W.; Zang, Z. L.; Gao, L. N.; Chen, Y. B.; Su, D. B.; Yan, P., Lidar data assimilation method based on CRTM and WRF-Chem models and its application in PM2.5 forecasts in Beijing. *Science of the Total Environment* **2019,** *682*, 541-552.
7. Wu, C. B.; Li, K.; Bai, K. X., Validation and Calibration of CAMS PM2.5 Forecasts Using In Situ PM2.5 Measurements in China and United States. *Remote Sensing* **2020,** *12*, (22), 19.
8. Flemming, J.; Benedetti, A.; Inness, A.; Engelen, R. J.; Jones, L.; Huijnen, V.; Remy, S.; Parrington, M.; Suttie, M.; Bozzo, A.; Peuch, V. H.; Akritidis, D.; Katragkou, E., The CAMS interim Reanalysis of Carbon Monoxide, Ozone and Aerosol for 2003–2015. *Atmos. Chem. Phys.* **2017,** *17*, (3), 1945-1983.
9. Keller, C. A.; Knowland, K. E.; Duncan, B. N.; Liu, J.; Anderson, D. C.; Das, S.; Lucchesi, R. A.; Lundgren, E. W.; Nicely, J. M.; Nielsen, E.; Ott, L. E.; Saunders, E.; Strode, S. A.; Wales, P. A.; Jacob, D. J.; Pawson, S., Description of the NASA GEOS Composition Forecast Modeling System GEOS-CF v1.0. *Journal of Advances in Modeling Earth Systems* **2021,** *13*, (4), e2020MS002413.
10. Cheng, F. Y.; Feng, C. Y.; Yang, Z. M.; Hsu, C. H.; Chan, K. W.; Lee, C. Y.; Chang, S. C., Evaluation of real-time PM2.5 forecasts with the WRF-CMAQ modeling system and weather-pattern-dependent bias-adjusted PM2.5 forecasts in Taiwan. *Atmos Environ* **2021,** *244*, 17.
11. Sayeed, A.; Lops, Y.; Choi, Y.; Jung, J.; Salman, A. K., Bias correcting and extending the PM forecast by CMAQ up to 7 days using deep convolutional neural networks. *Atmos Environ* **2021,** *253*, 9.
12. Zhang, H.; Wang, J.; García, L. C.; Ge, C.; Plessel, T.; Szykman, J.; Murphy, B.; Spero, T. L., Improving Surface PM2.5 Forecasts in the United States Using an Ensemble of Chemical Transport Model Outputs: 1. Bias Correction With Surface Observations in Nonrural Areas. *J Geophys Res-Atmos* **2020,** *125*, (14), e2019JD032293.
13. Feng, S. Z.; Jiang, F.; Jiang, Z. Q.; Wang, H. M.; Cai, Z.; Zhang, L., Impact of 3DVAR assimilation of surface PM2.5 observations on PM2.5 forecasts over China during wintertime. *Atmos Environ* **2018,** *187*, 34-49.

