Environmental Science & Technology

Article

Cite This: Environ. Sci. Technol. 2020, 54, 120−128

pubs.acs.org/est

Estimating Daily $PM_{2.5}$ and PM_{10} over Italy Using an Ensemble Model

Alexandra Shtein,[*,†] Itai Kloog,[†] Joel Schwartz,[‡] Camillo Silibello,[§] Paola Michelozzi,[∥] Claudio Gariazzo,[⊥] Giovanni Viegi,[#] Francesco Forastiere,[#,○] Arnon Karnieli,[¶] Allan C. Just,[▽] and Massimo Stafoggia[∥,♦]

[†]Department of Geography and Environmental Development, Ben-Gurion University of the Negev, Beer Sheva 8410501, Israel

[‡]Department of Environmental Health, Harvard T. H. Chan School of Public Health, Boston 02115, Massachusetts, United States

[§]ARIANET s.r.l., Milano 20128, Italy

[∥]Department of Epidemiology, Lazio Regional Health Service/ASL Roma 1, Rome 00147, Italy

[⊥]Occupational and Environmental Medicine, Epidemiology and Hygiene Department, Italian Workers' Compensation Authority (INAIL), Monte Porzio Catone (RM) 00078, Italy

[#]Institute for Biomedical Research and Innovation, National Research Council, Palermo 90146, Italy

[¶]Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Sede Boker Campus 84990, Israel

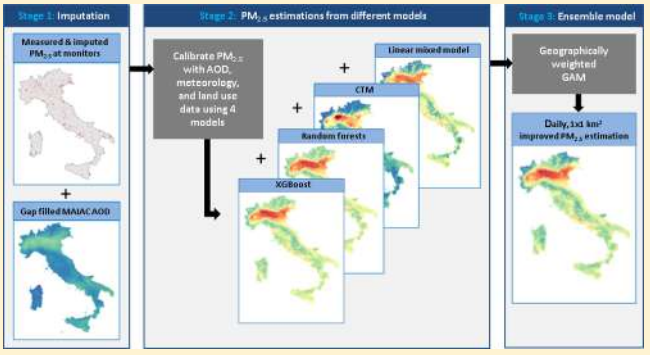
[▽]Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, New York 10029, United States

[○]Environmental Research Group, King's College, London SE1 9NH, U.K.

[♦]Institute of Environmental Medicine, Karolinska Institutet, Stockholm 171 77, Sweden

Supporting Information

ABSTRACT: Spatiotemporally resolved particulate matter (PM) estimates are essential for reconstructing long and short-term exposures in epidemiological research. Improved estimates of $PM_{2.5}$ and PM_{10} concentrations were produced over Italy for 2013−2015 using satellite remote-sensing data and an ensemble modeling approach. The following modeling stages were used: (1) missing values of the satellite-based aerosol optical depth (AOD) product were imputed using a spatiotemporal land-use random-forest (RF) model incorporating AOD data from atmospheric ensemble models; (2) daily PM estimations were produced using four modeling approaches: linear mixed effects, RF, extreme gradient boosting, and a chemical transport model, the flexible air quality regional model. The filled-in MAIAC AOD together with additional spatial and temporal predictors were used as inputs in the three first models; (3) a geographically weighted generalized additive model (GAM) ensemble model was used to fuse the estimations from the four models by allowing the weights of each model to vary over space and time. The GAM ensemble model outperformed the four separate models, decreasing the cross-validated root mean squared error by 1−42%, depending on the model. The spatiotemporally resolved PM estimations produced by the suggested model can be applied in future epidemiological studies across Italy.



1. INTRODUCTION

Particulate matter (PM) concentrations are regularly monitored worldwide inter alia because of associations that were found between long-term and short-term exposures to PM and adverse health effects. PM was ranked as the sixth leading cause of death in the Global Burden of Diseases study.[1] Assessing exposure to PM concentrations measured at monitoring stations that are spread sparsely and mainly in urban areas presents a limitation for health studies, especially for considering the exposure of rural and suburban populations. During the last years, different models have been developed to allow spatially and temporally continuous air pollution estimations, thus extending beyond the limited network of air pollution monitors. Satellite-based aerosol optical depth (AOD) products have been used in many recent studies as a predictor of PM,[2] as AOD constitutes a real physical measurement of the amount of light absorbed or scattered by suspended particles along the vertical atmospheric column. A mixed effects modeling approach was used to

Received: July 17, 2019
Revised: November 19, 2019
Accepted: November 21, 2019
Published: November 21, 2019

DOI: 10.1021/acs.est.9b04279
Environ. Sci. Technol. 2020, 54, 120−128

estimate PM[3−8] because of its ability to allow the relationship between PM and AOD to vary from day to day and produce accurate estimates of PM in different regions. These studies had to address AOD data not being spatially and temporally continuous because of various reasons such as cloud coverage, water/snow glint reflectance, and satellite calibration. For instance, the percent of missing data might range between ∼65 and 85% over Italy,[3] depending on the season and the geographic location. This issue led to the development of different interpolation and smoothing approaches to account for the missing data.[4−8] A recent study by Stafoggia et al. (2019) presented a novel approach for imputing missing AOD data over Italy from the Multiangle Implementation of Atmospheric Correction (MAIAC) algorithm by applying a random-forest (RF) model that uses modeled AOD estimates from atmospheric ensemble models as a predictor. The imputed MAIAC AOD estimation from this approach can be used as a predictor in any chosen modeling approach, allowing spatially and temporally continuous PM estimation.

Another modeling approach includes chemical transport models (CTMs) that have proved to be capable to reproduce atmospheric pollution phenomena and their reliability has been demonstrated by several single and multimodel evaluation studies. Nevertheless, there are many sources of uncertainty in their use for operational applications:[9,10] in emission data, in meteorological predictions, and an incomplete representation of the physical/chemical mechanisms that determine pollutant concentrations. These uncertainties can determine model errors[11,12] and consequently failures in air quality predictions. Moreover, CTMs have space resolution limitations that do not permit reproduction of subgrid-scale features that can determine hot-spot concentrations. For these reasons, CTM concentration estimations can be compared positively with rural and urban background air quality stations, whereas concentrations measured at traffic stations and in small towns can hardly be reconstructed. To overcome the above model limitations and uncertainties, different solutions were adopted like data fusion, assimilation techniques, or ensemble modeling.[13,14]

