Atmospheric Environment 310 (2023) 119967

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

High-resolution downscaling of source resolved $PM_{2.5}$ predictions using machine learning models

Brian T. Dinkelacker [a], Pablo Garcia Rivera [a], Julian D. Marshall [b], Peter J. Adams [c,d], Spyros N. Pandis [e,f,*]

[a] Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[b] Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, 98195, USA
[c] Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[d] Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[e] Institute of Chemical Engineering Sciences (FORTH/ICE-HT), 26504, Patras, Greece
[f] Department of Chemical Engineering, University of Patras, 26500, Patras, Greece

HIGHLIGHTS

- New ML method for downscaling chemical transport model predictions.
- Provides particle mass concentration, composition, and source contributions.
- Does not require high-resolution emission inventories.
- Uses as inputs available land use regression (LUR) variables.
- Encouraging results in its first application in Pittsburgh, US.

ABSTRACT

Accurate predictions of source resolved atmospheric $PM_{2.5}$ concentrations at high resolutions using chemical transport models (CTMs) require expensive CTM simulations and development of high-resolution emissions inventories. We use multiple machine learning (ML) approaches to downscale coarse-resolution (36×36 km^2) CTM predictions to 1×1 km^2 spatial resolution. ML predictions include concentrations of the major chemical components of $PM_{2.5}$ and the contributions of its major emissions sources. Inputs for the ML models include 36×36 km^2 source resolved CTM predicted concentrations of all $PM_{2.5}$ components, meteorological data, and several land-use (LU) variables. The output of our ML models is the 1×1 km^2 source-resolved concentrations of all major $PM_{2.5}$ components in southwestern Pennsylvania (5184 km^2 domain) during February and July 2017. Models were trained and validated using 1×1 km^2 resolution source- and species-resolved CTM predictions of $PM_{2.5}$ from recent complementary studies. The best overall performance was found using a random forest (RF) model, where species and source resolved $PM_{2.5}$ concentrations were reproduced with low normalized mean bias ($|NMB| < 0.01$). The downscaling model captures the spatial distribution of $PM_{2.5}$ both by component and source, with some discrepancies when predicting the plumes of large point sources that have long-range impacts. In a test of generalizability to unknown domains, the model differentiates well between areas that are primarily urban, rural, or industrial but faces challenges with the reproduction of the effects of large point sources of $PM_{2.5}$ when entire quadrants are removed from the training data. The results represent a proof of concept for downscaling low-resolution CTM predictions using native high-resolution CTM predictions in training.

1. Introduction

Particulate matter with aerodynamic diameter less than 2.5 μm ($PM_{2.5}$) is a major contributor to poor air quality throughout the United States. $PM_{2.5}$ directly impacts visibility (Seinfeld and Pandis, 2006) and is also a major public health concern. Exposure to $PM_{2.5}$ has been linked to both short- and long-term health effects including premature death due to increased risk of cardiovascular disease, increased chance of heart attacks and strokes, and hindered lung development and lung function in children and people with asthma (Dockery and Pope, 1994). Knowledge of individual source contributions to total $PM_{2.5}$ concentrations is vital for the development of effective emissions control policies. For environmental justice applications, it is also desirable to know these individual source contributions at high spatial resolution (Banzhaf et al.,

* Corresponding author. Institute of Chemical Engineering Sciences (FORTH/ICE-HT), 26504, Patras, Greece.
E-mail address: spyros@chemeng.upatras.gr (S.N. Pandis).

https://doi.org/10.1016/j.atmosenv.2023.119967
Received 24 April 2023; Received in revised form 16 July 2023; Accepted 17 July 2023
Available online 19 July 2023

2019). Simulations of high spatial resolution (1 × 1 km^2 grid cell size), source-resolved, speciated atmospheric $PM_{2.5}$ concentration using a state-of-the-art chemical transport model (CTM) are computationally expensive even for relatively small simulation domains (Garcia Rivera et al., 2022). This predicates an opportunity to apply machine learning algorithms, leveraging previous simulation results, to estimate the output from a CTM predicting high spatial resolution and source-resolved $PM_{2.5}$ concentrations.

Machine learning techniques have been used in the forecasting of future $PM_{2.5}$ concentrations based on past observed values at monitoring stations. Support vector machines and artificial neural networks have shown promising results for this prediction task (Voukantsis et al., 2011; Bai et al., 2016; Prasad et al., 2016; Zhou et al., 2019; Karimian et al., 2019). However, these applications are not designed for predicting the variability of $PM_{2.5}$ concentrations in space and are limited to predicting a value at the precise locations of individual monitoring sites. Shtein et al. (2019) implemented a model driven by aerosol optical depth observations in lieu of CTM data.

A handful of studies have sought to produce estimates of $PM_{2.5}$ concentrations with high temporal and spatial resolution (similar to that of CTM output) by incorporating CTM predictions in the modelling framework. Xue et al. (2019) combined data from $PM_{2.5}$ monitors, satellite aerosol optical depth data, meteorological fields calculated using the Weather Research and Forecasting (WRF) model, as well as predicted $PM_{2.5}$ concentrations and composition as predicted by the Community Multiscale Air Quality (CMAQ) model at a resolution of 36 × 36 km^2 into an elastic-net regression (Zou and Hastie, 2005) model to estimate $PM_{2.5}$ concentrations in China from 2000 to 2016. Estimates of $PM_{2.5}$ were found to be in reasonable agreement with measurements at large averaging timescales (R^2 = 0.77 for annual averaging). Vlasenko et al. (2021) implemented a 3-layer artificial neural network to produce daily average concentrations of NO_2, SO_2, and ethane over Europe at a resolution of 64 × 64 km^2. The predictive model in this study was trained on CMAQ simulation results from 1979 to 2012 with WRF meteorology and constant emissions calculated for 2012. The goal of that study was to develop a streamlined approach for testing future emissions scenarios. The authors found that the neural network was able to predict concentrations of the pollutants of interest with errors on the same order as those between two different CTMs.

Neural networks have also been used to downscale the spatial resolution of $PM_{2.5}$ predictions from a CTM. Di et al. (2016) used low-resolution (0.500° × 0.667°) GEOS-Chem $PM_{2.5}$ component concentrations, meteorological data, and land-use regression variables (population density, road density, etc.) as inputs to a neural network that predicts $PM_{2.5}$ at 1 × 1 km^2 grid resolution in the northeast United States. The meteorological and land-use terms were used to downscale the GEOS-Chem output to a higher resolution, and this model was trained using available speciated $PM_{2.5}$ monitoring data as well as total $PM_{2.5}$ measurements from 2001 to 2010. The high-resolution predictions were correlated well with monitoring data (R^2 = 0.85 for all data and R^2 = 0.70–0.80 for individual $PM_{2.5}$ components on an annual basis). This model was developed with the goal of assisting epidemiological analysis during the specific study period (2001–2010) and area (northeast United States) for which the model was trained. Generalizability to other time periods and locations was not considered. This study also lacked high temporal and spatial resolution in training data. EPA measurements for $PM_{2.5}$ composition are only available every three or six days and are quite sparse in space. A similar approach based on GEOS-Chem predictions was implemented by Yu et al. (2023).

