Published in final edited form as:
Environ Sci Technol. 2022 April 05; 56(7): 3871–3883. doi:10.1021/acs.est.1c04076.

A Flexible Bayesian Ensemble Machine Learning Framework for Predicting Local Ozone Concentrations

Xiang Ren[a,b], **Zhongyuan Mi**[a,c], **Ting Cai**[a], **Christopher G. Nolte**[d], **Panos G. Georgopoulos**[a,b,c,e,*]

[a]Environmental and Occupational Health Sciences Institute (EOHSI), Rutgers University, Piscataway, NJ 08854, USA

[b]Department of Chemical and Biochemical Engineering, Rutgers University, Piscataway, NJ 08854, USA

[c]Department of Environmental Sciences, Rutgers University, New Brunswick, NJ 08901, USA

[d]Center for Environmental Measurement and Modeling, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711, USA

[e]Department of Environmental and Occupational Health and Justice, Rutgers School of Public Health, Piscataway, NJ 08854, USA

Abstract

3D-grid-based chemistry-transport models, such as the Community Multiscale Air Quality (CMAQ) modeling system, have been widely used for predicting concentrations of ambient air pollutants. However, typical horizontal resolutions of nationwide CMAQ simulations (12×12 km^2) cannot capture local scale gradients for accurately assessing human exposures and environmental justice disparities. In this study, a Bayesian Ensemble Machine Learning (BEML) framework, that integrates thirteen learning algorithms, was developed for downscaling CMAQ estimates of ozone daily maximum 8-hr averages to census tract level, across the contiguous US and is demonstrated for 2011. Three-stage hyperparameter tuning and targeted validations were designed to ensure the ensemble model's ability to interpolate, extrapolate, and capture concentration peaks. The Shapley value metric from coalitional game theory was applied to interpret the drivers of subgrid gradients. The flexibility (transferability) of the 2011-trained BEML model was further tested by evaluating its ability to estimate fine-scale concentrations for other years (2012–2017) without re-training. To

*Corresponding author at: Environmental and Occupational Health Sciences Institute (EOHSI), Rutgers University, Piscataway, NJ 08854, USA. panosg@ccl.rutgers.edu (P.G. Georgopoulos).

Associated Content

Supporting Information

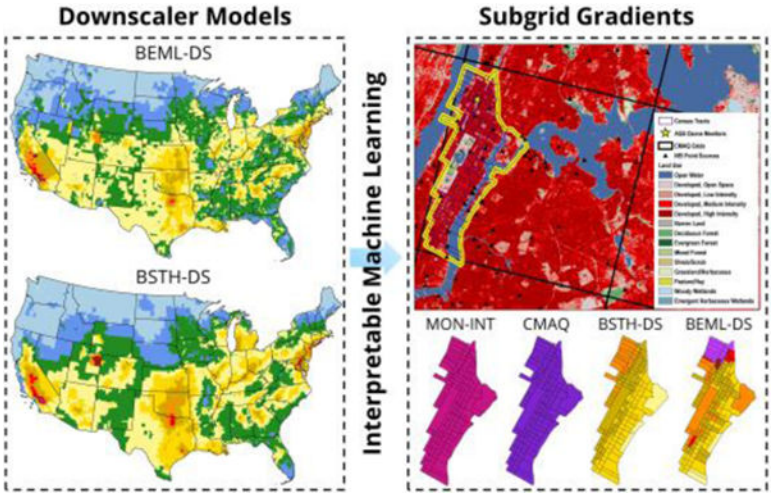
The Supporting Information is available free of charge at <TBD> and includes:
Additional algorithm details, evaluation statistics, and data description, additional explanations to extrapolation, three-stage hyperparameter tuning, and targeted validations, additional validation and supporting results, including subgrid gradient interpretation, downscaling results of simulations for climate change analyses, comparison with BSTH-DS, and spatial prediction maps (2011–2017 and 2051).

An online dashboard for this project is accessible at: https://ccl-eohsi.shinyapps.io/beml_ozone_dashboard/. Data generated for eight years (2011–2017 and 2051) and the code used for model training and validation (thirteen algorithms) are available at https://github.com/xr32/Code-for-es-2021-04076/.

The authors declare no competing financial interest.

demonstrate the feasibility of using the BEML approach to strictly “data-limited” situations, the model was applied to downscale CMAQ outputs for a future year scenario-based simulation that considers effects of variations in meteorology associated with climate change.

Graphical Abstract



Keywords

Ozone; interpretable machine learning; data fusion; exposure assessment; environmental and climate justice

1. Introduction

Ground-level ozone is one of the six criteria air pollutants for which the U.S. Environmental Protection Agency (USEPA) has established National Ambient Air Quality Standards (NAAQS). In 2015, USEPA strengthened the primary (health-based) NAAQS for ozone to 70 ppb and the corresponding Design Value (DV) is calculated as the 3-year average of the annual 4th highest Daily Maximum 8-Hour Average.[1] According to the latest review of the ozone NAAQS,[2] sufficient evidence is available to support the establishment of significant associations of respiratory and metabolic effects with ozone exposures. However, causal relationships with respect to other health outcomes such as cardiovascular, reproductive, and nervous system effects are still inconclusive. In this context, producing accurate fine-scale spatiotemporal concentration surfaces at community and neighborhood scales is important for assessing those potential health effects of long-term and short-term exposures to ambient ozone. Furthermore, there is a critical need to accurately assess environmental justice disparities for disadvantaged communities: such disparities are often “averaged out” over the grid scales of air quality and climate models.

Complementary monitoring networks such as the State or Local Air Monitoring Stations (SLAMS) and the Clean Air Status and Trends Network (CASTNET) have been established across the contiguous US (CONUS) to provide “ground truth” information for ozone exposures.[2] However, there are many areas, particularly rural and suburban locations, where

ozone monitors are not available (Figure S1). To obtain a complete spatial and temporal coverage across the CONUS, the USEPA has developed tools such as the Community Multiscale Air Quality (CMAQ) modeling system, which simulates the complex interactions of atmospheric chemistry and physics for multiple air pollutants (including ozone) at various scales. The reliability and performance of CMAQ have been evaluated extensively through an array of applications;[3] nonetheless, there are two problems typically associated with CMAQ ozone estimates: (1) Due to structural and parametric uncertainties,[4] those estimates for some areas deviate substantially from measured values (Figure S2); (2) Due to limitations in computational resources, CMAQ usually provides estimates at coarse spatial resolutions, ranging from 4×4 km^2 to 36×36 km^2, typically 12×12 km^2.[5] This limits the ability to characterize variations at finer resolutions, such as census tract scales, which are important for population-based exposure studies and essential for environmental justice assessments.