Recently, more studies in the field of air pollution modeling[3,15−19] have applied ensemble models because of their ability to incorporate predictions from multiple base learners, which allows combining their predictive power and create a final prediction that outperforms the predictions from each base learner. Stafoggia et al. (2019) used an RF approach to produce spatially and temporally continuous $PM_{2.5}$ and PM_{10} predictions using the filled in AOD and additional spatiotemporal predictors, showing promising results that add to recent studies[16,20−23] that presented the advantage of ensemble machine learning for air pollution modeling. Previous air pollution modeling studies usually applied a single modeling approach throughout the different stages, such as mixed effects modeling,[4−8] machine learning ensemble model,[3,20] or explored the performance of several models using eventually the one that performed the best.[17] The novelty of this research is the use of PM estimations from multiple different modeling approaches (called learners) in one ensemble model that accounts for geographical variation in the performance of these models, for the first time in Italy. The underlying assumption of this research is that each modeling approach has its merits and limitations, and an ensemble model that incorporates the PM predictions from a heterogeneous set of base learners generated from different modeling approaches is beneficial. The ensemble model used in this research applied a geographically weighted generalized additive model (GAM) to produce PM_{10} and $PM_{2.5}$ concentration estimates. In contrast to the standard approach for ensemble averaging, which uses a linear regression to estimate fixed weights for each learner, the GAM approach that was used here allows the weights for each learner to vary spatially, and also by concentration. For example, one learner may fit better at high PM concentrations and another at low concentrations. Or one learner may fit better in one region of Italy, and another in a different region. Italy is characterized by complex conditions for air pollution modeling because of its diverse geo-climatic zones and the complex mixture of anthropogenic and natural sources of air pollution. Therefore, the ensemble approach can be suitable for such areas, where the different models might have different performances throughout space and time. The primary objective of this research is to improve the estimation of daily concentrations of PM_{10} and $PM_{2.5}$ over Italy for the years 2013−2015 using a geographically weighted GAM ensemble model that incorporates the predictions from a unique combination of models [linear mixed effects model (LMM), machine-learning models, and CTM].

The study has been conducted within the project BEEP, "Big Data in Environmental and Occupational Epidemiology", funded by the National Institute for Insurance of Work-related Injuries (INAIL) and aimed at developing large-scale spatiotemporal estimates of environmental exposures for the evaluation of the short-term effects of air pollution and extreme temperatures on mortality and hospitalizations in Italy.

2. MATERIALS AND METHODS

2.1. Study Domain. Italy is a boot-shaped peninsula located in southern Europe with a total area is 307 635 km^2 (Figure S1). It is characterized by diverse geoclimatic areas, with two major mountain ranges (Alps and Apennines), one large plain (the Po valley), a long coastal line, and many medium-sized urban areas (46 municipalities above 100,000 inhabitants, 99 between 50,000 and 100,000, 165 between 30,000 and 50,000). Big metropolitan areas are also located along the territory with a population of over 500,000 inhabitants. Its wide variety of landscapes and climatic zones combined with a complex mixture of anthropogenic and natural sources of air pollution affect air quality differently across space (north to south, on the mountains vs coastal areas) and over seasons. Italy is affected by high concentrations of PM, particularly in the Po valley and in the main metropolitan areas, where concentrations often exceed the EU legal limit of PM_{10}.

2.2. Data. *2.2.1. PM_{10} and $PM_{2.5}$ Monitoring Data.* Daily (24 h mean) $PM_{2.5}$ and PM_{10} concentrations were obtained from the Italian Institute for Environmental Protection and Research (ISPRA). As there were considerably fewer $PM_{2.5}$ monitors before 2013, in this study the period 2013−2015 was considered, during which data were available from:

- * 198, 221, and 229 monitoring stations measuring $PM_{2.5}$ and PM_{10}, respectively, for the years 2013, 2014, and 2015;
- * 308, 298, and 295 monitoring stations measuring only PM_{10}, respectively, for the years 2013, 2014, and 2015.

2.2.2. AOD Data. AOD is a measure of the extinction of the solar beam by aerosol particles; it is a dimensionless number

DOI: 10.1021/acs.est.9b04279
Environ. Sci. Technol. 2020, 54, 120−128

that is related to the amount of aerosol in the vertical column of atmosphere over a given location and is therefore useful to estimate PM concentrations. The AOD product calculated by the MAIAC algorithm[24] based on collection 6 MODIS Aqua L1B data were downloaded for the period of research (2013−2015). This product was chosen because of its spatial resolution (1 km), temporal resolution (daily), long time coverage (2003−present for Aqua), and improved retrieval accuracy.[25] AOD observations were pre-processed to filter unreliable data as detailed in Stafoggia et al. (2017).

2.2.3. Spatial and Spatiotemporal Predictors. Various spatial and spatiotemporal predictors were computed for each 1×1 km^2 grid cell: (1) spatial predictors: geo-climatic zones, administrative regions, resident population, pollutants' emission data, mean elevation, impervious surface area, light at night data, land cover data, road density data, and distance from the closest road, airports, ports, sea, lakes; (2) spatiotemporal predictors: meteorological predictors (daily mean air temperature, sea-level barometric pressure, precipitations, relative humidity, wind speed, wind direction, and planetary boundary layer (PBL) height), normalized difference vegetation index (NDVI) at monthly temporal resolution, desert dust advection days,[26] and emission estimates (PM_{10}, PM_{25}, and NO_x) from traffic and heating. For a detailed description of these predictors, see Table S1.

2.3. Statistical Methods. An ensemble model, based on the combination of $PM_{2.5}$ estimations provided by different modeling approaches, has been developed to exploit the advantages of each model and to construct a final spatiotemporally resolved model, which will potentially outperform the individual models' results. The following three stages (Figure S2) have been used to develop the ensemble model: (1) increasing the observational $PM_{2.5}$ network and filling missing MAIAC AOD data; (2) estimating PM concentration using four models: LMM, RF, extreme gradient boosting (XGBoost), and the Italian CTM, namely the flexible air quality regional model (FARM); (3) fitting a GAM ensemble model based on the cross-validated PM estimations from the four models and producing spatiotemporally continuous estimations of PM for whole Italy. The framework for PM_{10} is similar except that the measured values are used as inputs for stage 2. All statistical analyses have been performed with the R statistical software, version 3.4.4.[27]