14. June, N.; Vaughan, J.; Lee, Y.; Lamb, B. K., Operational bias correction for PM2.5 using the AIRPACT air quality forecast system in the Pacific Northwest. *Journal of the Air & Waste Management Association* **2021,** *71*, (4), 515-527.
15. Kong, Y. W.; Sheng, L. F.; Li, Y. P.; Zhang, W. H.; Zhou, Y.; Wang, W. C.; Zhao, Y. H., Improving PM2.5 forecast during haze episodes over China based on a coupled 4D-LETKF and WRF-Chem system. *Atmos. Res.* **2021,** *249*, 14.
16. Liang, Y. F.; Zang, Z. L.; Liu, D.; Yan, P.; Hu, Y. W.; Zhou, Y.; You, W., Development of a three-dimensional variational assimilation system for lidar profile data based on a size-resolved aerosol model in WRF-Chem model v3.9.1 and its application in PM2.5 forecasts across China. *Geosci. Model Dev.* **2020,** *13*, (12), 6285-6301.
17. Peng, Z.; Liu, Z. Q.; Chen, D.; Ban, J. M., Improving PM2.5 forecast over China by the joint adjustment of initial conditions and source emissions with an ensemble Kalman filter. *Atmos Chem Phys* **2017,** *17*, (7), 4837-4855.
18. Zheng, H. T.; Liu, J. G.; Tang, X.; Wang, Z. F.; Wu, H. J.; Yan, P. Z.; Wang, W., Improvement of the Real-time PM2.5 Forecast over the Beijing-Tianjin-Hebei Region using an Optimal Interpolation Data Assimilation Method. *Aerosol Air Qual Res* **2018,** *18*, (5), 1305-1316.
19. Ge, C.; Wang, J.; Reid, J. S.; Posselt, D. J.; Xian, P.; Hyer, E., Mesoscale modeling of smoke transport from equatorial Southeast Asian Maritime Continent to the Philippines: First comparison of ensemble analysis with in situ observations. *J Geophys Res-Atmos* **2017,** *122*, (10), 5380-5398.
20. Chu, Y.; Liu, Y.; Li, X.; Liu, Z.; Lu, H.; Lu, Y.; Mao, Z.; Chen, X.; Li, N.; Ren, M.; Liu, F.; Tian, L.; Zhu, Z.; Xiang, H., A Review on Predicting Ground PM2.5 Concentration Using Satellite Aerosol Optical Depth. *Atmosphere-Basel* **2016,** *7*, (10), 129.
21. Lightstone, S. D.; Moshary, F.; Gross, B., Comparing CMAQ Forecasts with a Neural Network Forecast Model for PM2.5 in New York. *Atmosphere-Basel* **2017,** *8*, (9), 161.
22. Huang, C. J.; Kuo, P. H., A Deep CNN-LSTM Model for Particulate Matter (PM2.5) Forecasting in Smart Cities. *Sensors* **2018,** *18*, (7), 22.
23. Liu, H.; Duan, Z.; Chen, C., A hybrid multi-resolution multi-objective ensemble model and its application for forecasting of daily PM2.5 concentrations. *Inf. Sci.* **2020,** *516*, 266-292.
24. Zhou, Y.; Chang, F.-J.; Chang, L.-C.; Kao, I. F.; Wang, Y.-S., Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. *J. Clean Prod.* **2019,** *209*, 134-145.
25. Mahajan, S.; Chen, L. J.; Tsai, T. C., Short-Term PM2.5 Forecasting Using Exponential Smoothing Method: A Comparative Analysis. *Sensors* **2018,** *18*, (10), 15.
26. Niu, M.; Wang, Y.; Sun, S.; Li, Y., A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term PM2.5 concentration forecasting. *Atmos Environ* **2016,** *134*, 168-180.
27. Qin, S.; Liu, F.; Wang, J.; Sun, B., Analysis and forecasting of the particulate matter (PM) concentration levels over four major cities of China using hybrid models. *Atmos Environ* **2014,** *98*, 665-675.
28. Bai, Y.; Li, Y.; Zeng, B.; Li, C.; Zhang, J., Hourly PM2.5 concentration forecast using stacked autoencoder model with emphasis on seasonality. *J. Clean Prod.* **2019,** *224*, 739-750.
29. Franceschi, F.; Cobo, M.; Figueredo, M., Discovering relationships and forecasting PM10 and PM2.5 concentrations in Bogotá, Colombia, using Artificial Neural Networks,

Principal Component Analysis, and k-means clustering. *Atmos. Pollut. Res.* **2018,** *9*, (5), 912-922.

30. Sun, W.; Li, Z. Q., Hourly PM2.5 concentration forecasting based on feature extraction and stacking-driven ensemble model for the winter of the Beijing-Tianjin-Hebei area. *Atmos. Pollut. Res.* **2020,** *11*, (6), 110-121.

31. Ventura, L. M. B.; Pinto, F. D.; Soares, L. M.; Luna, A. S.; Gioda, A., Forecast of daily PM2.5 concentrations applying artificial neural networks and Holt-Winters models. *Air Qual. Atmos. Health* **2019,** *12*, (3), 317-325.

32. Zhou, Y. L.; Chang, F. J.; Chang, L. C.; Kao, I. F.; Wang, Y. S.; Kang, C. C., Multi-output support vector machine for regional multi-step-ahead PM2.5 forecasting. *Science of the Total Environment* **2019,** *651*, 230-240.

33. Bi, J.; Stowell, J.; Seto, E. Y. W.; English, P. B.; Al-Hamdan, M. Z.; Kinney, P. L.; Freedman, F. R.; Liu, Y., Contribution of low-cost sensor measurements to the prediction of PM2.5 levels: A case study in Imperial County, California, USA. *Environmental Research* **2020,** *180*, 108810.

34. Bi, J.; Wildani, A.; Chang, H. H.; Liu, Y., Incorporating Low-Cost Sensor Measurements into High-Resolution PM2.5 Modeling at a Large Spatial Scale. *Environ Sci Technol* **2020,** *54*, (4), 2152-2162.

35. Ma, J.; Ding, Y.; Cheng, J. C. P.; Jiang, F.; Wan, Z., A temporal-spatial interpolation and extrapolation method based on geographic Long Short-Term Memory neural network for PM2.5. *J. Clean Prod.* **2019,** *237*, 117729.

36. Lu, X. C.; Sha, Y. H.; Li, Z. N.; Huang, Y. Q.; Chen, W. Y.; Chen, D. H.; Shen, J.; Chen, Y.; Fung, J. C. H., Development and application of a hybrid long-short term memory - three dimensional variational technique for the improvement of PM2.5 forecasting. *Science of the Total Environment* **2021,** *770*, 10.