Geostatistical models have been developed to address the above issues, including Bayesian Melding,[6] Bayesian Maximum Entropy,[7] Spatiotemporal Data Fusion,[8] etc. These models incorporate “ground truth” from ozone monitors to improve the performance of chemistry-transport models. USEPA has implemented a Bayesian Spatio-Temporal Hierarchical Downscaler (BSTH-DS) model to downscale CMAQ grid-based estimates to census tracts for air pollutants such as ozone and $PM_{2.5}$ across the CONUS.[9,10] BSTH-DS produces better out-of-sample predictions compared to previous geostatistical methods.[10] However, due to the assumption of Gaussian random fields and fusion of only monitor measurements, the subgrid gradients simulated by BSTH-DS are not expected to fully represent fine-scale concentration gradients across census tracts.[4]

The ensemble method has recently been receiving increased attention for large-scale spatiotemporal estimation of ambient air pollution:[11–16] it involves training a “meta learner” by combining predictions from different single models, called “base learners”, to achieve better and more stable results. The meta learner takes advantage of different algorithms to capture complex patterns present in large spatiotemporal data sets. However, there are four major concerns regarding existing ensemble ML models. First, models in earlier studies were tuned and evaluated to maximize their interpolation ability, while their extrapolation ability was not fully considered. Interpolation-oriented tuning and validation cannot ensure prediction accuracy for many local areas at substantial distances from monitors. Second, previous studies mainly focused on predicting annual/seasonal averages instead of peaks. For air pollutants such as ozone, both averaged and peak concentrations are important exposure-relevant metrics, and should be considered in an ensemble model for achieving global and peak accuracy. Third, spatial patterns of fine-scale concentration learned by ML models are complex and usually difficult to interpret. Currently there is a lack of interpretation tools that can identify drivers of local concentration gradients, for improving transparency and credibility of ensemble models, and for developing pollution control strategies. Finally, previous studies focused on improving model accuracy, typically using cross validation results against ground observations, rather than investigating robustness and transferability. Transferable models are needed for downscaling CMAQ estimates beyond current conditions (e.g., for future years), when ground observations are unavailable: this

would allow use of ensemble ML methods in evaluating air pollution controls that take into account effects of climate change.[17,18]

We developed a Bayesian Ensemble Machine Learning (BEML) framework that flexibly selects base learners from thirteen algorithms to downscale the CMAQ estimates for daily maximum 8-hr average (DM8HA) ozone concentrations to census tract level across the CONUS. We employed three-stage hyperparameter tuning and targeted validation to ensure the ensemble model's ability to interpolate, extrapolate, and capture concentration peaks. The subgrid gradients learned from the BEML downscaler (BEML-DS) are interpreted and quantified using the Shapley value and compared with available state-of-the-art methods such as BSTH-DS. After training the model for year 2011, we tested its transferability in downscaling CMAQ ozone estimates across the CONUS from 2012 to 2017 without "re-training" with local inputs for those years. Finally, based on the positive outcomes of the aforementioned testing, we applied BEML-DS to downscale outcomes from a USEPA future year simulation[17,18] employing CMAQ to explore effects of climate change on photochemical air pollution.

2. Materials and Methods

2.1. Study Domain

The geographical area for our analysis covers the 48 CONUS states and the District of Columbia (Figure S1). Within this area, we used the nine climate regions defined by the National Oceanic and Atmospheric Administration (NOAA),[19] for spatial analysis and data splitting as described in the following. DM8HA estimates for the 72,283 census tracts of the CONUS, consistent with the estimates available from BSTH-DS, were calculated. The temporal domain includes each day of seven consecutive years (2011–2017) and one future year (2051); data from the earliest year (2011) were used for model construction and evaluation, and years 2012–2017 were used for transferability analysis.

2.2. Data

2.2.1. Ozone Monitor Measurements—Observed ozone DM8HA values were obtained from the USEPA Air Quality System (AQS),[20] for 1,482 unique monitors across the CONUS from 2011 to 2017. The 1,313 monitors reporting in 2011 were split into two parts: those reporting at least 75% of valid measurements (708 monitors with 249,197 observations) were used for model training and different internal validations; the remaining 605 monitors, with 125,218 observations independent of the model training process, were used for external validation (Figure S1). The final CONUS-wide model was trained using all 1,313 monitors in 2011 to estimate the census tract-based concentrations for the eight years.

2.2.2. CMAQ Model Simulations—The CMAQ estimates of ozone DM8HA are the primary inputs for the BEML-DS model, since our goal is to downscale CMAQ outputs to census tract level, fusing ground observations and available heterogeneous spatiotemporal information. CMAQ outputs for 2011–2017 were retrieved from the USEPA Remote Sensing Information Gateway Data Inventory[21] corresponding to each 12×12 km^2 cell within a domain consisting of 396×246 grid cells in the ground layer. We also used CMAQ

outputs from simulation for 2051, developed by USEPA researchers with meteorology obtained by dynamically downscaling the RCP8.5 scenario from the Community Earth System Model.[17,18] This simulation provided outputs at 36 km resolution and used unchanged (2011) anthropogenic emissions to isolate meteorological effects of climate change on air quality.[17,18]

2.2.3. Other Spatiotemporal Covariates—We considered the following spatiotemporal factors to enhance the accuracy and spatial resolution of CMAQ estimates: (1) Local meteorological factors, including temperature, relative humidity, solar radiation, precipitation, wind speed and direction; (2) Local land use and land cover, including elevation, population density, and twelve types of land coverage; (3) Local stationary point and nonpoint emissions (CO, SO_2, PM_{10}, $PM_{2.5}$, NO_x, VOC, and NH_3); (4) Local traffic, represented by vehicle miles traveled and road density; (5) Trending variables, including longitude, latitude, and day of the year. In variable selection, population density at both county and census-tract level was used for modeling, while for other factors we calculated buffers with radii of 0.5, 1, 5 and 10 km and selected the buffer variable corresponding to the highest correlations with the monitor measurements. Details on the relevant variables are presented in Table S1 and Ren, et al.[22]

2.3. Bayesian Ensemble Machine Learning Framework

The Bayesian Ensemble Machine Learning (BEML) model combines predictions from various “base learners” within a flexible framework (Figure 1). Base learners were trained with thirteen algorithms, including the Multiple Linear Regression Model (LM), Ridge Regression (RIDGE), Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net Regularization (ELASTICNET), Principal Component Regression (PCR), Partial Least Squares Regression (PLSR), *k*-Nearest Neighbors (KNN), Support Vector Regression (SVR), Back-Propagation Neural Network (BPNN), Deep Neural Network (DNN), Regression Tree (RT), Random Forest (RF), and Extreme Gradient Boosting (XGBOOST). These algorithms were selected from nine representative categories of statistical and machine learning methods, i.e., linear regression, regularization, dimensionality reduction, lazy learning, kernel trick, artificial neural networks, deep learning, decision tree, and ensemble learning. The rationale and configuration for each algorithm are summarized in the Supporting Information (Text S1). In the training process, each base learner was fitted to “learn” the underlying complex patterns between point-based monitor observations and associated area-based covariates. Based on the captured “area-to-point” relationships, concentrations were estimated for census tract centroids.