Land use regression (LUR) models use various independent variables (e.g. population density, restaurant count, road length, etc.) along with available pollutant measurements to capture the spatial variation in pollutant concentrations away from monitoring sites (Hoek et al., 2008). These models have difficulties in predicting the concentrations of major $PM_{2.5}$ components such as elemental carbon and organic aerosol and their sources (Wu et al., 2014) due to the high spatial variability of these species. In general, the smaller number of speciated $PM_{2.5}$ measurements is an important limitation. The addition of satellite data can improve the ability of a LUR model to predict species and sources (Rahman and Thurston, 2022), but performance is still moderate with R^2 = 0.67 for elemental carbon and R^2 = 0.73 for traffic $PM_{2.5}$ on an annual basis.

$PM_{2.5}$ prediction downscaling studies in the past have been hampered by the lack of high-resolution, source and species resolved $PM_{2.5}$ concentration fields to support the training of effective models. Instead, they have relied on monitoring data which are sparse in space and time. These models may or may not be able to predict in spatial and temporal domains that are not included in the training dataset. This is a difficult problem to address, however a generalizable model for high resolution $PM_{2.5}$ concentrations would be a useful tool for streamlined studies without requiring the computational resources associated with CTMs and the development time required for the appropriate high-resolution emission inventories.

In this study, we use land-use (LU) variables in combination with meteorological data and low-resolution CTM predictions (36 × 36 km^2) to predict source-resolved, speciated $PM_{2.5}$ concentrations at high-resolution (1 × 1 km^2). The goal of this work is to develop a predictive modeling framework that can provide high-resolution $PM_{2.5}$ concentrations from low resolution CTM simulations, eliminating the need to prepare high resolution emissions inventories and perform computationally expensive high resolution CTM simulations. The proposed modeling framework notably does not include any observations of $PM_{2.5}$ concentrations. While the addition of this type of data would likely improve the ability of the model to predict total $PM_{2.5}$ in some cases, this would extend beyond the scope of the current work which is to reproduce the 1 × 1 km^2 CTM source- and species-resolved $PM_{2.5}$ concentrations. Future work can include the addition of measurements to the proposed approach. We explore the application of several machine learning models for this task and evaluate the predicted speciated and source-resolved $PM_{2.5}$ concentrations by comparing them to the actual 1 × 1 km^2 resolution output of PMCAMx (Garcia Rivera et al., 2022; Dinkelacker et al., 2022). This work can be viewed as a first test of the feasibility of developing a generalizable downscaling approach using high-resolution CTM predictions in selected areas.

2. Model description

The information flow of the downscaling model developed in this study is illustrated in Fig. 1. Inputs to the downscaling model include low-resolution (36 × 36 km^2) source and species resolved $PM_{2.5}$ concentration predictions from PMCAMx, meteorological variables simulated using WRF, and LU variables described below. The downscaling model is trained and validated using high-resolution (1 × 1 km^2) source- and species-resolved $PM_{2.5}$ predictions from PMCAMx. These source- and species-resolved high-resolution $PM_{2.5}$ predictions are the target variables for the models described in the following sections.

2.1. Chemical transport model predictions

Speciated and source-resolved predictions of $PM_{2.5}$ in southwestern Pennsylvania at 36 × 36 km^2 and 1 × 1 km^2 grid resolution for February and July 2017 are available based on the work of Garcia Rivera et al. (2022). The high resolution 1 × 1 km^2 domain in southwestern Pennsylvania is used as the spatial domain for the following analyses. These concentration fields were produced using the Particulate Matter Comprehensive Air Quality Model with Extensions (PMCAMx) (Karydis et al., 2010; Murphy and Pandis, 2010; Tsimpidi et al., 2010), a state-of-the-science CTM that uses the framework of the CAMx model (Environ, 2006). Simulated components of $PM_{2.5}$ include primary organic aerosol (POA), secondary organic aerosol (SOA), elemental carbon (EC), crustal mass (CRST), nitrate, ammonium, and sulfate. Detailed descriptions of PMCAMx can be found in Fountoukis et al.

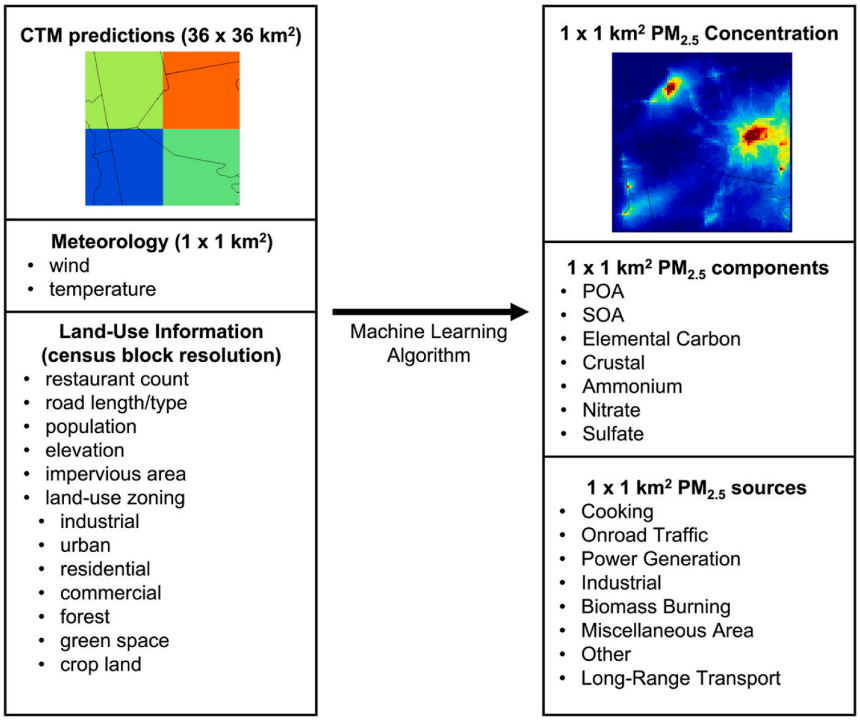

Fig. 1. Information flow for resolution downscaling of chemical transport model $PM_{2.5}$ concentration predictions. All predictors and target variables are of daily temporal resolution, with the exception of land-use information which is static.

(2011) and Zakoura and Pandis (2019). Seven emissions source categories are considered: (1) biomass burning, (2) cooking, (3) on-road vehicles, (4) power generation, (5) industrial activities, (6) miscellaneous area sources (solvent utilization, storage/transport of petroleum products, dry cleaning, waste disposal/incineration), and (7) "other" emissions sources (e.g. agricultural dust, river barges, off-road vehicles and equipment, rail activity, and oil-gas activities). Additionally, the contribution to $PM_{2.5}$ concentrations from sources located outside of the simulation domain is quantified and referred to as long-range transport (LRT) $PM_{2.5}$. The average contribution to total $PM_{2.5}$ concentrations in February and July 2017 by source are provided in Fig. S1. Daily average PMCAMx predictions of $PM_{2.5}$ concentration from 56 species/source combinations (i.e., six chemical components as well as total $PM_{2.5}$; seven source categories as well as LRT) are available for the 5184 km^2 simulation domain at 36 × 36 km^2 and 1 × 1 km^2 grid resolution. The simulation area is comprised of a large portion of southwestern Pennsylvania (including the urban area of Pittsburgh), as well as parts of eastern Ohio and northern West Virginia. The low-resolution $PM_{2.5}$ predictions are used as input for the downscaling model. while the high-resolution predictions are used to train and validate the downscaling model.