2.4. Stage 1: Data Imputation—Increasing the Observational $PM_{2.5}$ Network and Filling of Missing AOD Data. Two imputed data-sets from a previous work by Stafoggia et al. (2019) were used in this study: $PM_{2.5}$ concentrations (in PM_{10} monitoring sites) and MAIAC AOD across Italy. Because of the limited spread of $PM_{2.5}$ monitors across Italy, an RF model has been used to estimate daily mean $PM_{2.5}$ concentrations at locations where only PM_{10} measurements were available. A separate model was applied for each year where daily $PM_{2.5}$ concentrations were the target variable, and co-located PM_{10} concentrations were the main predictor. Additional predictors were monitoring location (traffic, industrial, or background), month, day of the week, and geographical coordinates. After fitting the model on monitors where both PM_{10} and $PM_{2.5}$ measurements were available, $PM_{2.5}$ concentrations were estimated at locations where only PM_{10} concentrations were measured, thus increasing the observational $PM_{2.5}$ network. This stage of the model performed well with a cross-validated R^2 of 0.87−0.90 and root mean squared error (RMSE) of 4.24−4.7 for the different years. Predictions on testing monitors were unbiased with intercepts close to 0 and slopes close to 1.

Satellite-based AOD data are often missing because of cloudiness and other limitations. MAIAC AOD were filled using a spatiotemporal land-use RF model based on modeled AOD data from the Copernicus Atmosphere Monitoring Service (CAMS), available from the European Centre for Medium-Range Weather Forecasts (ECMWF).[28] A separate model was applied for each year considering daily 1 km^2 MAIAC AOD as the target variable and the following as predictors: co-located multiband 3 h AOD estimates from CAMS, day of the year, and geographical coordinates. After fitting the RF model on locations where MAIAC AOD data are available, the same model was used to estimate MAIAC AOD at locations where satellite observations were absent, thus producing a continuous spatiotemporal surface of MAIAC AOD. The performance at this stage was very good with low variability between the years, cross-validated R^2 close to 0.95, low RMSE (0.02−0.03), an intercept of 0 and a slope very close to 1. The filled-in MAIAC AOD data alongside additional spatial and temporal predictors are used as inputs in three models (RF, XGBoost, and LMM) applied in this study, allowing spatially and temporally continuous estimation of PM on the next stage. A detailed description of $PM_{2.5}$ and the MAIAC AOD imputation process can be found in Stafoggia et al. (2019).

2.5. Stage 2: Estimating PM Using Different Models. *2.5.1. Linear Mixed Effects Model.* The LMM calibrates the AOD grid-level observations to the $PM_{2.5}$ or PM_{10} air quality monitoring stations using all daily observations with the closest available AOD value within 1 km during the study period, while adjusting for spatial and temporal predictors. Specifically, the following LMM (calibration stage) is used:

$$\mathrm{PM}_{ij} = (\alpha + u_j) + (\beta_1 + v_j)\mathrm{AOD}_{ij} + \sum_{m=1}^{11} \gamma_{1m} X_{1mi} + \sum_{m=1}^{10} \gamma_{2m} X_{2mij} + \varepsilon_{ij} \quad (1)$$

where PM_{ij} is the measured PM_{10} or $PM_{2.5}$ concentration at site i on day j; α and u_j are the fixed and random (day-specific) intercepts, respectively; AOD_{ij} is the AOD value at the grid cell corresponding to site i on day j; and β_1 and v_j are the fixed and day-specific random slopes, respectively. X_{1mi} is the value of the m-th spatial predictor at site i (i.e., elevation, light at night, population density, percent of certain land cover type, road density, total emissions, distance from water bodies, from airports, and from points emissions), and γ_{1m} is the corresponding fixed-effects slope of the m-th spatial predictor. X_{2mij} is the value of the m-th spatiotemporal predictor at site i, on day j (i.e., daily mean air temperature, sea-level, barometric pressure, precipitation, relative humidity, wind speed, wind direction, PBL height, NDVI, traffic- or heating-related emissions, dust classification) and γ_{2m} is the corresponding fixed-effects slope of the m-th spatiotemporal predictor.

2.5.2. Machine-Learning Ensemble Models (RF and XGBoost). RF[29] and XGBoost[30] are examples of ensemble learning methods that train multiple decision trees for one dataset and construct an ensemble of those trees using different approaches. The randomness in the RF is reflected in the fact that each tree is built using a bootstrap sample of the data, and each node of the tree is split according to the best of

DOI: 10.1021/acs.est.9b04279
Environ. Sci. Technol. 2020, 54, 120−128

a subset of randomly chosen predictors.[31] After training the desirable number of trees, an average of their outputs is used to get a final RF ensemble prediction. XGBoost implements the gradient boosting decision tree algorithm. As opposed to RF, where trees are independent, in XGBoost the individual trees are dependent because each tree focuses its learning on what has not been well modeled by the previous tree, as each new tree is created to predict the residuals or errors of prior trees and then added together to make the final prediction. RF is relatively simple to use with only three hyper-parameters available for tuning compared to XGBoost, which has additional tuning parameters related to regularization to avoid overfitting of the boosted model. The first stage of the ensemble learning methods is the so-called hyper parameter tuning (see the Supporting Information for further details). The second stage of this modeling approach is to fit a model using the chosen hyper-parameters and the various spatial and temporal predictors. The "Ranger" package[32] was used to fit the RF model, and the "Caret"[33] package was used for the XGBoost model.

2.5.3. CTM FARM. A modeling system based on the CTM FARM[13] and on meteorology, emission, and boundary-condition modules has been used to perform high spatial resolution (5 × 5 km) simulations over Italy of PM for a 3 year period (2013−2015). This modeling system accounts for the main processes involved in air quality, such as emission, dispersion, transformation, and deposition. Conversely to the above-described models, it adopts a deterministic approach to estimate ground PM concentrations. Detailed descriptions of this model and its inputs[34−36]are available in the Supporting Information. The capability of the system to capture the spatiotemporal distribution of PM_{10} and $PM_{2.5}$ has been assessed through a comparison with monitoring network data (see Figure S4 for further details). The PM estimations from the FARM model were averaged in each grid cell over 24 h and downscaled to the 1 km AOD grid using the estimation from the closest grid. Additional details of this model are available in the Supporting Information.