37. Pu, Q.; Yoo, E. H., Ground PM2.5 prediction using imputed MAIAC AOD with uncertainty quantification. *Environmental Pollution* **2021,** *274*, 9.

38. Brokamp, C.; Jandarov, R.; Hossain, M.; Ryan, P., Predicting Daily Urban Fine Particulate Matter Concentrations Using a Random Forest Model. *Environ Sci Technol* **2018,** *52*, (7), 4173-4179.

39. Hu, X.; Belle, J. H.; Meng, X.; Wildani, A.; Waller, L. A.; Strickland, M. J.; Liu, Y., Estimating PM2.5 Concentrations in the Conterminous United States Using the Random Forest Approach. *Environ Sci Technol* **2017,** *51*, (12), 6936-6944.

40. Bi, J.; Belle, J. H.; Wang, Y.; Lyapustin, A. I.; Wildani, A.; Liu, Y., Impacts of snow and cloud covers on satellite-derived PM2.5 levels. *Remote Sens Environ* **2019,** *221*, 665-674.

41. Huang, K.; Bi, J.; Meng, X.; Geng, G.; Lyapustin, A.; Lane, K. J.; Gu, D.; Kinney, P. L.; Liu, Y., Estimating daily PM2.5 concentrations in New York City at the neighborhood-scale: Implications for integrating non-regulatory measurements. *Science of the Total Environment* **2019,** *697*, 134094.

42. Wright, M. N.; Ziegler, A., ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **2017,** *77*, (1), 1-17.

43. Di, Q.; Kloog, I.; Koutrakis, P.; Lyapustin, A.; Wang, Y.; Schwartz, J., Assessing PM2.5 Exposures with High Spatiotemporal Resolution across the Continental United States. *Environ Sci Technol* **2016,** *50*, (9), 4712-4721.

44. Kooreman, M. L.; Stammes, P.; Trees, V.; Sneep, M.; Tilstra, L. G.; de Graaf, M.; Stein Zweers, D. C.; Wang, P.; Tuinder, O. N. E.; Veefkind, J. P., Effects of clouds on the UV Absorbing Aerosol Index from TROPOMI. *Atmos. Meas. Tech.* **2020,** *13*, (12), 6407-6426.
45. Wang, S.; Hao, J., Air quality management in China: Issues, challenges, and options. *Journal of Environmental Sciences* **2012,** *24*, (1), 2-13.
46. Xiao, Q.; Chang, H. H.; Geng, G.; Liu, Y., An Ensemble Machine-Learning Model To Predict Historical PM2.5 Concentrations in China from Satellite Data. *Environ Sci Technol* **2018,** *52*, (22), 13260-13269.
47. Barkjohn, K. K.; Gantt, B.; Clements, A. L., Development and application of a United States-wide correction for PM2.5 data collected with the PurpleAir sensor. *Atmos. Meas. Tech.* **2021,** *14*, (6), 4617-4637.
48. Liou, N. C.; Luo, C. H.; Mahajan, S.; Chen, L. J., Why is Short-Time PM2.5 Forecast Difficult? The Effects of Sudden Events. *IEEE Access* **2020,** *8*, 12662-12674.
49. Stein, A.; Draxler, R. R.; Rolph, G. D.; Stunder, B. J.; Cohen, M.; Ngan, F., NOAA's HYSPLIT atmospheric transport and dispersion modeling system. *B Am Meteorol Soc* **2015,** *96*, (12), 2059-2077.
50. Li, Y.; Tong, D. Q.; Ngan, F.; Cohen, M. D.; Stein, A. F.; Kondragunta, S.; Zhang, X.; Ichoku, C.; Hyer, E. J.; Kahn, R. A., Ensemble PM2.5 Forecasting During the 2018 Camp Fire Event Using the HYSPLIT Transport and Dispersion Model. *J. Geophys. Res.-Atmos.* **2020,** *125*, (15), 19.
51. Malings, C.; Knowland, K. E.; Keller, C. A.; Cohn, S. E., Sub-City Scale Hourly Air Quality Forecasting by Combining Models, Satellite Observations, and Ground Measurements. *Earth and Space Science* **2021,** *8*, (7), e2021EA001743.

Table 1: The forecast model building process by matching ground $PM_{2.5}$ observations with the $PM_{2.5}$ convolutional layer and GEOS-CF forecast data ($PM_{2.5}$ pollution and meteorology forecast) as training and prediction data. N is the rolling period (N = 60 days). Day 0 is the present day, Day 1 is the next day, *etc*. The CTM running date is the day on which the CTM is run. The CTM forecast date is the day the forecast is made for.