2.3.1. Three-Stage Hyperparameter Tuning—Three-stage hyperparameter tuning was introduced to balance the model’s ability to interpolate, extrapolate and capture peak concentrations (Text S2). For Stage 1, intrinsic hyperparameters of each base learner were tuned using coarse/fine grid search with expert knowledge (Text S1.0). An extrapolation-oriented 5-fold leave-cluster-out sample set, rather than the commonly used interpolation-oriented random sample hold-out set, was used to assess tuning performance; optimum hyperparameters minimizing the 5-fold validated Root Mean Squared Error (RMSE) were selected. To balance global accuracy and peak accuracy, Stage 2 implements sample weight

tuning, assigning larger weights to peak values and nudging base learners to focus on learning peak patterns.[22] Stage 3 further optimizes the ensemble model via tuning the hyperparameter q (see eq. 6 below) of the BEML meta-learner, with emphasis on trade-offs between interpolation, extrapolation, and peak accuracy, instead of simply minimizing the cross-validated RMSE of the ensemble outcomes that may increase the risk of overfitting.

2.3.2. Statistical Indices—Three statistical indices were constructed to provide metrics of model accuracy, robustness and diversity, and to inform the calculation of ensemble weights for each base learner.

The predicted RMSE assesses model accuracy:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\mathrm{Obs}_n - \mathrm{Pred}_n)^2} \tag{1}$$

where Obs_n and Pred_n are the n^{th} observed and predicted responses, and N is the number of samples for prediction. For base learner i, the RMSE can be variable for different sample sets; therefore, a Robustness Index (RI) was defined to measure robustness/stability of prediction performance:

$$\mathrm{RI}_i = \frac{a}{\overline{\mathrm{Median}(\mathrm{RMSE}_i)}} + \frac{b}{\overline{\mathrm{IQR}(\mathrm{RMSE}_i)}} \tag{2}$$

where $\overline{\mathrm{Median}(\mathrm{RMSE}_i)}$ and $\overline{\mathrm{IQR}(\mathrm{RMSE}_i)}$ denote the normalized median and interquartile range of the RMSE. Herein, we simply set $a = 0.7$, $b = 0.3$ to account for central tendency and spread of predictions.

To simulate the uncertainty of the RMSE mentioned above, a climate region-based resampling method reflecting spatial extrapolation was implemented: Among the 708 monitors, each time we selected monitors from five out of the nine climate regions to train a submodel. Then we used the submodel to calculate the predicted RMSE for the remaining samples using eq. 1. This climate region-based resampling generates 126 different subsample sets and correspondingly 126 RMSEs. The median and interquartile range of the RMSEs were simulated, and finally, the RI was calculated for each base learner with eq. 2 (Figure S3).

Model diversity is important for ensemble learning. Combining accurate base learners with distinct weak learners has been shown to be better than combining similarly strong learners.[23] Accordingly, we defined a Correlation Index (CI) to measure the similarity of base learners (Figure S4):

$$\mathrm{CI}_{ij} = \frac{1}{126}\cdot\sum_{l=1}^{126}\mathrm{RMSE}_i^{(l)}\cdot\mathrm{RMSE}_j^{(l)} \tag{3}$$

where CI_{ij} denotes the CI between base learner i and base learner j and $RMSE_i^{(l)}$ denotes the RMSE of the l^{th} climate region-based subsample set for the i^{th} base learner. The paired RMSEs for any two base learners are derived from the same subsample set. Smaller CI tends to have smaller RMSE (strong learner) with weaker correlation (distinct learner), indicating a good combination for the ensemble model.

2.3.3. BEML Meta Learner—The BEML "meta learner" is designed as a linear combination of thirteen base learners:

$$Pred_{\mathrm{BEML}} = \sum_{k=1}^{13} w_k^l \cdot \mathrm{Pred}_k \tag{4}$$

where Pred_k and $\mathrm{Pred}_{\mathrm{BEML}}$ are predictions from base learner k and the BEML meta learner, respectively, and w_k^l is the ensemble weight for base learner k in location l. Earlier studies have used regression methods to optimize w_k^l ("stacking"). Herein, we treated w_k^l as the posterior probability $\mathrm{Prob}(\mathrm{Learner}\ k \mid \mathrm{Target}, l)$ that base learner k can perfectly emulate the underlying air pollutant system (called target function) in location l:

$$\begin{aligned} w_k^l &= \mathrm{Prob}\left(\mathrm{Learner}\ k \,\middle|\, \mathrm{Target}, l\right) \\ &\propto \mathrm{Likelihood}\left(\mathrm{Learner}\ k \,\middle|\, \mathrm{Target}, l\right) \cdot \mathrm{Prior}(\mathrm{Learner}\ k) \end{aligned} \tag{5}$$

where $\mathrm{Prior}\,(\mathrm{Learner}\ k)$ is the prior probability, and $\mathrm{Likelihood}\,(\mathrm{Learner}\ k \mid \mathrm{Target}, l)$ is the likelihood function. The prior probability is determined via the RI and expert knowledge. In Figure S3, base learners with smaller RI than LM were excluded. RT was also excluded due to its highly variable predictions. For linear models from the same category, the one having the largest RI was selected. Therefore, seven base learners (BPNN, PLSR, ELASTICNET, SVR, DNN, RF, and XGBOOST) were assigned a uniform prior $1/7$, while the remaining were all set to 0. The likelihood is related to the CI:

$$\mathrm{Likelihood}\left(\mathrm{Learner}\ k \,\middle|\, \mathrm{Target}, l\right) = \frac{\sum_{i=1}^{13} \mathrm{CI}_{ki}^{-q}}{\sum_{i=1}^{13}\sum_{j=1}^{13} \mathrm{CI}_{ij}^{-q}} \tag{6}$$

The exponent q in eq. 6 is the hyperparameter in Stage 3, which is tuned by states (l) to account for the potential spatial heterogeneity of the ensemble weights. According to eq. 6, a strong learner with lower correlation with other base learners tends to have a large likelihood and can potentially improve the ensemble performance.