2.6. Stage 3: GAM Ensemble Model. The geographically weighted GAM ensemble model aims to produce an improved PM estimation compared to the estimations from each base learner (LMM\ RF\XGBoost\FARM). This model uses the cross-validated PM estimations (i.e., at held out monitoring sites) from the other models as predictors. The cross-validated estimations are used, as opposed to estimations in the training data, because they better approximate the predictive performance of the models in locations without data. The ensemble stage is applied using GAM, which is suitable for cases where it is desired to predict from complex, nonlinear, and possibly interacting relationships. This approach allows flexibility in the weights that vary through space for each model and assigning higher weights to a model that performs better for certain locations. The model is based on the following relationship

$$\begin{aligned}\mathrm{GAM(PM)}_{ij} = {} & s(X, Y, \text{ by} = \mathrm{pred}_{\mathrm{LMM}})_{ij} + s \\ & (X, Y, \text{ by} = \mathrm{pred}_{\mathrm{RF}})_{ij} + s(X, Y, \text{ by} = \mathrm{pred}_{\mathrm{XGBoost}})_{ij} \\ & + s(X, Y, \text{ by} = \mathrm{pred}_{\mathrm{FARM}})_{ij} + \varepsilon_{ij} \end{aligned} \quad (2)$$

where PM_{ij} is the PM concentration on monitor i and in day j, X and Y are the universal transverse mercator (UTM) coordinates of each monitor i, whereas $\mathrm{pred}_{\mathrm{LMM}}$, $\mathrm{pred}_{\mathrm{RF}}$, $\mathrm{pred}_{\mathrm{XGBoost}}$ are the cross-validated estimations from the LMM, RF, XGBoost, $\mathrm{pred}_{\mathrm{FARM}}$ is the PM estimation from the FARM model at monitor i and on day j, and ε_{ij} is the error of the model.

The GAM ensemble model uses the PM estimations from the four models as predictors. Therefore, first, PM estimations are produced from these models for all the locations and days using the models fitted in stage 2 (Sections 2.5.1−2.5.3). The GAM model that is fitted in this stage is used for the final PM estimation over all Italy.

2.7. Model Performance Evaluation. The model performance process was based on cross-validation (CV) by monitors, for example, splitting of the monitors into training and testing groups, with model evaluation done in the testing only. As our modeling procedure entailed the fitting of individual learners first, and the implementation of an ensemble model afterward, our CV scheme was designed to make sure data were not overfitted in any of these phases. Specifically, the monitoring database was divided into 10 random groups of monitors. The process of CV is based on two substages (Figure S5):

(1) This substage creates an out-of-sample prediction of the three models (LMM, RF, and XGBoost). At each iteration, the three models are trained on 80% of the monitors and estimate PM for a separate test dataset, which is another group of random 10% of the monitors. The remaining 10% of the monitors (validation dataset) is kept aside for later evaluation of the GAM model.

(2) At each iteration in this substage, the GAM ensemble model is trained while using the out-of-sample estimations of PM from the different learners (which are created in stage 1 of the CV). Then, it predicts for the 10% validation set that was kept aside during stage 1. This ensures that the GAM model is trained on unbiased PM estimations and predicts for a dataset that was not used during the training of the other learners, thus simulating the prediction process in places with no monitoring stations. The other learners are also trained using the same training data (90% of the monitors) and predicts for the validation set. The process is repeated 10 times until there is complete PM estimation for all the monitors. After aggregating the entire out-of-sample estimation of PM from a certain model, these estimations were compared to the actual measurements in the monitors and the following performance measures were computed:

* Coefficient of determination (R^2): the observed and predicted PM values were regressed and the percent of explained variance was computed;

* RMSE: the square root of the mean quadratic differences between observed and predicted PM values. It is a summary measure of the prediction error, and it is on the same scale as the measured observation (PM, $\mu g/m^3$);

* Slope: the coefficient from the linear regression between PM observed and PM predicted. It represents the multiplicative bias;

* Intercept: the intercept from the linear regression between PM observed and PM predicted. It represents the additive bias in the model.

DOI: 10.1021/acs.est.9b04279
Environ. Sci. Technol. 2020, 54, 120−128

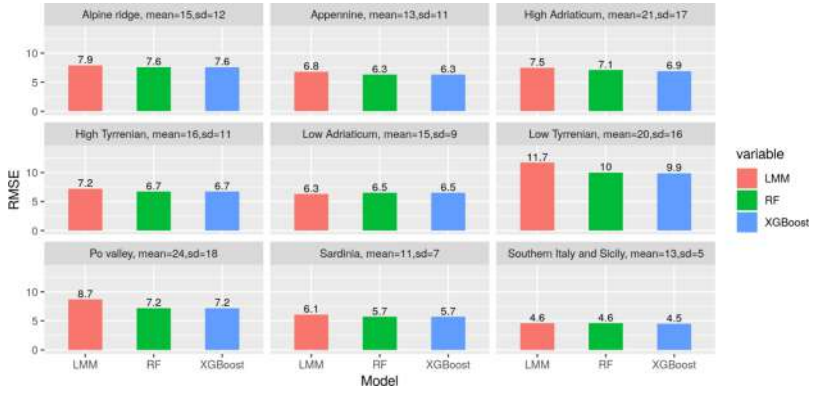


Figure 1. Variability in cross-validated RMSE of the three models across the nine climatic zones. The mean and sd of $PM_{2.5}$ concentrations are detailed for each zone.

Table 1. Performance Measures of the Different Models for 2013−2015[a]

		$PM_{2.5}$					PM_{10}				
year	measure	GAM	XGBoost	RF	LMM	FARM	GAM	XGBoost	RF	LMM	FARM
2013	R^2	0.79	0.78	0.77	0.71	0.44	0.73	0.71	0.7	0.65	0.36
	RMSE	6.56	6.66	6.84	7.73	10.67	9.42	9.73	9.91	10.74	14.56
	intercept	0.13	−0.91	−1.49	−1.05	7.32	0.25	−0.48	−3.14	−0.73	13.57
	slope	0.99	1.04	1.07	1.06	0.76	0.99	1.01	1.1	1.05	0.82
2014	R^2	0.79	0.79	0.77	0.71	0.36	0.74	0.74	0.73	0.68	0.24
	RMSE	5.29	5.34	5.49	6.18	9.21	8.54	8.66	8.74	9.6	14.72
	intercept	0.18	−0.8	−1.33	−0.84	7.73	0.59	−0.66	−2.56	0.03	14.48
	slope	0.99	1.04	1.07	1.04	0.64	0.98	1.02	1.08	1.01	0.7
2015	R^2	0.81	0.8	0.79	0.72	0.48	0.76	0.75	0.74	0.69	0.42
	RMSE	6.34	6.46	6.62	7.67	10.42	8.95	9.07	9.32	10.21	13.9
	intercept	0.47	−0.7	−1.53	0.06	8.61	0.62	−0.39	−2.2	−0.62	15.03
	slope	0.99	1.03	1.07	1.06	0.72	0.98	1.01	1.07	1.04	0.79

[a]The results for RF, XGBoost, and linear mixed model (LMM) are cross-validated, whereas those for the FARM model are not.