<table>
<tr><th rowspan="2">Forecast Day</th><th rowspan="2">Date of $PM_{2.5}$ Convolutional Layer</th><th colspan="2">Training</th><th colspan="2">Prediction</th></tr>
<tr><th>CTM Running Date</th><th>$PM_{2.5}$ Observation & CTM Forecast Date</th><th>CTM Running Date</th><th>CTM Forecast Date</th></tr>
<tr><td>Day 1</td><td rowspan="5">Day 0</td><td>Day -N to -1</td><td rowspan="5">Day -(N-1) to 0</td><td rowspan="5">Day 0</td><td>Day 1</td></tr>
<tr><td>Day 2</td><td>Day -(N+1) to -2</td><td>Day 2</td></tr>
<tr><td>Day 3</td><td>Day -(N+2) to -3</td><td>Day 3</td></tr>
<tr><td>Day 4</td><td>Day -(N+3) to -4</td><td>Day 4</td></tr>
<tr><td>Day 5</td><td>Day -(N+4) to -5</td><td>Day 5</td></tr>
</table>

Table 2: The overall validation performance of our forecast framework (RF + GEOS-CF) and the original CTM forecast model (GEOS-CF) for the validation period of March 11th, 2019 to March 10th, 2020. The rolling period was 60 days.

| Forecast Day | N of Test Sample | $R^2$ | RMSE ($\mu g/m^3$) | MAPE (%) | NMB |
|---|---|---|---|---|---|
| RF + GEOS-CF | | | | | |
| Day 1 | 78378 | 0.76 | 18.70 | 34.3 | 0.003 |
| Day 2 | 78398 | 0.64 | 23.07 | 43.2 | 0.008 |
| Day 3 | 78158 | 0.56 | 25.48 | 46.8 | 0.005 |
| Day 4 | 78372 | 0.51 | 26.70 | 48.9 | -0.001 |
| Day 5 | 78372 | 0.47 | 27.81 | 52.4 | -0.003 |
| GEOS-CF | | | | | |
| Day 1 | 78378 | 0.56 | 140.76 | 278.1 | 2.23 |
| Day 2 | 78398 | 0.56 | 141.46 | 277.3 | 2.224 |
| Day 3 | 78158 | 0.53 | 145.55 | 279.4 | 2.246 |
| Day 4 | 78372 | 0.50 | 144.09 | 280.0 | 2.216 |
| Day 5 | 78372 | 0.45 | 141.87 | 278.8 | 2.139 |

Table 3: The 10-fold spatial CV performance of our forecast framework (RF + GEOS-CF) for the validation period of March 11th, 2019 to March 10th, 2020. The rolling period was 60 days.

| Forecast Day | N of Test Sample | $R^2$ | RMSE ($\mu g/m^3$) | MAPE (%) | NMB |
|---|---|---|---|---|---|
| Day 1 | 78378 | 0.74 | 19.47 | 36.7 | 0.001 |
| Day 2 | 78398 | 0.63 | 23.53 | 45.0 | 0.009 |
| Day 3 | 78158 | 0.55 | 25.82 | 48.4 | 0.006 |
| Day 4 | 78372 | 0.50 | 27.02 | 50.5 | 0.001 |
| Day 5 | 78372 | 0.46 | 28.08 | 54.0 | 0 |

Table 4: The categorical forecast performance metrics of our forecast framework (RF + GEOS-CF) and the original CTM forecast model (GEOS-CF), including positive predictive value (PPV) and negative predictive value (NPV), for clean (N of training sample = ~66000, varied on different forecast days), moderate-pollution (N of training sample = ~9800), and heavy-pollution categories (N of training sample = ~2300).

| Forecast Day | Clean | | Moderate Pollution | | Heavy Pollution | |
|---|---|---|---|---|---|---|
| | PPV | NPV | PPV | NPV | PPV | NPV |
| RF + GEOS-CF | | | | | | |
| Day 1 | 94.1% | 81.6% | 64.9% | 94.0% | 71.5% | 98.2% |
| Day 2 | 92.2% | 77.2% | 56.9% | 92.3% | 54.5% | 97.7% |
| Day 3 | 91.5% | 72.7% | 53.8% | 92.0% | 44.3% | 97.4% |
| Day 4 | 90.5% | 71.1% | 52.3% | 91.3% | 43.1% | 97.5% |
| Day 5 | 89.6% | 66.8% | 48.4% | 90.6% | 33.0% | 97.3% |
| GEOS-CF | | | | | | |
| Day 1 | 98.8% | 20.6% | 3.4% | 82.3% | 7.4% | 99.7% |
| Day 2 | 98.9% | 20.7% | 3.1% | 81.9% | 7.7% | 99.7% |
| Day 3 | 98.8% | 20.6% | 3.2% | 81.6% | 7.9% | 99.8% |
| Day 4 | 98.4% | 20.5% | 3.3% | 81.6% | 7.8% | 99.7% |
| Day 5 | 98.1% | 20.8% | 4.5% | 82.4% | 8.1% | 99.7% |