2.3.4. Validation for Base/Meta Learners—Base/meta learners were evaluated with respect to seven validation criteria using the same sample sets, and seven statistics (Text S4), i.e., RMSE, coefficient of determination, mean bias, peak RMSE, Pearson correlation, variance of errors, and variance of model predictions, were calculated for assessing model performance. (Descriptions of the validation strategies are provided in Text S3 of the Supplement)

2.3.5. Interpretation of Subgrid Gradients—Unlike global interpretation tools, such as permutation variable importance that measures variable contribution across all predictions,[22] local interpretation tools aim to quantify variable contribution to individual predictions, helping to understand why a model makes a certain prediction. Local Interpretable Model-diagnostic Explanations (LIME) simulate local variable importance via fitting a number of "easily-interpretable" local surrogate models (e.g., linear models) around given predictions;[24] however this tool does not control the quality of the local fit to the data and becomes biased and unreliable for small sample sizes in local approximation.[25] To alleviate this limitation, a unified framework, called SHapley Additive exPlanations (SHAP), is available.[26] SHAP simulates variable contribution to individual predictions by calculating the Shapley value[27] for each variable per sample. Extensively used in coalitional game theory, Shapley value currently provides a unique solution to satisfying properties (local accuracy, consistency, missingness) desired for explanatory machine learning analysis.[26,28]

The Shapley value was incorporated in our BEML framework to explain local scale gradients. This approach considers the learning/regression process as a game played by covariates, where each can generate "payout", i.e., contribution to the prediction. For a specific sample, the Shapley value of the j^{th} covariate $\varphi^{(j)}$ is the average of marginal contributions across all possible coalitions of covariates; $\varphi^{(j)}$ measures the magnitude of the j^{th} covariate contribution to the prediction compared to the reference level. Herein, the Shapley value for BEML-DS was approximated as the weighted average of Shapley values from the base learners:

$$\varphi_{\text{BEML}}^{(j)} = \sum_{k=1}^{13} w_k^l \cdot \varphi_k^{(j)} \tag{7}$$

where $\varphi_k^{(j)}$ denotes the Shapley value of the j^{th} covariate from base learner k. It should be noted that a feature's Shapley value is a relative value which changes with the reference level selected; the reference level is set as the average of predictions for monitoring sites at specific spatial and temporal domains.

3. Results

3.1. Spatiotemporal Evaluation and Analysis of Subgrid Predictions

The performance of each base/meta learner is presented in Table S4. As expected, 10-fold leave-sample-out validation (interpolation) provided more optimistic results than 10-fold leave-cluster-out validation (extrapolation). Peak accuracy significantly improved after sample weight tuning; for instance, the peak-validated RMSE for XGBOOST decreased from 11.61ppb to 8.80ppb (−24%), though it caused 2%−3% increase in RMSE for other validations. External validation exhibited better results than leave-cluster-out validation, since leave-cluster-out validation examined prediction accuracy for locations farther than external validation (Figure S6b). BEML achieved significantly better performance than the ensemble method (AVERAGING) that simply averages the predictions from the selected base learners (Figure S3). Among the thirteen base learners, RF achieved the highest global

accuracy and XGBOOST achieved the highest peak accuracy - BEML achieved global accuracy closer to RF and peak accuracy closer to XGBOOST.

The performances of CMAQ and of BEML-DS before and after sample weight tuning are presented in Figure 2. In estimating DM8HA for the 708 monitors, CMAQ exhibited reasonable prediction performance (RMSE=9.80 ppb, R^2=0.52) with points scattered around the 1:1 line. The performance for the annual 4th highest DM8HA was lower (RMSE=7.86 ppb, R^2=0.42). For the ensemble method, global accuracy was improved for BEML-DS without sample weight tuning (RMSE=5.78 ppb, R^2=0.83). However, this model cannot ensure peak accuracy, and the annual 4th highest DM8HA was underestimated for almost all locations. The performance for ozone peak estimation (RMSE=7.37 ppb, R^2=0.49) was only slightly better than CMAQ. In contrast, BEML-DS with sample weight tuning achieved significant improvements in global accuracy for DM8HA (RMSE=5.88 ppb, R^2=0.83) and peak accuracy for the annual 4th highest DM8HA (RMSE=4.64 ppb, R^2=0.80). Here, BEML-DS, unless otherwise specified, denotes the model trained based on complete three-stage hyperparameter tuning. For external validation, BEML-DS performed better than CMAQ across different climate regions; for instance, the RMSE in each climate region decreased by ~20% (Figure S7).

Spatial distributions of the predicted annual 4th highest DM8HA ozone concentrations from four methods (BEML-DS, BSTH-DS, CMAQ, MON-INT) are depicted in Figure 3 (see Figure S8 for analogous results from the thirteen base learners). MON-INT, a simple interpolation method often used in epidemiological studies, assigns closest monitor measurements to neighboring counties. MON-INT generated a spatial surface with similar or identical values in many neighboring areas, leading to unreliable estimates, especially for those counties and census tracts without monitors (Figure S1). CMAQ modeling considers the chemistry and transport of atmospheric pollutants, predicting concentrations independently of monitor observations. While CMAQ captured spatial patterns consistent with MON-INT, it exhibits bias (Figure S2) and cannot provide subgrid scale information important for exposure and environmental justice assessments. BSTH-DS fuses CMAQ with monitor measurements to ensure perfect fit with available “ground truth” and provides estimates for each census tract. However, since spatial patterns are adjusted solely using monitor observations and extrapolated via “non-informative” Gaussian random fields, the reliability of predicted subgrid gradients cannot be guaranteed.[10] BEML-DS fuses CMAQ outputs with monitor measurements and a wide spectrum of heterogeneous spatiotemporal information, combining different algorithms to capture underlying patterns.

Both BEML-DS and BSTH-DS were found to produce consistent predictions of ozone DM8HA across all 72,283 CONUS census tracts for 2011 (Figure S9a). However, significant differences are observed in particular ranges, and these differences exhibit a strong relationship with census tract distance from the nearest monitor (Figure S9b). Both approaches had similar predictions for census tracts close to monitors, e.g., discrepancy R^2=0.95 for distances <5 km, but diverged for greater distances, with R^2 0.85–0.90 for distances between 5 and 270 km, and R^2 0.65–0.70 for distances exceeding 270 km.