3. RESULTS AND DISCUSSION

The descriptive statistics of the PM_{10} and $PM_{2.5}$ concentrations as measured by monitors and the imputed $PM_{2.5}$ from stage 1 are presented in Table S2. Observed mean concentrations across Italy are between 24 and 27 $\mu g/m^3$ for PM_{10} and 16−18 $\mu g/m^3$ for $PM_{2.5}$. The imputed $PM_{2.5}$ concentrations show similar descriptive statistics and these are the actual values that are used as inputs in the following stages of the model.

To strengthen the hypothesis that the performance of the different models might vary across space and an ensemble model is beneficial, the CV performance of three modeling approaches was compared for $PM_{2.5}$ in the nine different geo-climatic zones for the year 2015. Figure 1 shows the variability in CV rmse of the LMM, the RF, and the XGBoost across the geo-climatic zones, and the mean and standard deviation (sd) of measured $PM_{2.5}$ concentrations for each zone. The spread of $PM_{2.5}$ monitors within the geo-climatic zones can be found in Figure S6. The varying performance of each model for the different geo-climatic areas across Italy supports the use of the GAM for the ensemble model that provides different weights to the estimations from these models smoothly across space. The spatial variation in the effect of different learners on $PM_{2.5}$ concentrations within the GAM ensemble model is presented in Figure S7 (from left to right), supporting the idea that varying levels of the four learners effect differently across space.

The CV performance of the different models is summarized in Table 1. The range of CV R^2 results for the GAM ensemble model was 0.73−0.76 for PM_{10} and 0.79−0.81 for $PM_{2.5}$. In all years, the ensemble model showed the best performance or at least very similar to the XGBoost model, which performed the best among the different base learners in some years, with highest R^2, lowest rmse, an intercept that is closest to 0 and slope that is closest to 1, indicating a minimum bias of this model. The errors of the ensemble model are within a reasonable range in comparison to the measured PM values, with 50% of the model residuals ranging between −14 and 25% (percent out of the measured PM_{10} values), and between −7 and 62% in the case of $PM_{2.5}$ (Figure S8 and Table S3). Moreover, the ensemble model shows an improved CV performance in comparison to the previous models in Italy that applied mixed effects modeling[4] or spatiotemporal land-use RF model.[3] The FARM model showed the worst performance, followed by the LMM, the RF, and the XGBoost. Table S4 divides the performance by monitor type (traffic, industrial, background) showing that in most cases and in all types of monitors, the ensemble performs better than the single learners with lowest RMSE and highest R^2. This implies that this modeling approach is preferable through diverse areas where the PM is originated from different pollution sources and has different chemical characteristics. These findings are in line with the research assumption that ensemble learners might have an advantage over other models because of their ability to incorporate multiple predictions from different learners. The results presented here add up to a growing body of literature in the field of air pollution modeling, showing that ensemble modeling is a strong tool that allows high predictive power.

DOI: 10.1021/acs.est.9b04279
Environ. Sci. Technol. 2020, 54, 120−128

Other studies[3,15−17] that implemented ensemble models showed similar findings that strengthen the power of these models to improve air pollution modeling. Several recent studies[3,20,37] applied machine-learning ensemble models (RF) to predict PM concentration over different regions (Italy, China, and the US), showing high performance and an advantage over the often used parametric regression models. A study carried out in the Southeastern US by Murray et al. (2018) showed that for estimating daily $PM_{2.5}$ a Bayesian ensemble approach outperformed other statistical downscalers that use either AOD or CTM. Their ensemble approach performs data fusion of multiple sources of information (CTMs simulation or satellite AOD) to predict $PM_{2.5}$ concentrations with complete spatiotemporal coverage. Differently from previous studies that use one or few modeling approaches, the strength of the GAM ensemble methodology is its ability to combine estimations from several different modeling approaches while accounting for geographical differences in the performance of these models.

The FARM model is not a data-driven model, but a deterministic model based on a priori equations accounting for the transport, dispersion, chemical conversion, and deposition of atmospheric pollutants. Therefore, it is expected that it will show a different performance compared to the other models. Because of the lowest performance of the FARM model, a sensitivity test was carried out for the ensemble model for year 2015 to explore if it is beneficial to include the predictions provided by this model as a base learner in the ensemble. The CV results of this analysis are summarized in Table 2. This exploration shows that for $PM_{2.5}$, the results of the ensemble model with and without the FARM are identical, meaning that the GAM model provides very low weights to this model. Figure S7 presents the low effect of the FARM estimations on $PM_{2.5}$ concentrations in comparison to other models. As for PM_{10}, there is a slight difference in RMSE, slope, and intercept, which indicates that it is better to exclude the FARM prediction from the ensemble.

Table 2. Analyzing the GAM Ensemble Model Performance for the Year 2015 Depending on the Inclusion of the FARM Model

	$PM_{2.5}$		PM_{10}	
	GAM with FARM	GAM without FARM	GAM with FARM	GAM without FARM
R^2	0.81	0.81	0.76	0.76
RMSE	6.31	6.31	8.98	8.95
intercept	0.49	0.49	0.8	0.62
slope	0.99	0.99	0.97	0.98

The spatial patterns of the GAM ensemble model for the 3 years of modeling are presented in Figure 2. The general spatial pattern is similar between the 3 years, showing the highest concentrations of PM_{10} around the Po valley in the northern part, and within the major cities of Rome, Milan, Naples, and Turin, with the highest values encountered during 2015 in these locations. The mean annual $PM_{2.5}$ concentrations show also the highest values around the Po valley and in the city of Rome, as well as around Frosinone in proximity to the Sacco river valley, with the highest values identified in 2013. These spatial patterns are similar to those derived from model estimations developed by Stafoggia et al., (2019, 2017) using LMM and RF models, showing the highest values of PM concentration in these major metropolitan areas and their surrounding industrial areas. More detailed maps of these areas can be found in Figures S10 and S11.