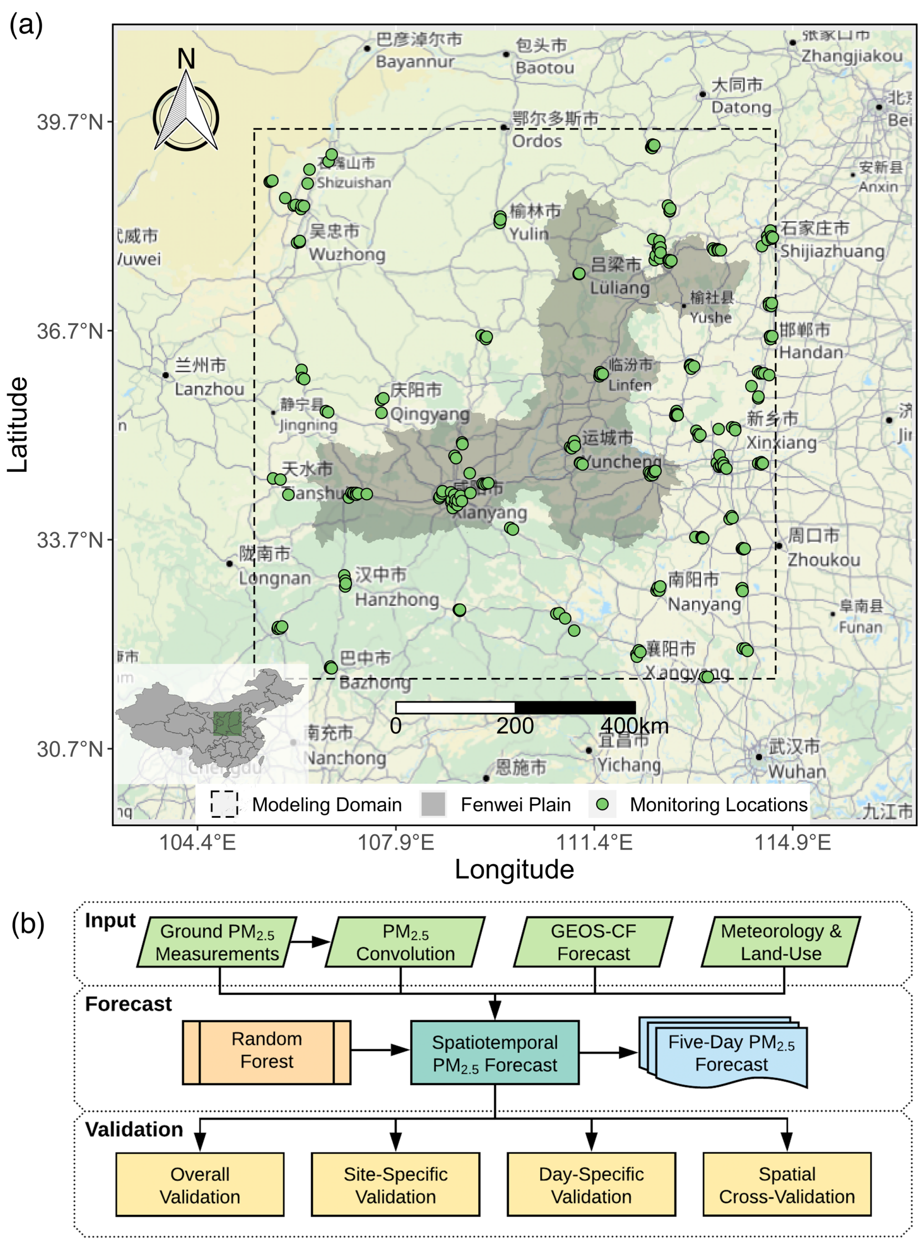



Figure 1: (a) Our study domain (the dashed box) with the locations of $PM_{2.5}$ monitoring sites (at the 1-km grid cells; N = 226); the shadow region shows the municipality boundary of the Fenwei Plain. (b) The workflow of our $PM_{2.5}$ forecast modeling and validation processes.