The source- and species-resolved CTM predictions of total $PM_{2.5}$ mass and all major $PM_{2.5}$ components at both grid resolutions have been evaluated using all available regulatory data in the southwestern Pennsylvania as well as aerosol mass spectrometer (AMS) data available from the Carnegie Mellon University Supersite. Importantly, the CTM reproduced urban-rural $PM_{2.5}$ gradients at 1 × 1 km^2 resolution, with performance against stationary monitors improving with increasing grid resolution. Total $PM_{2.5}$ mass concentrations were predicted well in the winter with low fractional error (0.3) and fractional bias (+0.05). In the summer period, total $PM_{2.5}$ was underpredicted (fractional bias = −0.39) due to corresponding underpredictions of organic aerosol (OA). Improvement of biogenic secondary OA formation mechanisms in PMCAMx is in active development to address this issue. Further detailed evaluation of this data can be found in Dinkelacker et al. (2022). The subsequently developed machine learning models inherit the performance of the CTM data used in training, but the methodology described in this work is general and can be applied to any other CTM or take advantage of future improvements to PMCAMx.

2.2. Meteorology

Meteorological fields were first calculated using the Weather Research and Forecasting model (WRF-v3.6.1) with horizontal resolution of 12 × 12 km^2. Initial and boundary conditions were obtained from the ERA-Interim global climate re-analysis database. Required WRF input data including terrain, land-use, and soil type were retrieved from the United States Geological Survey database. WRF output was then interpolated to 1 × 1 km^2 grid resolution. Temperature, wind velocity, and wind direction from these meteorological simulations are used at inputs for the downscaling models tested in this study. An evaluation of the performance of these spatially interpolated meteorological variables was performed to accompany the original analysis of CTM results. A summary of the evaluation of the high-resolution meteorological variables can be found in Dinkelacker et al. (2022). In general, the evaluation showed that the errors in magnitude and phasing of the diurnal cycles of the interpolated variables are appropriately small for use in CTM simulations.

2.3. Land use variables

The LU variables used in this study are from a database compiled by Kim et al. (2020). For each variable, values are available for every census block across the contiguous United States at a range of buffer sizes (100 m–15 km^2), referring to the radius of aggregation of that variable with respect to the census block centroid. In total, we selected 24 variables for use in this study (as inputs to the downscaling models): road length and type, restaurant count, population, elevation,

impervious land area, and 14 characterizations of land use (e.g. residential land, industrial land, urban green space, forested land, etc.). To map the data to the 1 × 1 km^2 resolution simulation grid, data from the nearest census block to each computational cell in the downscaling model was used (smallest centroid-to-centroid distance). For each variable, we selected only one radius: when a 1000 m radius was available in the dataset, we used that value; a 3000 m radius was used for four variables (length of A2 road, industrial area, mixed-urban area, and residential area) when 1000 m data was not available.

2.4. Algorithms used for resolution downscaling

Four algorithms were tested in this study to determine their suitability as estimators for high-resolution, source- and species-resolved $PM_{2.5}$ predictions. (1) A multiple linear regression (MLR) scheme was tested as a zeroth level approach. With the MLR model, an individual linear model is fit to describe the relationship between one of the output variables and all of the input variables. This type of approach cannot capture nonlinear relationships between inputs and outputs. (2) A 3-hidden layer artificial neural network (NN) similar to the architecture used in previous studies (Di et al., 2016; Vlasenko et al., 2021), where the model includes multiple layers of nodes. The output of each node is computed by some non-linear function of its inputs. The weights that determine these functions are updated during training, resulting in a more accurate model. (3) A simple decision tree (DT) framework (Breiman et al., 1984) was tested. In this framework, the dataset gets divided into smaller and smaller subsets based on subsequent conditional statements. These conditional statements are made by determining what split can be made on the data that results in the greatest decrease in standard deviation of the data after splitting. (4) A random forest (RF) model (Breiman, 2001) was tested. This type of model is trained to return the average prediction of a large number of decision trees that have been trained on random parts of the training dataset. Use of these machine learning algorithms is motivated by the need for capturing inherent nonlinearities associated with air quality modeling. Scikit-Learn (Pedregosa et al., 2011) implementations for linear regression, decision tree, and random forest models were used. The Keras TensorFlow (Abadi et al., 2015) package was used to formulate the neural network model.

In all cases, a scaling function was applied to all variables ($PM_{2.5}$ concentrations, meteorological variables, LUR variables) of the downscaling model in order to set the range of these variables to [−1, 1]. This is done to prevent training bias towards a variable with a larger range of values. For example, the value of population ranges from 0 to 18000 people, while the value of $PM_{2.5}$ POA at coarse resolution (36 × 36 km^2) ranges from 0 to 4 μg m^{-3} even though both of these inputs are provided to the same model. These input variables may be of equal importance for making predictions, but the magnitudes of their values differ tremendously. The variable with the higher magnitude will tend to have a disproportionate impact on model predictions, so scaling is important before training a predictive model. The variables were scaled according to:

$$X_{scaled} = \frac{2(X - X_{min})}{X_{max} - X_{min}} - 1 \quad (1)$$

where X_{scaled} is the scaled vector of all values for a single variable, X is the vector pre-scaling, $X_{\min}$ is the minimum value in the vector, and $X_{\max}$ is the maximum value in the vector. After predictions are made, an inverse function is applied to calculate the predicted concentrations to their original ranges.

For model training, the high-resolution cells were randomly divided into ten 10%–90% splits. Models were trained on the 90% splits and tested on the remaining 10% of cells. Training and testing were repeated on each set (ten total) and performance metrics were averaged for the formal analysis. Cells were randomly split at first, however nonrandom splits were used as a tool for model evaluation. Data from both February and July 2017 periods are included in the training dataset, with the only obvious indicator variable of which month a data point is from being temperature. The same model is used for both February and July to strive towards temporal generalizability. To visually present the ability of the model to reproduce the spatial distribution of source-resolved $PM_{2.5}$ and its components, five unique and random 20%–80% splits were used so that the test predictions cover the entire spatial domain. This enables map generation while avoiding making predictions on grid cells that exist in any of the training sets.

2.5. Predictive performance metrics

All predictions made on testing data inputs are evaluated by comparing predictions of species- and source-resolved $PM_{2.5}$ with the daily high-resolution PMCAMx simulation results. Evaluation metrics considered are the normalized root mean squared error (NRMSE), normalized mean error (NME), and normalized mean bias (NMB):

$$NRMSE = \sqrt{\frac{\sum_{k=1}^{N}(P_k - O_k)^2}{N}} \Big/ O_{mean} \quad (2)$$

$$NMB = \sum_{k=1}^{N}(P_k - O_k) \Big/ \sum_{k=1}^{N} O_k \quad (3)$$

$$NME = \sum_{k=1}^{N}|P_k - O_k| \Big/ \sum_{k=1}^{N} O_k \quad (4)$$

where P_k is the kth predicted daily value of the output $PM_{2.5}$ concentration (aggregated by species or source), O_k is the kth "true" daily value of the $PM_{2.5}$ concentration, N is the total number of data points, and O_{mean} is the mean value of the "true" $PM_{2.5}$ concentration. In the context of this methodology, the "true" value refers to the $PM_{2.5}$ concentration that was calculated by the CTM at 1 × 1 km^2 resolution. The evaluation metrics are calculated after performing the necessary inverse variable scaling. We evaluate here the ability of each algorithm to downscale predictions of each $PM_{2.5}$ component as well as total $PM_{2.5}$ from each source category.