Figure 3 shows the spatial pattern of the five models used in this study. The three data-driven models (GAM, XGBoost, RF, and LMM) show generally a similar pattern. RF and XGBoost show very similar results, compared to LMM, which shows in some area, lower PM estimations (Sicily, Sardinia, and over major mountain ranges). The spatial pattern of the FARM model is the most exceptional, showing high values of PM focused in the main metropolitan areas and considerably less variability in PM values through all other regions in comparison to the other models. Such a different pattern is expected given its lowest performance and because of the uncertainties in input data (mainly emission data) and in the model assumptions and the adopted spatial resolution (5 km). The GAM ensemble incorporates the PM estimations from the three models (LMM, RF, and XGBoost) and its map shows how in some areas there is distinct integration between the spatial patterns of LMM and the two ensemble models (RF and XGBoost).

Although the proposed GAM ensemble model showed promising results and outperformed other learners, there are some limitations that should be considered. First, in this study daily mean PM concentrations are estimated, whereas the MAIAC AOD product that is used as one of the main predictors is derived from Aqua satellite that is characterized by sun-synchronous orbit, meaning that measurements are taken at a specific time of the day (afternoon overpass). Deriving AOD products from a geostationary satellite platforms, which image the same area all the time, allows high spatial temporal resolution, with daytime availability of up to every 15 min (i.e., the SEVIRI sensor onboard the Meteosat satellite). Such products allow estimation of the daytime mean AOD, which potentially might be a better predictor of daily (24 h) mean PM. Second, the clustered spread of PM monitors, mainly in proximity to populated areas means that the model is trained and tested mainly on measurements taken from urban areas with specific air pollution characteristics and that remote rural areas are under-represented (Figure S9). This limitation is marginally relevant for most epidemiological studies that use data from populated areas. For other implementations, a possible solution might be integration of measurements from low-cost PM monitors and sensors that will increase their spread to remote areas.

The presented approach has also many strengths. By the integration of the different models whose performance might vary across space, the ensemble model exploits the advantages of each model by allowing the weights of each model to vary over space and time and constructs a more reliable spatiotemporal PM estimation which outperforms the individual models' results. With respect to deterministic models (e.g., the FARM model), which are widely used to assess air quality impacts at different spatial scales (from continental to urban), the individual machine-learning techniques and their ensemble approach have demonstrated, when properly fed with spatiotemporal data, to better perform in assessing air quality, particularly in areas where the uncertainties in emissions are high as in rural areas. The GAM ensemble model can be used to improve the accuracy of air pollution models, and the estimations from such models can be applied confidently to study the association with health outcomes.

DOI: 10.1021/acs.est.9b04279
Environ. Sci. Technol. 2020, 54, 120−128

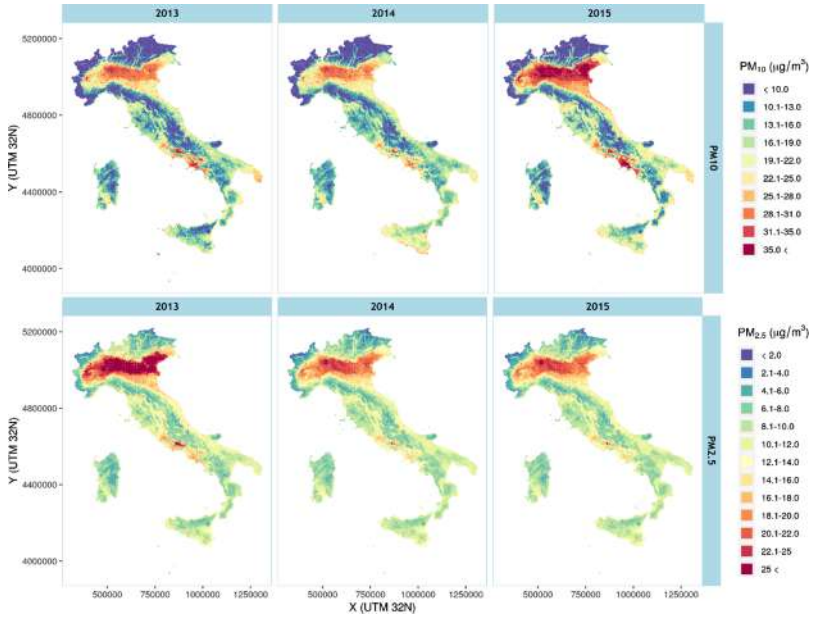


Figure 2. Mean concentrations ($\mu g/m^3$) of PM_{10} (upper panel) and $PM_{2.5}$ (lower panel) estimated by the GAM ensemble model for years 2013–2015.

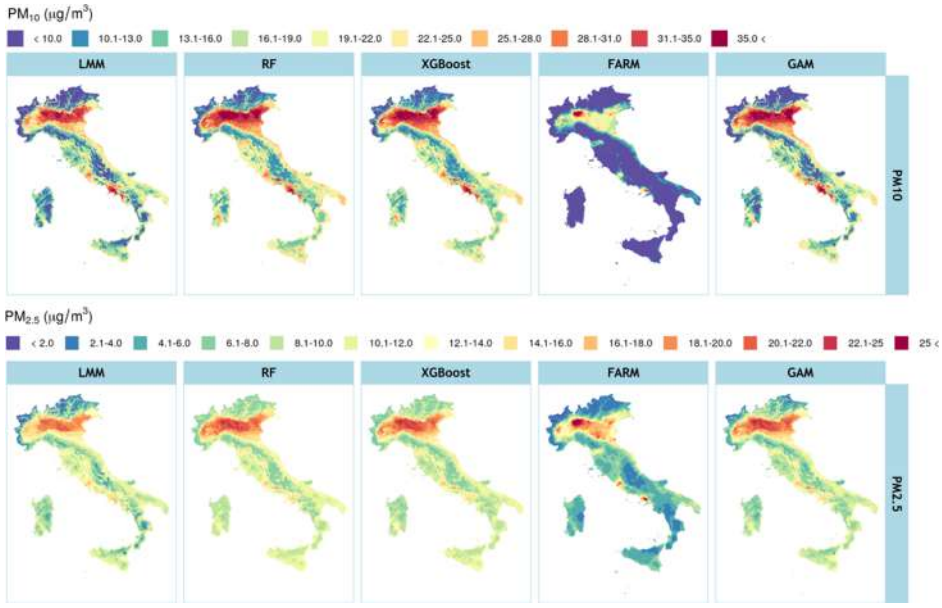


Figure 3. Maps of mean concentrations ($\mu g/m^3$) of PM_{10} (upper panel) and $PM_{2.5}$ (lower panel) for the year 2015 for the five different models.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.9b04279.