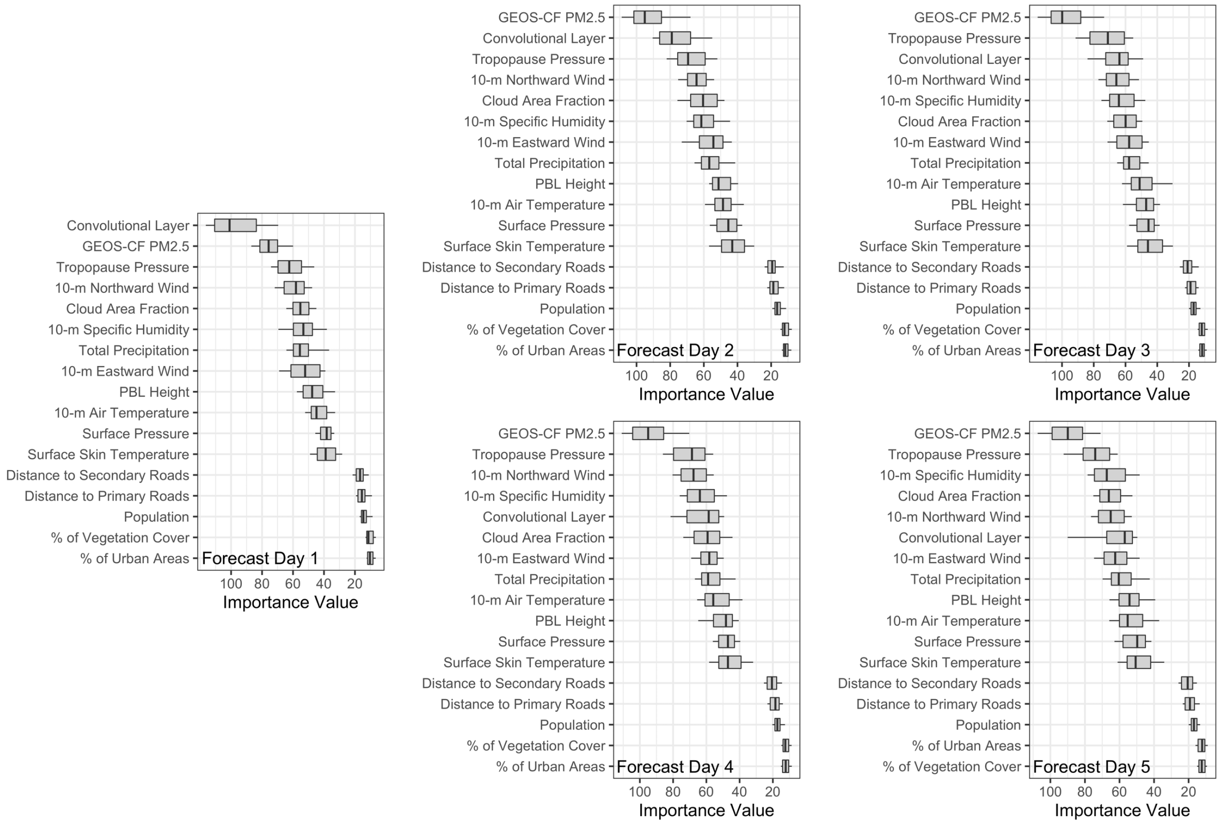



Figure 2: RF variable importance values for the five forecast days. The box plots summarize the importance values of the daily models from March 11th, 2019 to March 10th, 2020. The boxes represent the 25th and 75th percentile ranges; the whiskers represent the 10th and 90th percentile ranges; the bars within the boxes represent the 50th percentiles.