3. Results and discussion

3.1. Algorithm screening

Fig. 2 illustrates the NRMSE of test set predictions of $PM_{2.5}$ components and total $PM_{2.5}$ mass. As expected, the MLR model has overall the weakest performance in terms of NRMSE. The linear model does exhibit better performance with NRMSE <0.1 for SOA predictions (NRMSE <0.1) because this component has much less variation in space compared to other $PM_{2.5}$ components. The NN and DT algorithms offer significant improvements compared to the linear model. NN and DT produce similar results in terms of NRMSE by species, with the neural network predicting some species (POA, EC, CRST) with slightly lower error and the decision tree performing better with the rest (SOA, ammonium, nitrate, sulfate). The RF algorithm predicts all species with the lowest error, although in most cases the difference is small compared to NN and DT. For total $PM_{2.5}$ predictions, the NRMSE achieved with the random forest algorithm is around 0.01 less than that of the neural network and decision tree algorithms.

The performance (NRMSE) of the various algorithms for source-resolved $PM_{2.5}$ predictions is summarized in Fig. 3. NME and NMB values are available in Table S2 in the supplementary material. The linear model again has the worst performance, while the NN and RF algorithms have the best. The NRMSE of predictions made using RF is lower than NN for power generation (0.2 lower), but higher for cooking

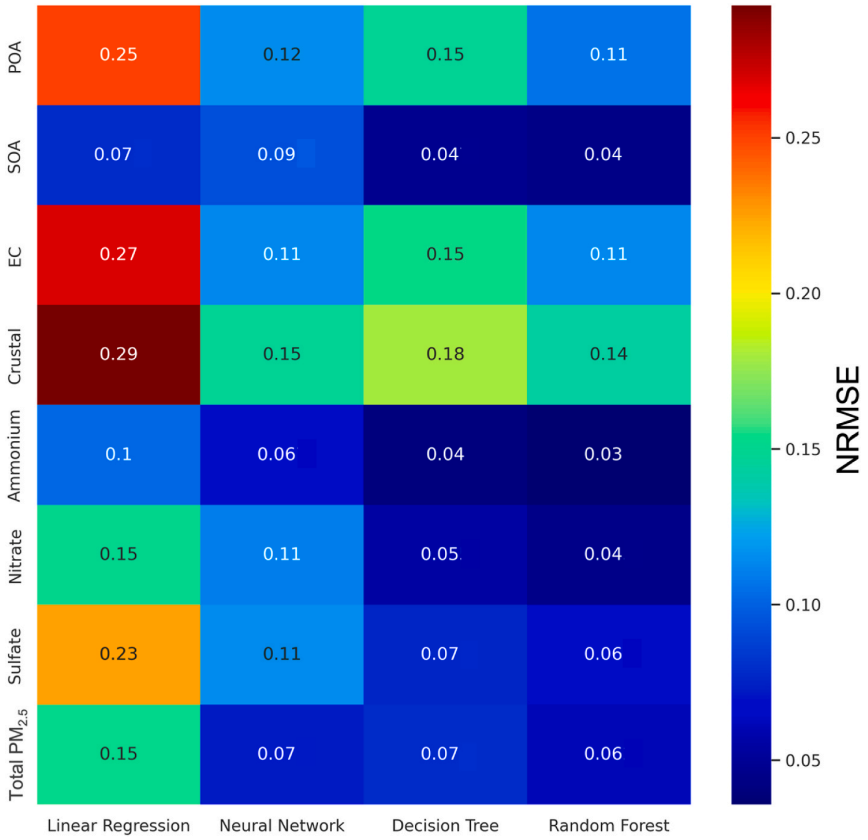

Fig. 2. Normalized root mean squared error (NRMSE) of testing set $PM_{2.5}$ predictions by species for various algorithms when compared to daily high-resolution PMCAMx predictions.

(0.2 higher) and industrial sources (0.13 higher). The rest of the predictions are similar or show slightly better performance using the RF. Scatter plots depicting the performance of each algorithm with regards to total $PM_{2.5}$ concentration are shown in Fig. 4. Total $PM_{2.5}$ is predicted very well by the NN, DT, and RF models (NME <0.05; |NMB| < 0.01). For cooking $PM_{2.5}$, the NN and RF algorithms predictions have little absolute normalized bias (less than 0.012). With industrial $PM_{2.5}$ we see enormous improvements from the linear model, especially at high concentrations. Here, the lowest bias (|NMB| = 0.002–0.003) is achieved using any of the NN, DT, or RF algorithms.

Our analysis up to this point supports the use of the random forest algorithm due to consistently low error and bias in predictions, regardless of $PM_{2.5}$ species or source category. Additional scatter plots for all other $PM_{2.5}$ species and source categories are included in the supplementary material (Figs. S2–S12). In the remaining sections we will focus on the results of the RF algorithm.

3.2. Spatial distribution of monthly averaged predictions

In general, the downscaling model reproduces the spatial distribution of $PM_{2.5}$ in southwestern Pennsylvania well for both February and July 2017. The downscaling model captures both the peak $PM_{2.5}$ concentrations in downtown Pittsburgh and those near the industrial facilities in the northwestern part of the domain (Fig. 5). The urban to rural transition outside of the city of Pittsburgh is also captured. Some inconsistencies on the order of 0.1 μg m^{-3} are seen with $PM_{2.5}$ species with considerable emissions from large point sources. An example of this is $PM_{2.5}$ sulfate which has large emissions from power generation in the upper left and lower left corners of the domain (Fig. 6).

The ability of the downscaling model to differentiate between two distinct time periods is shown in the monthly average maps for biomass burning $PM_{2.5}$ (Fig. 7). This source shows the greatest variability between months. In February the domain-average, minimum, and maximum concentration of biomass burning $PM_{2.5}$ as predicted by PMCAMx are 0.8 μg m^{-3}, 0.02 μg m^{-3}, and 3.31 μg m^{-3} respectively. In July, the corresponding values are 0.007 μg m^{-3}, 0.0005 μg m^{-3}, and 0.036 μg m^{-3}. In February, the downscaling model reproduces both the large peak values as well as the large sections of the domain where the concentration is low. In July, it is encouraging that the downscaling model does not predict any large concentrations of biomass burning $PM_{2.5}$ anywhere throughout the simulation domain. With power generation $PM_{2.5}$, small inconsistencies in predictions follow those seen with $PM_{2.5}$ sulfate (Fig. 8). This occurs because the emissions from sources like power generation originate from just a few large localized sources, but travel large distances across the entire domain. The LUR variables provide the downscaling model with high-resolution information about emissions sources (restaurants, population, industrial land-use area, etc.) within a radius of 1 km. This means that far away from a large point source, the machine learning algorithm has no direct knowledge of an emissions source like a power plant that could be impacting $PM_{2.5}$ concentrations at the location of interest far beyond the 1 km radius. Despite this lack of information, the RF model reproduces well the major plumes of $PM_{2.5}$ from power generation sources, however with the small inconsistencies present throughout the domain. This is due to the random nature of the training and testing data splits. Increasing the radius of LUR variables such as industrial land use may improve performance for these sources. Maps of monthly average $PM_{2.5}$ concentrations for all other major species and sources are included in the supplementary material (Figs. S13–S23).