Study area map, detailed description of predictors, modelling stages flowchart, details of machine learning models and the CTM, cross-validation framework, summary statistics of PM_{10} and $PM_{2.5}$ concentrations, map of $PM_{2.5}$ monitors spread along geoclimatic zones, spatial variation in the effect of different learners on $PM_{2.5}$ concentrations within the GAM model, variably in performance between different station types and different zones, distribution of the percent of relative errors for the GAM ensemble model estimations, maps of mean PM_{10} concentrations for 2013–2015 around metropolitan areas (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: shtien@post.bgu.ac.il.

ORCID

Alexandra Shtein: 0000-0002-2852-4477
Camillo Silibello: 0000-0002-0400-6755
Allan C. Just: 0000-0003-4312-5957

Notes

The authors declare no competing financial interest.

DOI: 10.1021/acs.est.9b04279
Environ. Sci. Technol. 2020, 54, 120–128

■ ACKNOWLEDGMENTS

This work has received funding from INAIL, contract ID 04/2016. The authors would like to thank the Ministry of Science and Technology, Israel, for supporting the research [grant number: 3-13142] and A.S. with a PhD scholarship, U.S. EPA grant sRD-83479801 and RD-83587201 for supporting J.S, and NIH grants R00 ES023450 and P30 ES023515 for supporting A.C.J. BEEP Collaborative Group: Ancona C., Bucci S., de' Donato F., Michelozzi P., Renzi M., Scortichini M., Stafoggia M.; Bonafede M., Gariazzo C., Marinaccio A.; Argentini S., Sozzi R.; Bonomo S., Fasola S., Forastiere F., La Grutta S., Viegi G.; Cernigliaro A., Scondotto S.; Baldacci S., Maio S; Licitra G,. Moro A.; Angelini P., Bonvicini L., Broccoli S., Ottone M., Giorgi Rossi P., Ranzi Andrea.; Galassi C., Migliore E.; Bisceglia L., Chieti A.; Brusasca G., Calori G., Finardi S., Nanni A., Pepe N., Radice P., Silibello C., Tinarelli G., Uboldi F., Carlino G.

■ REFERENCES

(1) Gakidou, E. Global, Regional, and National Comparative Risk Assessment of 84 Behavioural, Environmental and Occupational, and Metabolic Risks or Clusters of Risks, 1990-2016: A Systematic Analysis for the Global Burden of Disease Study 2016. *Lancet* **2017**, *390*, 1345−1422.

(2) Sorek-Hamer, M.; Just, A. C.; Kloog, I. Satellite Remote Sensing in Epidemiological Studies. *Curr. Opin. Pediatr.* **2016**, *28*, 228−234.

(3) Stafoggia, M.; Bellander, T.; Bucci, S.; Davoli, M.; de Hoogh, K.; de' Donato, F.; Gariazzo, C.; Lyapustin, A.; Michelozzi, P.; Renzi, M.; Scortichini, M.; Shtein, A.; Viegi, G.; Kloog, I.; Schwartz, J. Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013-2015, using a spatiotemporal land-use random-forest model. *Environ. Int.* **2019**, *124*, 170−179.

(4) Stafoggia, M.; Schwartz, J.; Badaloni, C.; Bellander, T.; Alessandrini, E.; Cattani, G.; de' Donato, F.; Gaeta, A.; Leone, G.; Lyapustin, A.; Sorek-Hamer, M.; de Hoogh, K.; Di, Q.; Forastiere, F.; Kloog, I. Estimation of Daily PM10 Concentrations in Italy (2006-2012) Using Finely Resolved Satellite Data, Land Use Variables and Meteorology. *Environ. Int.* **2017**, *99*, 234−244.

(5) de Hoogh, K.; Héritier, H.; Stafoggia, M.; Künzli, N.; Kloog, I. Modelling Daily PM2.5 Concentrations at High Spatio-Temporal Resolution across Switzerland. *Environ. Pollut.* **2018**, *233*, 1147−1154.

(6) Kloog, I.; Sorek-Hamer, M.; Lyapustin, A.; Coull, B.; Wang, Y.; Just, A. C.; Schwartz, J.; Broday, D. M. Estimating daily PM2.5 and PM10 across the complex geo-climate region of Israel using MAIAC satellite-based AOD data. *Atmos. Environ.* **2015**, *122*, 409−416.

(7) Shtein, A.; Karnieli, A.; Katra, I.; Raz, R.; Levy, I.; Lyapustin, A.; Katra, M.; Dorman, D. M.; Kloog, I. Estimating Daily and Intra-Daily PM10 and PM2.5 in Israel Using a Spatio-Temporal Hybrid Modeling Approach. *Atmos. Environ.* **2018**, *191*, 142−152.

(8) Just, A. C.; Wright, R. O.; Schwartz, J.; Coull, B. A.; Baccarelli, A. A.; Tellez-rojo, M. M.; Moody, E.; Wang, Y.; Lyapustin, A.; Kloog, I. Using High-Resolution Satellite Aerosol Optical Depth To Estimate Daily PM2.5 Geographical Distribution in Mexico City. *Environ. Sci. Technol.* **2015**, *49*, 8576−8584.

(9) Kukkonen, J.; Olsson, T.; Schultz, D. M.; Baklanov, A.; Klein, T.; Miranda, A. I.; Monteiro, A.; Hirtl, M.; Tarvainen, V.; Boy, M.; Peuch, V. H.; Poupkou, A.; Kioutsioukis, I.; Finardi, S.; Sofiev, M.; Sokhi, R.; Lehtinen, K. E. J.; Karatzas, K.; San José, R.; Astitha, M.; Kallos, G.; Schaap, M.; Reimer, E.; Jakobs, H. A Review of Operational, Regional-Scale, Chemical Weather Forecasting Models in Europe. *Atmos. Chem. Phys.* **2012**, *12*, 1−87.

(10) Zhang, Y.; Bocquet, M.; Mallet, V.; Seigneur, C.; Baklanov, A. Real-Time Air Quality Forecasting, Part II: State of the Science, Current Research Needs, and Future Prospects. *Atmos. Environ.* **2012**, *60*, 656−676.

(11) Chang, J. C.; Hanna, S. R. Air Quality Model Performance Evaluation. *Meteorol. Atmos. Phys.* **2004**, *87*, 167−196.