The extrapolation performance of the three approaches (CMAQ, BSTH-DS, BEML-DS) is summarized in Figure S10. Compared with BSTH-DS in predicting DM8HA (Table S5), BEML-DS had higher global accuracy (RMSE=8.55 ppb, R^2=0.66), larger association (r=0.82), and smaller bias (MB = 1.23 ppb). BSTH-DS and BEML-DS had peak accuracy lower than CMAQ and the peak RMSE of BSTH-DS was slightly smaller than that of BEML-DS. In predicting ozone warm-season mean, BEML-DS captured the highest variations of observed data (varZ=27.44 ppb^2) and retained the smallest error variations (varE=10.76 ppb^2): it achieved much better performance than BSTH-DS, with 41.54% reduction in RMSE and 54.04% reduction in MB.

Figure 4 depicts the time series of estimated DM8HA at three locations that have been historically of interest: New York, NY (urban area), Los Angeles, CA (urban area), and Houston, TX (urban and suburban areas). Concentration series for multiple census tracts were plotted to show subgrid variations predicted by BEML-DS. The highlighted area in New York (Figure 5a) consists of 145 census tracts and correspondingly 145 independent concentration series. This area is part of Manhattan and located within the 12×12 km^2 grid cell with the monitor site ID 360610135. BEML-DS reduced bias by calibrating CMAQ estimates to ground truth, and simulated subgrid gradients by considering effects of heterogeneous spatiotemporal information: these gradients can be interpreted using Shapley values (Figure 5, Figure S12–S17); please refer to Text S5 for further discussion.

3.2. Model Transferability to Data Limited Situations

Table 1 lists the transferred RMSE and R^2 values for different models by year (see Tables S6–S7 for five other evaluation metrics). With the exception of 2014, the 2011-trained BEML model can predict well for the next six “future” years, with R^2 ranging from 0.66 to 0.75. It should be pointed that here we report a “pseudo-R^2”, as explained in the Note of Table 1, instead of the ordinary least squares R^2 (OLS-R^2) that is typically used for linear regression. The pseudo-R^2 is more sensitive to bias, and thus is less “optimistic” than OLS-R^2 for (nonlinear) model evaluation. The CMAQ simulation for 2014 had significantly larger bias, lower accuracy, and smaller correlation; this resulted in worse BEML performance for that year. The 2011-trained BEML model achieved 49% decrease of MB, 33% decrease of RMSE, and 22% increase of r compared to CMAQ in 2014, indicating BEML’s robustness. Among the fourteen ML models, BEML had the best global accuracy for 2012 and 2015–2017, and accuracy close to the best base learner (RF) for 2013–2014. In addition, BEML achieved performance close to the best base learners for bias (RF) and peak accuracy (XGBOOST).

Figure S18 shows the spatial distributions of predicted mean of ozone DM8HA during warm season (May to September) for six years. The 2011-trained BEML model captures spatial patterns consistent with ground observations (MON-INT) and the monitor-based calibration model (BSTH-DS) in each year except 2014. The 2014 CMAQ simulation overestimated substantially over large areas such as the Southwest region, causing insufficient calibration for the 2011-trained BEML model compared to MON-INT and BSTH-DS that use year-specific corresponding observations. Higher uncertainties in CMAQ inputs for 2014 resulted

in outcomes that exhibited larger positive bias than other years; however CMAQ still captured important spatiotemporal patterns (r = 0.67, OLS-R^2=0.45).

To assess potential effects of local spatiotemporal covariates on prediction performance, we compared predictions at monitoring sites during 2012–2017 using the 2011-trained model with different inputs: (1) same year CMAQ estimates and same year fine-scale covariates; (2) same year CMAQ estimates and 2011 fine-scale covariates. Performances of the RF models (evaluated with seven metrics) in 2012–2017 for the two types of inputs are summarized in Table S8. Compared to predictions using reference year covariates with which the model was trained, use of actual year fine-scale covariates achieved significant reduction of RMSE, peak error, error variance, and significant increase of coefficient of determination, correlation. This shows that the 2011-trained ML model is not simply debiasing CMAQ outcomes, but that the corresponding local spatiotemporal factors play significant roles in improving model accuracy and transferability for “future” years.

Figure 6 shows the simulated spatial distributions of the warm-season (May to September) mean ozone DM8HA across the CONUS for a simulation corresponding to a hypothetical future year (2051) scenario adopted in earlier studies.[17,18] BEML-DS captures complex nonlinear patterns that differ by climate region (Figures S19–S24); please refer to Text S6 for further discussion.

4. Discussion

We developed a robust and flexible/transferable Bayesian Ensemble Machine Learning framework for downscaling CMAQ estimates of ozone DM8HA to census tracts across the CONUS and considered both past (2011–2017) and future (2051) years. The transferability of the model was demonstrated by training it with CMAQ outputs and local data for 2011 and then applying it to predict concentrations for the other years. New concepts and tools were applied to support the design, evaluation, and interpretation of Machine Learning for fine-scale air quality modeling.

4.1. Multi-objective Ensemble Method

Previous ensemble models[13–15,29–31] typically involved a small number of Machine Learning algorithms and estimated ensemble weights via single objective optimization to improve performance for a particular validation, not necessarily outperforming learners for different validation criteria. The performance of an ensemble model is related to both the accuracy and diversity of base learners; furthermore, a robust ensemble model must be balanced and evaluated using multiple criteria and objectives.[23,32] The present study considered a rich set of algorithms and constructed statistical indices to measure robustness, accuracy, and diversity of base learners, and fused this knowledge via Bayesian inference to update ensemble weights by location. Three-stage hyperparameter tuning and targeted validations (Figure 1) were introduced to improve the model’s ability to interpolate, extrapolate, and capture peak concentrations.

4.2. Interpolation vs Extrapolation

Previous ML/ensemble models[33–36] have achieved cross-validated (either leave-sample-out or leave-monitor-out) R^2 ranging from 0.64 to 0.78. A recent paper,[31] considering 169 covariates, reported cross-validated R^2=0.9 in the predicting of 1×1 km^2 ozone concentrations across the CONUS. Some state-of-the-art geostatistical models,[4,7,8] that downscale CMAQ estimates solely with ground observations, have achieved similar results with cross-validated R^2=0.7–0.9. However, most reported validation results are "interpolation-oriented", thus not reflecting the model's extrapolation ability.