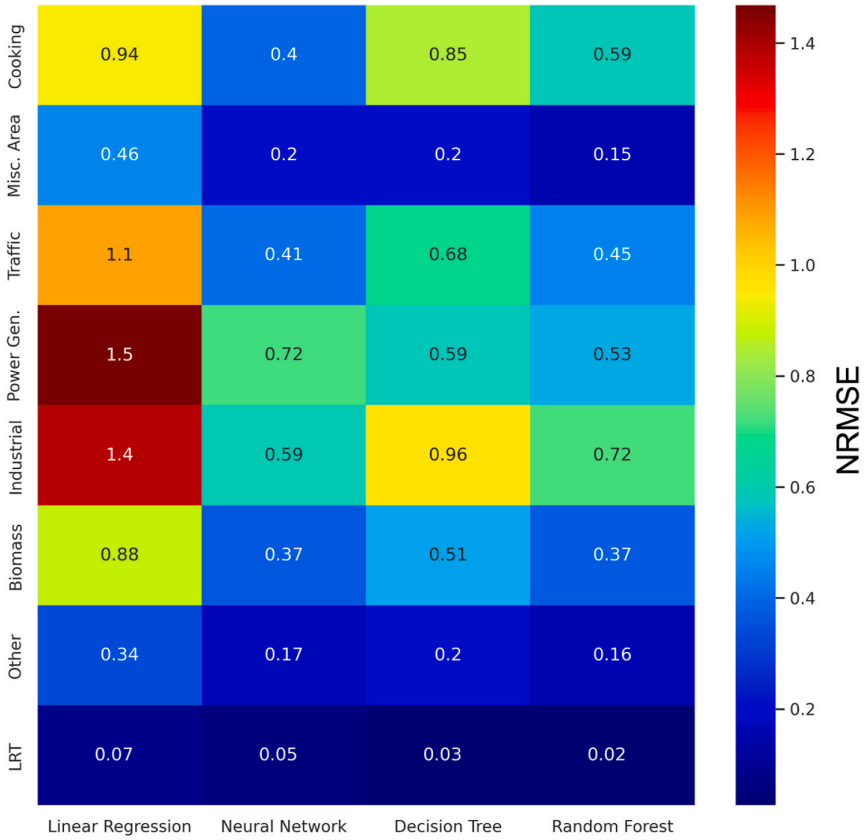

Fig. 3. Normalized root mean squared error (NRMSE) of testing set $PM_{2.5}$ predictions by source category for various algorithms when compared to daily high-resolution PMCAMx predictions.

3.3. Additional model testing for "new" areas

As a preliminary test of the generalizability of this downscaling approach, we trained the model by holding out entire quadrants of the domain as testing data, rather than a randomly selected subset of 25% of the 1×1 km^2 cells. The training of the model did not include any information for this held-out area. The quadrant-based approach was chosen to challenge the model in terms of predicting $PM_{2.5}$ concentrations in unknown, relatively large, contiguous domains while also aligning with the grid cells in the low-resolution CTM predictions. This procedure was followed for each quadrant, and the results of this test are shown in Fig. 9 (February) and Fig. 10 (July). In the northwest and southwest quadrants, we see the expected effects from long range sources of $PM_{2.5}$. The downscaling model has difficulties in reproducing large emissions plumes due to the relatively small radius associated with the land-use information (1–3 km) compared to the range of the visible plumes (up to 40 km). In the northeast quadrant, the downscaling model tends to spread the effects of the downtown Pittsburgh area to a wider area. The downtown area is a unique feature in this domain, so it is not surprising that the downscaling model has difficulty reproducing the spatial distribution of $PM_{2.5}$ in this area without an additional urban center in the training data. The downscaling model reproduces the spatial distribution of $PM_{2.5}$ in the southeast quadrant, which has both significant population and industrial activity. With its obvious challenges to reproduce certain large emissions features (large emissions stacks, discrete urban centers, etc.), the model clearly differentiates between areas that are primarily urban, rural, or industrial when faced with the challenge of predicting over contiguous subdomains that contain valuable information for training a machine learning model. This result is encouraging ahead of the application of this downscaling model to other locations in the United States. The improvement of the treatment of the effects of large point emission sources will be a topic of future study.

4. Conclusions

Four algorithms were tested to downscale the resolution of CTM (PMCAMx) predictions of source-resolved speciated $PM_{2.5}$ concentration from a coarse resolution of 36×36 km^2 grid cell size to 1×1 km^2 in southwestern Pennsylvania during the months of February and July 2017. Additional inputs to the downscaling model include high-resolution meteorological variables and LU variables to capture high-resolution spatial variability in emissions sources. The RF approach reproduces the high-resolution CTM output for all $PM_{2.5}$ species except crustal $PM_{2.5}$ with low NRMSE (<0.11), with the most obvious inconsistencies present with species associated with large point sources such as $PM_{2.5}$ sulfate, a major component of power generation emissions. The RF model does especially well at predicting total $PM_{2.5}$ (NRMSE <0.1, NMB <0.001). The RF model also does well at capturing the major emissions features in the source-resolved predictions. In particular, the ability of the model to accurately reproduce $PM_{2.5}$ from sources that are significantly different in contribution between the two months is encouraging. The RF algorithm was chosen to further test the ability of a predictive model to reproduce high-resolution predictions of speciated and source-resolved $PM_{2.5}$ as predicted by PMCAMx. Minor inconsistencies in model predictions are only seen for $PM_{2.5}$ from sources that have long-range effects far from the point of emission. With these sources (power generation, industrial), the 1–3 km radius for land-use information may not be far enough for the model to learn well the long-range effects. To improve performance with these sources, larger

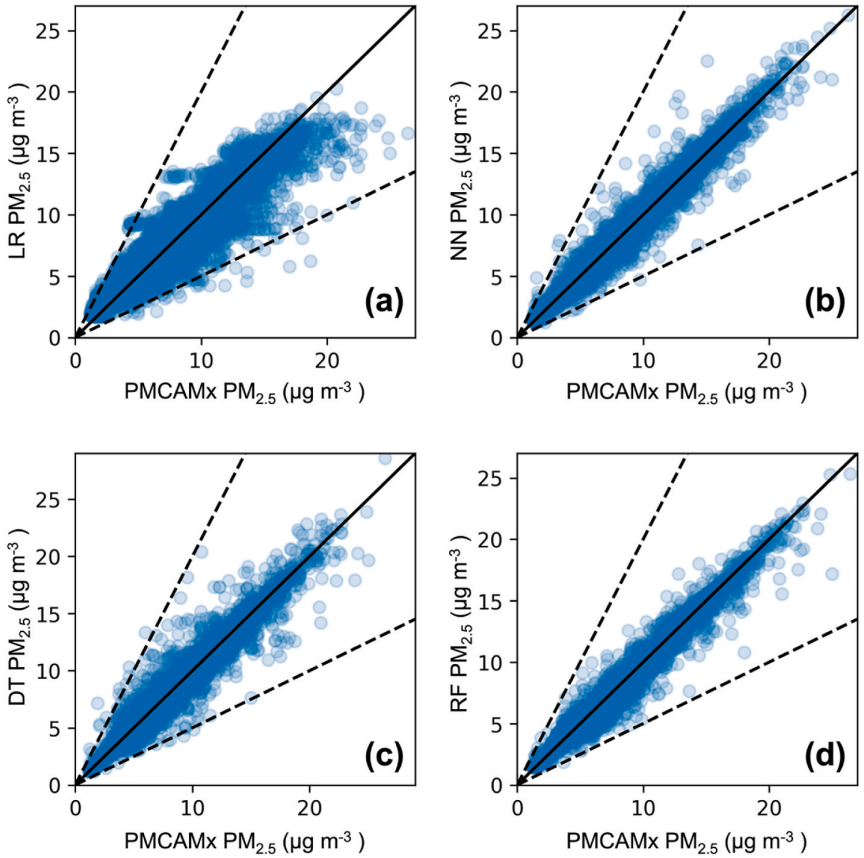

Fig. 4. Prediction performance of (a) linear regression, (b) neural network, (c) decision tree, and (d) random forest algorithms on test set total $PM_{2.5}$ concentrations compared to the "true" daily concentrations predicted by PMCAMx. n = 26950.