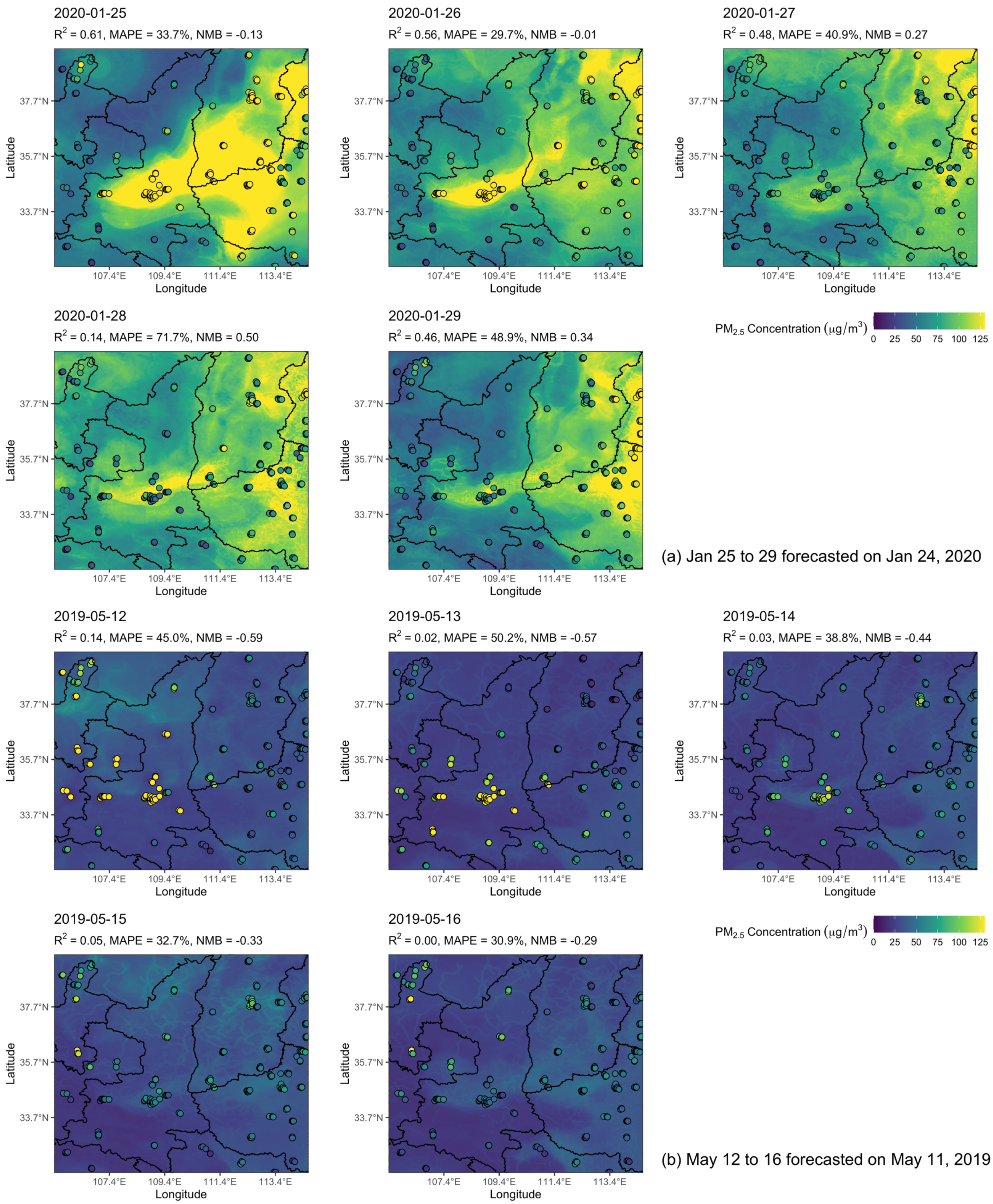



Figure 3: Spatial $PM_{2.5}$ forecast concentrations in two example periods: (a) January 25th to 29th forecasted on January 24th, 2020, and (b) May 12th to 16th forecasted on May 11th, 2019. The

colored dots show the observed $PM_{2.5}$ concentrations at the monitoring locations, which share the same color scheme with the forecast concentrations.

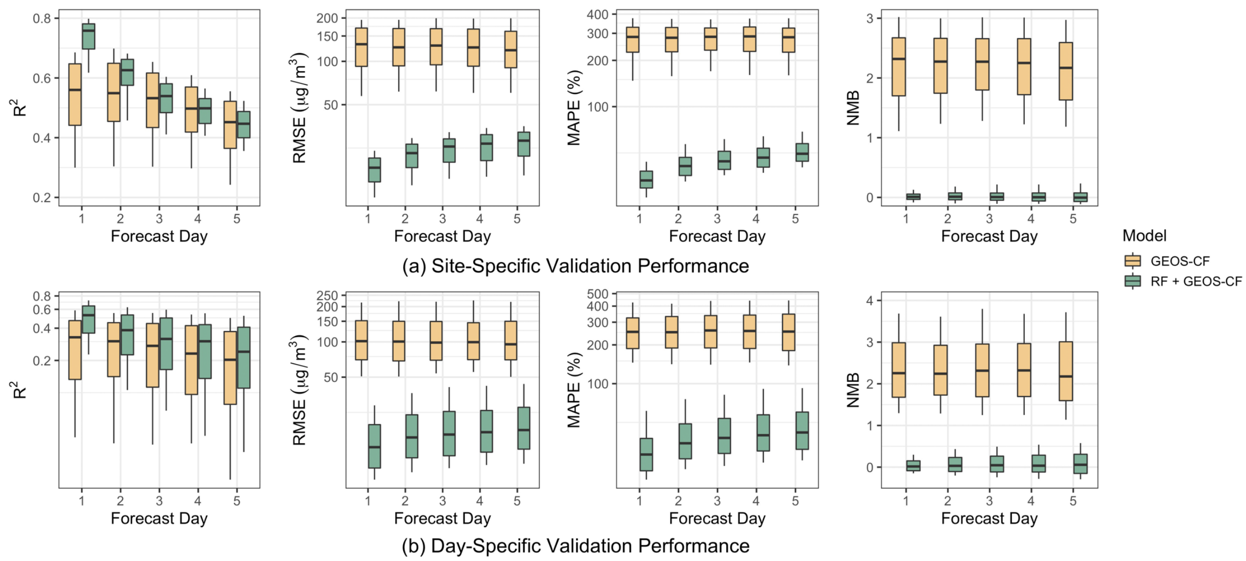



Figure 4: The (a) site-specific and (b) day-specific validation performance from March 11th, 2019 to March 10th, 2020. The boxes represent the 25th and 75th percentile ranges; the whiskers represent the 10th and 90th percentile ranges; the bars within the boxes represent the 50th percentiles. The RMSE and MAPE plots are on the log scale.