Leave-monitor-out validation has advantages over leave-sample-out, as it can provide less optimistic results;[35] however, it reflects prediction accuracy for locations within short distances from monitoring sites (up to 36 km). We observed that ~17% of census tracts, or equivalently, ~40% of 1×1 km^2 grid cells (a resolution used in previous ML studies) in the CONUS exceeded the extrapolation distance range for leave-monitor-out validation (Figure S6). A traditional geostatistical downscaler model (BSTH-DS) can achieve prediction performance similar to the ensemble model (BEML-DS) for census tracts close to monitors (interpolation), but significant differences occur for census tracts far from monitors (Figure S9). We used targeted leave-cluster-out validation to assess prediction accuracy for areas far away from monitors (36–330 km), showing that BEML-DS can improve spatial extrapolation accuracy compared to BSTH-DS (Figure S10, Table S5).

Unfortunately, comprehensive spatial extrapolation evaluations for ML/ensemble models have been rare (e.g., Huang, et al.[37]). Compared to representative ensemble and geostatistical models, the present study obtained excellent interpolation performance (10-fold leave-sample-out R^2=0.85), and prominent extrapolation performance (10-fold leave-cluster-out R^2=0.69).

4.3. Peak Accuracy

In order to refine earlier ML/ensemble approaches[33–36] that focused on annual and seasonal averages of ozone concentrations, the present study introduced sample weight tuning to balance global and peak accuracy, thus developing the first Machine Learning model that can be comparable to state-of-the-art geostatistical methods (BSTH-DS) in predicting simultaneously annual peaks and seasonal means of ozone levels (Figure 3, Figure S11).

4.4. Interpretability

Typical Machine Learning modeling is considered a "black-box" approach;[38] however, interpretable ML provides a framework that can help explain the reasoning of the "learning system" to improve prediction credibility.[22,39–41] Previous studies have produced predictions with very fine resolutions, ranging from 1 km to 100 m for certain areas, but the quality of those fine-scale predictions is unknown. In the estimation and evaluation of subgrid gradients, the following five issues matter: (1) availability of local spatiotemporal information (e.g., microscale measures[42]) to support the finer scale considered; (2) targeted validations to reflect not only interpolation but also extrapolation capability for substantial distances; (3) specialized interpretation tools to measure the contributions of local variations; (4) locally dense monitoring networks (that benefit from advanced measurement and sensing

techniques[43,44]) for direct evaluation; (5) applicability to exposure and health studies and environmental justice assessments.[45–48]

The present study used the Shapley value, for the first time in this type of application, to link the "logic" of subgrid gradients to available local spatiotemporal covariates. The Shapley value measures the contribution of inputs (including CMAQ estimates and multiple fine-scale covariates) to each concentration prediction. CMAQ estimates, as expected, had the highest contribution to predictions; however, contributions of certain fine-scale covariates are also significant (Figure 6d, Figure S15). The contribution differences for each of those covariates across neighboring census tracts are in essence the "forces" acting to generate local concentration gradients (Figure S16).

Fusing the spatiotemporal information with explainable effects, the local concentrations predicted by BEML-DS exhibited higher credibility (especially for areas far from monitors, Figures S9–S10) than the estimates derived through BSTH-DS, which conducts "extrapolation" using the assumption of Gaussian random fields without spatiotemporal covariates. The data sets generated from the different approaches employed in the present article are available at: https://ccl-eohsi.shinyapps.io/beml_ozone_dashboard/, for further comparisons and for use in health and environmental justice studies.

4.5. Transferability and CMAQ Downscaling for Future Years

Previous studies[15,34,49,50] evaluated prediction performance with back-extrapolation validation, estimating historical concentrations, for periods before current monitoring networks were operational. The present study focused on "forward-extrapolation" performance for years after the training period: we obtained robust results (with consistent improvements that were verified by seven evaluation statistics) for predicting six consecutive years 2012–2017 using the "historically" trained (2011) BEML-DS model.

In essence, Machine Learning aims to find a "mapping" that can capture complex patterns[22] between monitor observations, CMAQ estimates and multiple fine-scale spatiotemporal covariates. A carefully tuned ML model with targeted validations can ensure that these patterns are applicable for predictions across the modeling domain, where data scales for prediction are consistent with those for training. On one hand, grid-based CMAQ estimates were "calibrated" to "adjust" towards "ground truth"; on the other hand, these calibrated concentrations (reflecting coarse scale variation) were further "adjusted" locally based on fine-scale spatiotemporal covariates that are considered in the training model (local scale variation). According to counterfactual analysis performed (Table S8), ML predictions for 2012–2017 using 2011 local spatiotemporal covariates had accuracy close to CMAQ estimates, while predictions using same year local spatiotemporal covariates can achieve further improvements. This finding indicates that local spatiotemporal covariates (main sources of subgrid concentration gradients) are important for improving local accuracy and transferability for future years.

The BEML model was also applied to downscale CMAQ outcomes for a simulation using meteorology for a future year (2051), while employing a hypothetical scenario of "fixed" anthropogenic emissions, aiming to isolate the effect of climate on air quality.[51] It should be

clarified that the objective of this analysis is not to produce accurate predictions of ozone concentrations for the future year considered, as this is precluded by the uncertainties in the "trajectories" of actual future emissions. Instead, it aims to show that Machine Learning can capture complex nonlinear patterns among different variables for specific alternative scenarios; those patterns are robust and can be utilized to downscale scenario-based CMAQ simulations in data-limited situations. The Shapley value provides a metric that helps to quantitatively explain why the "historically" trained ML model makes a certain estimation for a local area in specific future scenarios (Figures S19–S24). This "proof of concept" demonstration provides a pathway for developing modeling tools required for pursuing studies related to environmental and climate justice, thus responding to a critical need for already disadvantaged communities.