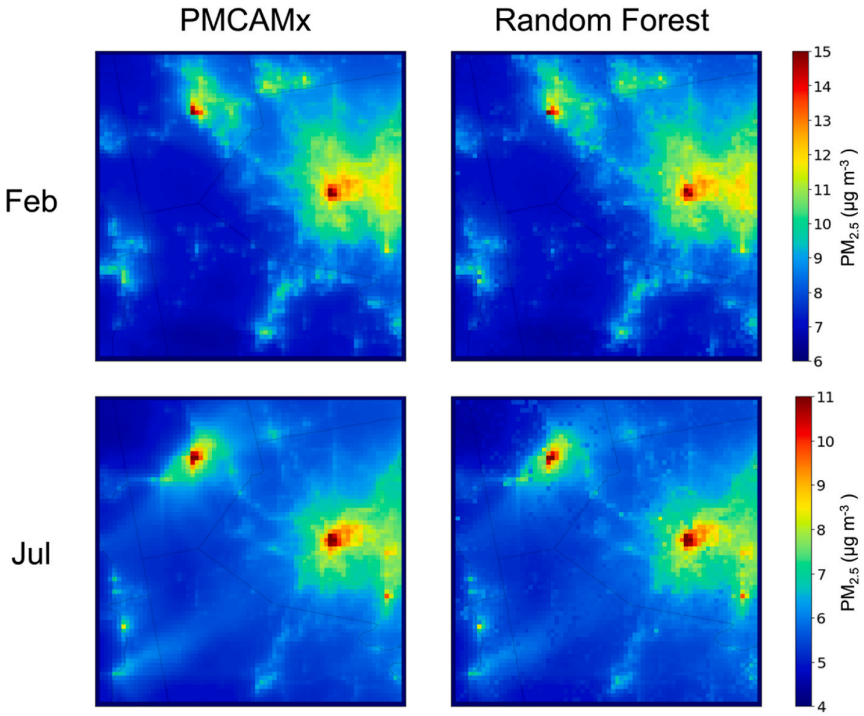

Fig. 5. Maps of monthly average ground level total $PM_{2.5}$ for February and July 2017 as predicted by PMCAMx and the random forest model.

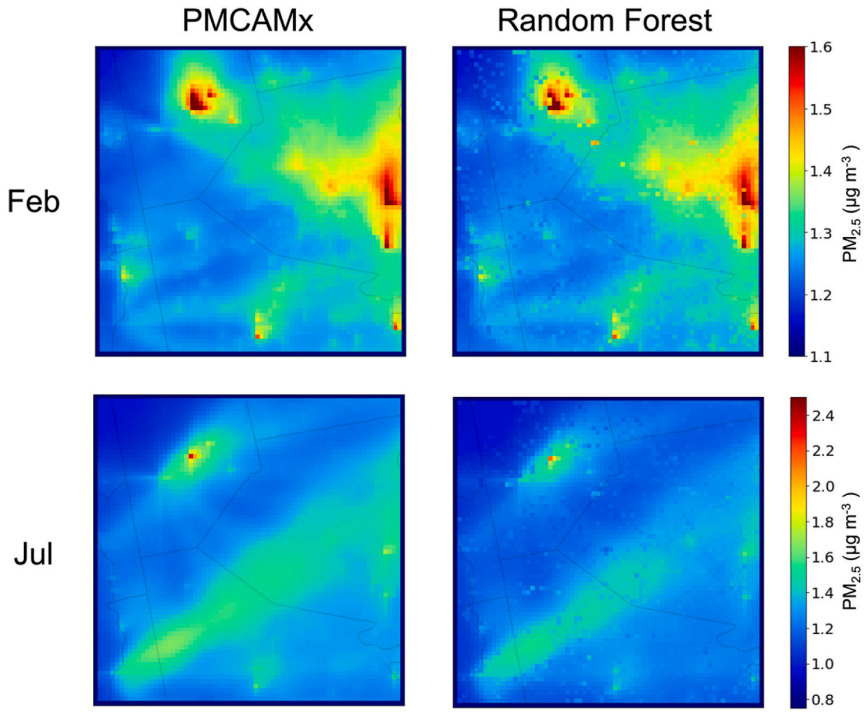

Fig. 6. Maps of monthly average ground level $PM_{2.5}$ sulfate for February and July 2017 as predicted by PMCAMx and the random forest model.

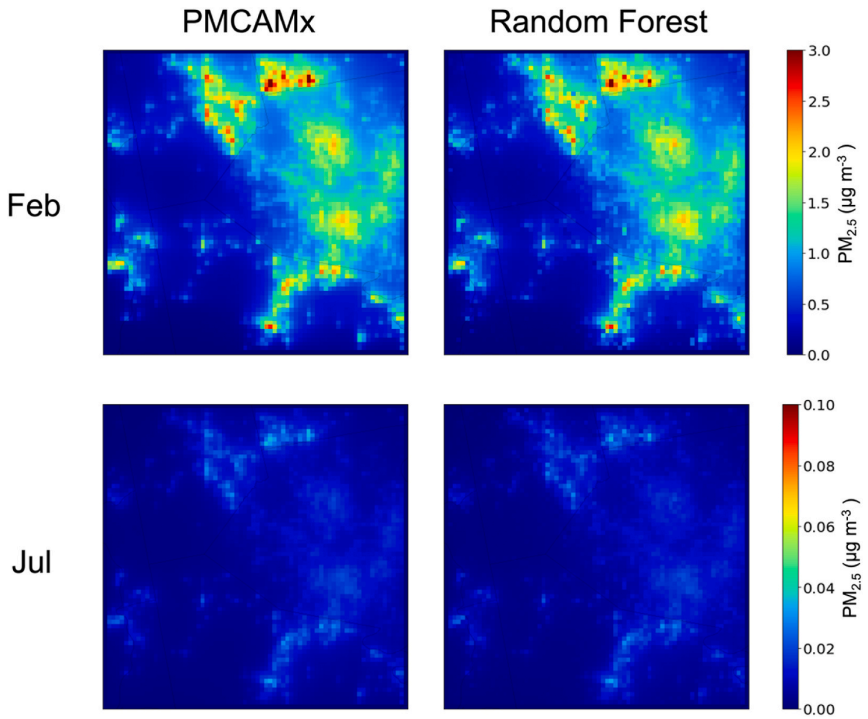

Fig. 7. Maps of monthly average ground level total $PM_{2.5}$ from biomass burning sources for February and July 2017 as predicted by PMCAMx and the random forest model.