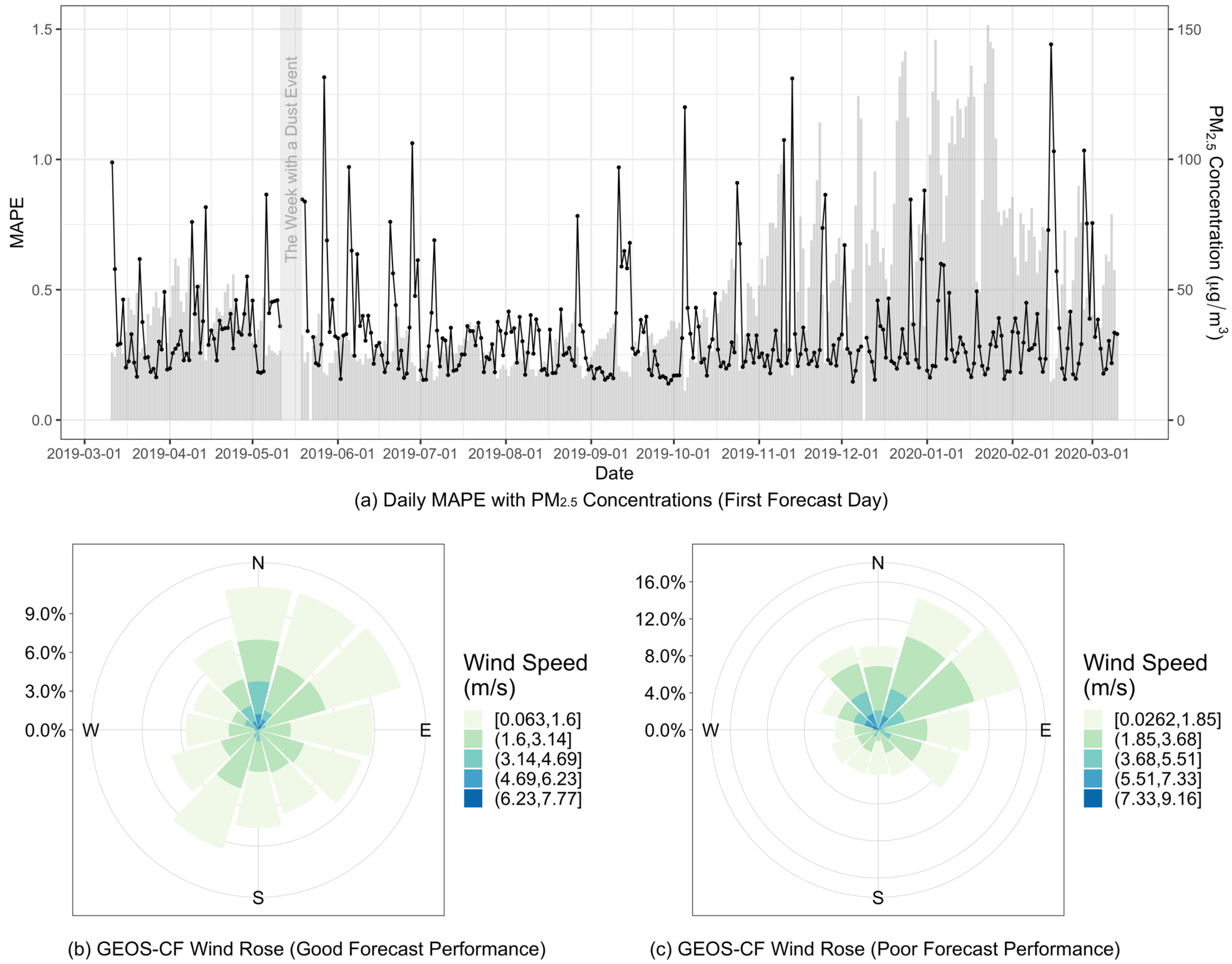



Figure 5: (a) Daily validation MAPE values (black dots) with domain-average $PM_{2.5}$ concentrations (grey bars) using the first forecast day as an example; (b) GEOS-CF wind rose on days with validation $R^2$ > its 95th percentile (good forecast performance); (c) GEOS-CF wind rose on days with validation $R^2$ < its 5th percentile (poor forecast performance).