4.6. Limitations

Naturally, the tools presented in this study must be considered and understood in the context of their intended applications. Specifically, the framework and methods introduced here have been designed to be robust, flexible and transferable to "data-limited" situations, where re-training of the ML model with high resolution data may not be an option. This differs from the approach and objectives of other related efforts: for example, a recent study[31] obtained excellent cross-validation results (reflecting interpolation and extrapolation with short distances) for ozone concentrations across the CONUS using a broad set of spatiotemporal covariates (169 variables including remote sensing data) to improve prediction accuracy. In the present study we focused (a) on improving robustness by balancing interpolation, extrapolation, and peak accuracy of the downscaling model, combined with basic available spatiotemporal information, and (b) on evaluating transferability of the BEML-DS approach to data-limited situations, by applying our framework without retraining, to multiple past years and evaluating it with available monitor observations. We have further assessed model transferability by demonstrating the feasibility of employing BEML-DS in downscaling scenario-based CMAQ simulations for a future year, a step that addresses a need in climate change and environmental justice studies. Of course, application of the "pre-trained" BEML-DS framework in a data-limited setting, would only be reasonable for conditions that do not alter drastically the dynamics of the air pollution system studied. For situations where conditions (emission patterns and levels, meteorology, etc.) may have changed from their historically "normal ranges" to an extent that would cause the air pollution dynamics simulated by CMAQ to exhibit a drastically different behavior (e.g. to move from a NO_x-limited to a VOC-limited regime of ozone formation in a particular area) one cannot expect applicability of the pre-trained BEML-DS. Nevertheless, the good transfer validation results (Table 1) appear to support the robustness of our model with respect to reasonable levels of structural and parametric model and data uncertainties and fluctuations.

Due to lack of locally dense monitoring networks (a problem for most studies), the generated fine-scale concentrations/variations in this national study cannot be directly compared and tested. However, relevant demonstrations and analyses have been performed, with efforts in the following four aspects: (a) the excellent validation results obtained in this study indicate stable and effective mappings between monitor measurements and fine-scale variables learned via different algorithms; (b) the Shapley value provides

reasonable quantification and explanation regarding driving forces (i.e., potential fine-scale attributes) of local variations; (c) inclusion of fine-scale variables can improve extrapolation accuracy, compared to BSTH-DS; (d) contributions of fine-scale variables are important to predictions, and as expected, inclusion of same year information improves model accuracy and transferability. Shapley values provide promising metrics for quantitative interpretation of captured local variations and future patterns, but require a substantial computational effort, especially for large samples: It takes ~2 min (Intel Xeon 6130) to calculate the Shapley values for all 37 features for each estimate; the calculation for ~200,000 estimates (Figure 6) for the national study requires high performance computing with tens or hundreds of CPUs. However, the computational load is not an issue for smaller regional/community studies or for national/global studies using subsets of samples (Text S5.1). Finally, it should be noted that the current BEML-DS model employs a deterministic approach (i.e., an ensemble of discriminative algorithms) and therefore does not provide confidence intervals for each estimate. Expanding the current framework with probabilistic generative ML models such as deep probabilistic graphical models,[52] Bayesian networks,[53] or vine copulas[54] is currently being evaluated for future enhancements of the information derived from outputs produced by spatiotemporal ambient air pollution models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This study was supported by the Ozone Research Center (funded by the State of New Jersey Department of Environmental Protection under Grant AQ05-011), by the Center for Environmental Exposures and Disease at EOHSI (NIEHS Grant P30ES005022) and by the NJ Alliance for Clinical Translational Science (NIH Grant UL1TROO3017). The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency or any of the funding agencies. We sincerely acknowledge the help of Adam Reff for running the validation experiment of the EPA-BSTH downscaler.

References

1. USEPA, Integrated science assessment of ozone and related photochemical oxidants (Final Report, Feb 2013) In U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-10/076F: 2013.
2. USEPA, Integrated science assessment for ozone and related photochemical oxidants (Final) In U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-20/012: 2020.
3. USEPA, CMAQ publications and peer review https://www.epa.gov/cmaq/cmaq-publications-and-peer-review (accessed 2019-10-30).
4. USEPA, Bayesian space-time downscaling fusion model (downscaler) -derived estimates of air quality for 2011 In U.S. Environmental Protection Agency, Washington, DC, EPA-454/S-15–001: 2015.
5. USEPA, CMAQ-v5.3 user manual In U.S. Environmental Protection Agency: Community Modeling and Analysis System: 2019.
6. Fuentes M; Raftery AE, Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. Biometrics 2005, 61, 36–45. [PubMed: 15737076]
7. Xu Y; Serre ML; Reyes J; Vizuete W, Bayesian maximum entropy integration of ozone observations and model predictions: A national application. Environ. Sci. Technol 2016, 50 (8), 4393–4400. [PubMed: 26998937]

8. Friberg MD; Zhai X; Holmes HA; Chang HH; Strickland MJ; Sarnat SE; Tolbert PE; Russell AG; Mulholland JA, Method for fusing observational data and chemical transport model simulations to estimate spatiotemporally resolved ambient air pollution. Environ. Sci. Technol 2016, 50 (7), 3695–3705. [PubMed: 26923334]
9. Berrocal VJ; Gelfand AE; Holland DM, A spatio-temporal downscaler for output from numerical models. J. Agr. Biol. Envir. St 2010, 15 (2), 176–197.
10. USEPA, Downscaler model for predicting daily air pollution https://www.epa.gov/air-research/downscaler-model-predicting-daily-air-pollution (accessed 2018-12-03).
11. Feng L; Li Y; Wang Y; Du Q, Estimating hourly and continuous ground-level $PM_{2.5}$ concentrations using an ensemble learning algorithm: The ST-stacking model. Atmos. Environ 2019, 117242.
12. Li L; Lurmann F; Habre R; Urman R; Rappaport E; Ritz B; Chen JC; Gilliland FD; Wu J, Constrained mixed-effect models with ensemble learning for prediction of nitrogen oxides concentrations at high spatiotemporal resolution. Environ. Sci. Technol 2017, 51 (17), 9920–9929. [PubMed: 28727456]
13. Lyu B; Hu Y; Zhang W; Du Y; Luo B; Sun X; Sun Z; Deng Z; Wang X; Liu J; Wang X; Russell AG, Fusion method combining ground-level observations with chemical transport model predictions using an ensemble deep learning framework: application in China to estimate spatiotemporally-resolved $PM_{2.5}$ exposure fields in 2014–2017. Environ. Sci. Technol 2019, 53 (13), 7306–7315. [PubMed: 31244060]
14. Shtein A; Kloog I; Schwartz J; Silibello C; Michelozzi P; Gariazzo C; Viegi G; Forastiere F; Karnieli A; Just AC; Stafoggia M, Estimating daily $PM_{2.5}$ and PM_{10} over Italy using an ensemble model. Environ. Sci. Technol 2019, 54 (1), 120–128. [PubMed: 31749355]
15. Xiao Q; Chang HH; Geng G; Liu Y, An ensemble machine-learning model to predict historical $PM_{2.5}$ concentrations in China from satellite data. Environ. Sci. Technol 2018, 52 (22), 13260–13269. [PubMed: 30354085]
16. Zhai B; Chen J, Development of a stacked ensemble model for forecasting and analyzing daily average $PM_{2.5}$ concentrations in Beijing, China. Sci. Total Environ 2018, 635, 644–658. [PubMed: 29679837]
17. Fann NL; Nolte CG; Sarofim MC; Martinich J; Nassikas NJ, Associations between simulated future changes in climate, air quality, and human health. JAMA Netw. Open 2021, 4 (1), e2032064. [PubMed: 33394002]
18. Nolte CG; Spero TL; Bowden JH; Sarofim MC; Martinich J; Mallard MS, Regional temperature-ozone relationships across the U.S. under multiple climate and emissions scenarios. J. Air Waste Manag. Assoc 2021, 71 (10), 1251–1264. [PubMed: 34406104]
19. Cai T; Zhang Y; Ren X; Bielory L; Mi Z; Nolte CG; Gao Y; Leung LR; Georgopoulos PG, Development of a semi-mechanistic allergenic pollen emission model. Sci. Total Environ 2019, 653, 947–957. [PubMed: 30759620]
20. USEPA, Air data: air quality data collected at outdoor monitors across the US https://www.epa.gov/outdoor-air-quality-data (accessed 2019-01-27).
21. USEPA, RSIG-related downloadable data files https://www.epa.gov/hesc/rsig-related-downloadable-data-files (accessed 2019-01-27).
22. Ren X; Mi Z; Georgopoulos P, Comparison of machine learning and land use regression for fine scale spatiotemporal estimation of ambient air pollution: modeling ozone concentrations across the contiguous United States. Environ. Int 2020, 142, 105827. [PubMed: 32593834]
23. Zhou Z, Ensemble methods: foundations and algorithms CRC press: 2012; pp 99–118.
24. Ribeiro MT; Singh S; Guestrin C In " Why should I trust you?" Explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016; p 1135–1144.
25. Biecek P; Burzykowski T, Explanatory model analysis: explore, explain, and examine predictive models CRC Press: 2021; pp 107–122.
26. Lundberg SM; Lee SI In A unified approach to interpreting model predictions, Advances in neural information processing systems, 2017; p 4765–4774.
27. Roth AE, The Shapley value: essays in honor of Lloyd S. Shapley Cambridge University Press: 1988; pp 1–30.