radii on the order of the range of these sources (~10 km) will be tested in future work. While the performance of the downscaling model is encouraging, it is important that the model be viable in other locales provided sufficient input information. Preliminary tests of model generalizability were explored within the southwestern Pennsylvania domain by training downscaling models using entire domain quadrants as testing data. The downscaling model struggles to reproduce large emissions plumes in the northwest and southwest quadrants, again

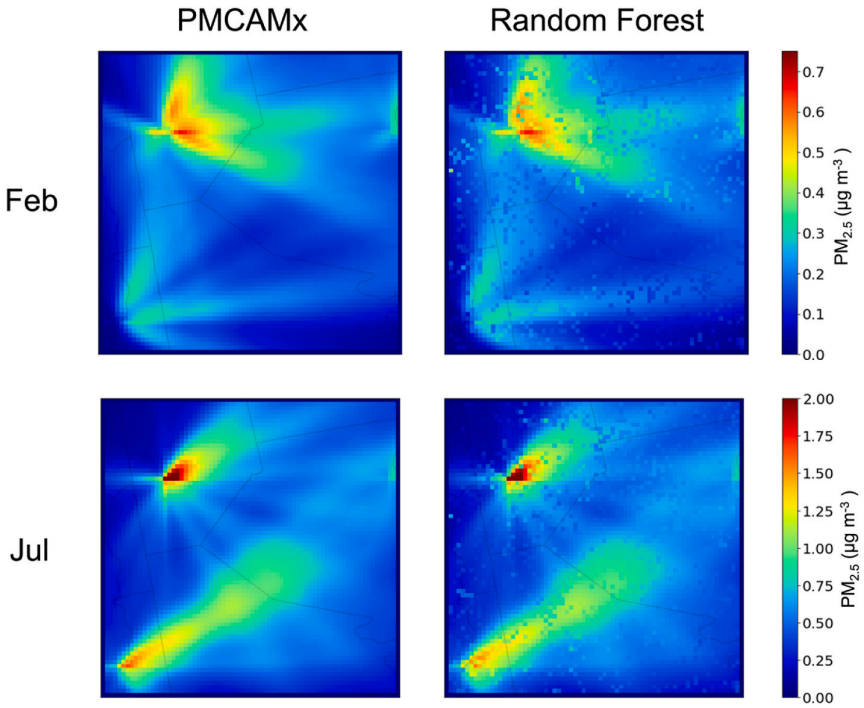

Fig. 8. Maps of monthly average ground level total $PM_{2.5}$ from power generation sources for February and July 2017 as predicted by PMCAMx and the random forest model.

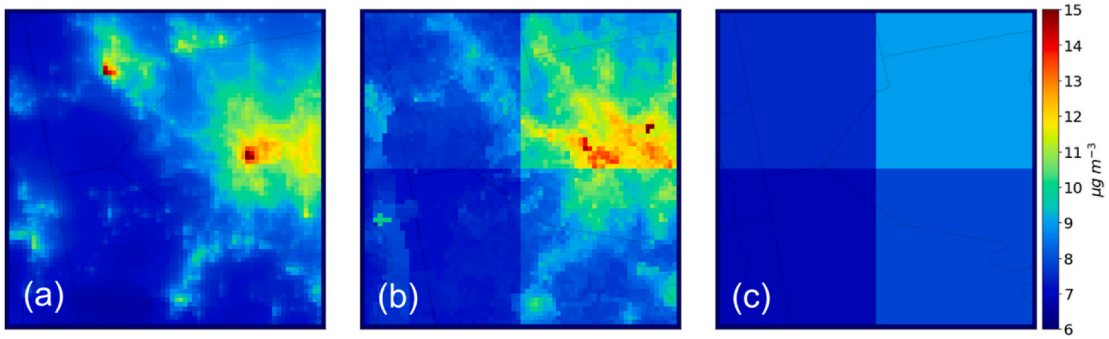

Fig. 9. Maps of monthly average ground level total $PM_{2.5}$ in southwestern Pennsylvania for February 2017 as predicted by (a) PMCAMx high-res, (b) RF model with entire quadrant testing sets, and (c) PMCAMx low-res.

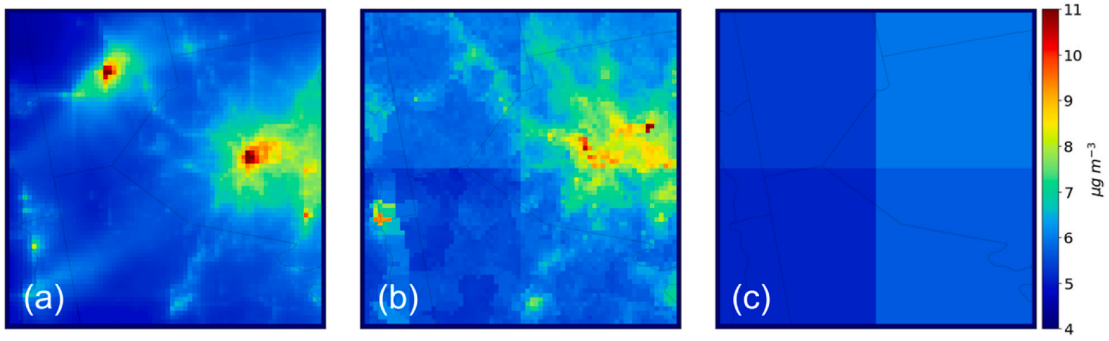

Fig. 10. Maps of monthly average ground level total $PM_{2.5}$ in southwestern Pennsylvania for July 2017 as predicted by (a) PMCAMx high-res, (b) RF model with entire quadrant testing sets, and (c) PMCAMx low-res.

suggesting the need for longer range land-use information for power generation and industrial sources. The downscaling model also tends to spread out the urban $PM_{2.5}$ in the northeast quadrant. The ability of the downscaling model to differentiate between largely urban, rural, and industrial areas is encouraging, understanding that important features had to be removed from training datasets to carry out these tests. Future work will also involve compiling the requisite input data for application of this model to other cities in the United States, where the predictions will be evaluated against available $PM_{2.5}$ measurements.

Financial support

This work was supported by the Center for Air, Climate, and Energy Solutions (CACES) which was supported under Assistance Agreement No. R835873 awarded by the U.S. Environmental Protection Agency and the Horizon-2020 Project REMEDIA of the European Union under grant agreement No 874753.

CRediT authorship contribution statement

BTD performed the PMCAMx simulations, prepared all data for downscaling models, implemented the predictive models in Python, analyzed the results, and wrote the manuscript. PGR prepared anthropogenic emissions and other inputs for the PMCAMx simulations and assisted in preparation of PMCAMx output. SNP, PJA, and JDM designed and coordinated the study and helped in the writing of the paper. All authors reviewed and commented on the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2023.119967.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moores, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems software available at: tensorflow.org.

Bai, Y., Li, Y., Wang, X., Xie, J., Li, C., 2016. Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. Atmos. Pollut. Res. 7, 557–566. https://doi.org/10.1016/j.apr.2016.01.004.