(12) Borrego, C.; Monteiro, A.; Ferreira, J.; Miranda, A. I.; Costa, A. M.; Carvalho, A. C.; Lopes, M. Procedures for Estimation of Modelling Uncertainty in Air Quality Assessment. *Environ. Int.* **2008**, *34*, 613−620.

(13) Silibello, C.; Bolignano, A.; Sozzi, R.; Gariazzo, C. Application of a Chemical Transport Model and Optimized Data Assimilation Methods to Improve Air Quality Assessment. *Air Qual., Atmos. Health* **2014**, *7*, 283−296.

(14) Zhang, Y.; Bocquet, M.; Mallet, V.; Seigneur, C.; Baklanov, A. Real-Time Air Quality Forecasting, Part I: History, Techniques, and Current Status. *Atmos. Environ.* **2012**, *60*, 632−655.

(15) Murray, N.; Chang, H. H.; Holmes, H.; Liu, Y. *Combining Satellite Imagery and Numerical Model Simulation to Estimate Ambient Air Pollution: An Ensemble Averaging Approach*, Ann. Appl. Stat. **2018**; No. 2010, pp 1−18.

(16) Zhai, B.; Chen, J. Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China. *Sci. Total Environ.* **2018**, *635*, 644−658.

(17) Just, A.; De Carli, M.; Shtein, A.; Dorman, M.; Lyapustin, A.; Kloog, I. Correcting Measurement Error in Satellite Aerosol Optical Depth with Machine Learning for Modeling PM2.5 in the Northeastern USA. *Remote Sens.* **2018**, *10*, 803−817.

(18) Li, L.; Zhang, J.; Qiu, W.; Wang, J.; Fang, Y. An Ensemble Spatiotemporal Model for Predicting PM2.5 Concentrations. *Int. J. Environ. Res. Public Health* **2017**, *14*, 549.

(19) Di, Q.; Amini, H.; Shi, L.; Kloog, I.; Silvern, R.; Kelly, J.; Sabath, M. B.; Choirat, C.; Koutrakis, P.; Lyapustin, A.; Wang, Y.; Mickley, L. J.; Schwartz, J. An Ensemble-Based Model of PM2.5 Concentration across the Contiguous United States with High Spatiotemporal Resolution. *Environ. Int.* **2019**, *130*, 104909.

(20) 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*, 6936−6944.

(21) 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*, 4712−4721.

(22) Brokamp, C.; Lemasters, G. K.; Ryan, P. H. Residential Mobility Impacts Exposure Assessment and Community Socioeconomic Characteristics in Longitudinal Epidemiology Studies. *J. Exposure Sci. Environ. Epidemiol.* **2016**, *26*, 428−434.

(23) 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*, 4173.

(24) Lyapustin, A.; Wang, Y.; Korkin, S.; Huang, D. MODIS Collection 6 MAIAC Algorithm. *Atmos. Meas. Tech.* **2018**, *11*, 5741−5765.

(25) Mhawish, A.; Banerjee, T.; Sorek-Hamer, M.; Lyapustin, A.; Broday, D. M.; Chatfield, R. Comparison and Evaluation of MODIS Multi-Angle Implementation of Atmospheric Correction (MAIAC) Aerosol Product over South Asia. *Remote Sens. Environ.* **2019**, *224*, 12−28.

(26) Pey, J.; Querol, X.; Alastuey, A.; Forastiere, F.; Stafoggia, M. African dust outbreaks over the Mediterranean Basin during 2001−2011: PM10 concentrations, phenomenology and trends, and its relation with synoptic and mesoscale meteorology. *Atmos. Chem. Phys.* **2013**, *13*, 1395−1410.

(27) *R. R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.

(28) Dee, D. P.; Uppala, S. M.; Simmons, A. J.; Berrisford, P.; Poli, P.; Kobayashi, S.; Andrae, U.; Balmaseda, M. A.; Balsamo, G.; Bauer, P.; Bechtold, P.; Beljaars, A. C. M.; van de Berg, L.; Bidlot, J.; Bormann, N.; Delsol, C.; Dragani, R.; Fuentes, M.; Geer, A. J.; Haimberger, L.; Healy, S. B.; Hersbach, H.; Hólm, E. V.; Isaksen, L.; Kållberg, P.; Köhler, M.; Matricardi, M.; McNally, A. P.; Monge-Sanz, B. M.; Morcrette, J.-J.; Park, B.-K.; Peubey, C.; de Rosnay, P.; Tavolato, C.; Thépaut, J.-N.; Vitart, F.; Lehtinen, K. E. J.; Karatzas, K.; San José, R.; Astitha, M.; Kallos, G.; Schaap, M.; Reimer, E.;

DOI: 10.1021/acs.est.9b04279
Environ. Sci. Technol. 2020, 54, 120−128

Jakobs, H.; Eben, K. The ERA-Interim Reanalysis: Configuration and Performance of the Data Assimilation System. *Q. J. R. Meteorol. Soc.* **2011**, *137*, 553−597.
(29) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(30) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: San Francisco, CA, USA, 2016; pp 785−794.
(31) Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R News*, 2002; Vol. *2*, pp 18−22. http://cran.r-_project.org/doc/Rnews/.
(32) Wright, M. N.; Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **2017**, *77*, 1−17.
(33) Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **2008**, *28*, 1−26.
(34) Kuenen, J. J. P.; Visschedijk, A. J. H.; Jozwicka, M.; Denier Van Der Gon, H. A. C. TNO-MACC_II emission inventory; a multi-year (2003−2009) consistent high-resolution European emission inventory for air quality modelling. *Atmos. Chem. Phys.* **2014**, *14*, 10963−10976.
(35) Carter, W. P. L. *Documentation of the SAPRC-99 Chemical Mechanism for VOC Reactivity Assessment; Riverside, California*, 2000; Vol. *1*.
(36) Binkowski, F. S.; Roselle, S. J. Models-3 Community Multiscale Air Quality (CMAQ) model aerosol component 1. Model description. *J. Geophys. Res.* **2003**, *108*, 4183.
(37) Chen, G.; Wang, Y.; Li, S.; Cao, W.; Ren, H.; Knibbs, L. D.; Abramson, M. J.; Guo, Y. Spatiotemporal patterns of PM10 concentrations over China during 2005-2016: A satellite-based estimation using the random forests approach. *Environ. Pollut.* **2018**, *242*, 605−613.

DOI: 10.1021/acs.est.9b04279
Environ. Sci. Technol. 2020, 54, 120−128