Banzhaf, S., Ma, L., Timmins, C., 2019. Environmental justice: the economics of race, place, and pollution. JEP (J. Environ. Psychol.) 33, 185–208. https://doi.org/10.1257/jep.33.1.185.

Breiman, L., Freidman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees, first ed. Routledge. https://doi.org/10.1201/9781315139470.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Di, Q., Koutrakis, P., Schwartz, J., 2016. A hybrid prediction model for $PM_{2.5}$ mass and components using a chemical transport model and land use regression. Atmos. Environ. 131, 390–399. https://doi.org/10.1016/j.atmosenv.2016.02.002.

Dinkelacker, B.T., Garcia Rivera, P., Kioutsioukis, I., Adams, P.J., Pandis, S.N., 2022. Evaluation of high-resolution predictions of fine particulate matter and its composition in an urban area using PMCAMx-v2.0, Geosci. Model Dev. 15, 8899–8912. https://doi.org/10.5194/gmd-15-8899-2022.

Dockery, D.W., Pope, C.A., 1994. Acute respiratory effects of particulate air pollution. Annu. Rev. Publ. Health 15, 107–132. https://doi.org/10.1146/annurev.pu.15.050194.000543.

Environ, 2006. Comprehensive Air Quality Model with Extensions Version 4.40, Users Guide. ENVIRON Int. Corp., Novato, CA. http://www.camx.com. May, 2022.

Fountoukis, C., Racherla, P.N., Denier van der Gon, H.A.C., Polymeneas, P., Charalampidis, P.E., Pilinis, C., Wiedensohler, A., Dall'Osto, M., O'Dowd, C., Pandis, S.N., 2011. Evaluation of a three-dimensional chemical transport model (PMCAMx) in the European domain during the EUCAARI May 2008 campaign. Atmos. Chem. Phys. 11, 10331–10347. https://doi.org/10.5194/acp-11-10331-2011.

Garcia Rivera, P., Dinkelacker, B.T., Kioutsioukis, I., Adams, P.J., Pandis, S.N., 2022. Source-resolved variability of fine particulate matter and human exposure in an urban area. Atmos. Chem. Phys. 22 https://doi.org/10.5194/acp-22-2011-2022.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., Briggs, D., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. 42, 7561–7578. https://doi.org/10.1016/j.atmosenv.2008.05.057.

Karimian, H., Li, Q., Wu, C., Qi, Y., Mo, Y., Chen, G., Zhang, X., Sachdeva, S., 2019. Evaluation of different machine learning approaches to forecasting $PM_{2.5}$ mass concentrations. Aerosol Air Qual. Res. 19, 1400–1410. https://doi.org/10.4209/aaqr.2018.12.0450.

Karydis, V.A., Tsimpidi, A.P., Fountoukis, C., Nenes, A., Zavala, M., Lei, W., Molina, L.T., Pandis, S.N., 2010. Simulating the fine and coarse inorganic particulate matter concentrations in a polluted megacity. Atmos. Environ. 44, 608–620. https://doi.org/10.1016/j.atmosenv.2009.11.023.

Kim, S., Bechle, M., Hankey, S., Sheppard, L., Szpiro, A.A., Marshall, J.D., 2020. Concentrations of criteria pollutants in the contiguous U.S., 1979-2015: role of prediction model parsimony in integrated empirical geographic regression. PLoS One 15, e0228535. https://doi.org/10.1371/journal.pone.0228535.

Murphy, B.N., Pandis, S.N., 2010. Exploring summertime organic aerosol formation in the eastern United States using a regional-scale budget approach and ambient measurements. J. Geophys. Res. 115 https://doi.org/10.1029/2010JD014418.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. JMLR 12, 2825–2830.

Prasad, K., Gorai, A.K., Goyal, P., 2016. Development of ANFIS models for air quality forecasting and input optimization for reducing the computational cost and time. Atmos. Environ. 128, 246–262. https://doi.org/10.1016/j.atmosenv.2016.01.007.

Rahman, M.M., Thurston, G., 2022. A hybrid satellite and land use regression model of source-specific $PM_{2.5}$ and $PM_{2.5}$ constituents. Environ. Int. 163 https://doi.org/10.1016/j.envint.2022.107233.

Seinfeld, J.H., Pandis, S.N., 2006. Atmospheric Chemistry and Physics: from Air Pollution to Climate Change, second ed. John Wiley and Sons, Inc., Hoboken. ISBN: 978-0471720188.

Shtein, A., Kloog, I., Schwartz, J., Silibello, C., Michelozzi, P., Gariazzo, C., Viegi, G., Forastiere, F., Karnieli, A., Just, A.C., Stafoggia, M., 2019. Estimating daily $PM_{2.5}$ and PM_{10} over Italy using an ensemble model. Environ. Sci. Technol. 54 (1), 120–128. https://doi.org/10.1021/acs.est.9b04279.

Tsimpidi, A.P., Karydis, V.A., Zavala, M., Lei, W., Molina, L.T., Ulbrich, I.M., Jimenez, J. L., Pandis, S.N., 2010. Evaluation of the volatility basis-set approach for the simulation of organic aerosol formation in the Mexico City metropolitan area. Atmos. Chem. Phys. 10, 525–546.

Vlasenko, A., Matthias, V., Callis, U., 2021. Simulation of chemical transport model estimates by means of a neural network using meteorological data. Atmos. Environ. 254 https://doi.org/10.1016/j.atmosenv.2021.118236.

Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A., Kolehmainen, M., 2011. Intercomparison of air quality data using principal component analysis, and forecasting of PM_{10} and $PM_{2.5}$ using artificial neural networks, in Thessaloniki and Helsinki. Sci. Total Environ. 409, 1266–1276. https://doi.org/10.1016/j.scitotenv.2010.12.039.

Wu, C., Lin, H., Ho, C., Yang, T., Chen, C., Chang, C., 2014. Modeling horizontal and vertical variation in intraurban exposure to $PM_{2.5}$ concentrations and compositions. Environ. Res. 133, 96–102. https://doi.org/10.1016/j.envres.2014.04.038.

Xue, T., Zheng, Y., Tong, D., Zheng, B., Li, X., Zhu, T., Zhang, Q., 2019. Spatiotemporal continuous estimates of $PM_{2.5}$ concentrations in China, 2000-2016: a machine learning method with inputs from satellites, chemical transport model, and ground observations. Environ. Int. 123, 345–357. https://doi.org/10.1016/j.envint.2018.11.075.

Yu, W., Ye, T., Zhang, Y., Xu, R., Lei, Y., Chen, Z., 2023. Global estimates of daily ambient fine particulate matter concentrations and unequal spatiotemporal distribution of population exposure: a machine learning modelling study. Lancet Planet. Health 7, E209–E218. https://doi.org/10.1016/S2542-5196(23)00008-6.

Zakoura, M., Pandis, S.N., 2019. Improving fine aerosol nitrate predictions using a Plume-in-Grid modeling approach. Atmos. Environ. 187, 390–400. https://doi.org/10.1016/j.atmosenv.2019.116887.

Zhou, Y., Chang, F., Chang, L., Kao, I., Wang, Y., Kang, C., 2019. Multi-output support vector machine for regional multi-step-ahead $PM_{2.5}$ forecasting. Sci. Total Environ. 651, 230–240. https://doi.org/10.1016/j.scitotenv.2018.09.111.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.