28. Lundberg SM; Erion G; Chen H; DeGrave A; Prutkin JM; Nair B; Katz R; Himmelfarb J; Bansal N; Lee SI, From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell 2020, 2 (1), 56–67. [PubMed: 32607472]
29. Di Q; Amini H; Shi L; Kloog I; Silvern R; Kelly J; Sabath MB; Choirat C; Koutrakis P; Lyapustin A, An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution. Environ. Int 2019, 130, 104909. [PubMed: 31272018]
30. Di Q; Amini H; Shi L; Kloog I; Silvern R; Kelly J; Sabath MB; Choirat C; Koutrakis P; Lyapustin A, Assessing NO_2 concentration and model uncertainty with high spatiotemporal resolution across the contiguous United States using ensemble model averaging. Environ. Sci. Technol 2019, 54 (3), 1372–1384.
31. Requia WJ; Di Q; Silvern R; Kelly JT; Koutrakis P; Mickley LJ; Sulprizio MP; Amini H; Shi L; Schwartz J, An ensemble learning approach for estimating high spatiotemporal resolution of ground-level ozone in the contiguous United States. Environ. Sci. Technol 2020, 54 (18), 11037–11047. [PubMed: 32808786]
32. Tang EK; Suganthan PN; Yao X, An analysis of diversity measures. Mach. Learn 2006, 65 (1), 247–271.
33. Di Q; Rowland S; Koutrakis P; Schwartz J, A hybrid model for spatially and temporally resolved ozone exposures in the continental United States. J. Air Waste Manag. Assoc 2017, 67 (1), 39–52. [PubMed: 27332675]
34. Liu R; Ma Z; Liu Y; Shao Y; Zhao W; Bi J, Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: A machine learning approach. Environ. Int 2020, 142, 105823. [PubMed: 32521347]
35. Watson GL; Telesca D; Reid CE; Pfister GG; Jerrett M, Machine learning models accurately predict ozone exposure during wildfire events. Environ. Pollut 2019, 254, 112792. [PubMed: 31421571]
36. Zhan Y; Luo Y; Deng X; Grieneisen ML; Zhang M; Di B, Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. Environ. Pollut 2018, 233, 464–473. [PubMed: 29101889]
37. Huang C; Hu J; Xue T; Xu H; Wang M, High-resolution spatiotemporal modeling for ambient $PM_{2.5}$ exposure assessment in China from 2013 to 2019. Environ. Sci. Technol 2021, 55 (3), 2152–2162. [PubMed: 33448849]
38. Schmidt CW, Into the black box: What can machine learning offer environmental health research? Environ. Health Perspect 2020, 128 (2), 022001.
39. Doshi-Velez F; Kim B, Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 2017, 1–13.
40. Xing J; Zheng S; Ding D; Kelly JT; Wang S; Li S; Qin T; Ma M; Dong Z; Jang C, Deep learning for prediction of the air quality response to emission changes. Environ. Sci. Technol 2020, 54 (14), 8589–8600. [PubMed: 32551547]
41. Yan X; Zang Z; Luo N; Jiang Y; Li Z, New interpretable deep learning model to monitor real-time PM2.5 concentrations from satellite data. Environ. Int 2020, 144, 106060. [PubMed: 32920497]
42. Lu T; Marshall JD; Zhang W; Hystad P; Kim S-Y; Bechle MJ; Demuzere M; Hankey S, National empirical models of air pollution using microscale measures of the urban environment. Environ. Sci. Technol 2021, 55 (22), 15519–15530. [PubMed: 34739226]
43. Hankey S; Marshall JD, Land use regression models of on-road particulate air pollution (particle number, black carbon, $PM_{2.5}$, particle size) using mobile monitoring. Environ. Sci. Technol 2015, 49 (15), 9194–9202. [PubMed: 26134458]
44. Schneider P; Bartonova A; Castell N; Dauge FR; Gerboles M; Hagler GS; Huglin C; Jones RL; Khan S; Lewis AC, Toward a unified terminology of processing levels for low-cost air-quality sensors. Environ. Sci. Technol 2019, 53 (15), 8485–8487. [PubMed: 31353903]
45. Berrocal VJ; Guan Y; Muyskens A; Wang H; Reich BJ; Mulholland JA; Chang HH, A comparison of statistical and machine learning methods for creating national daily maps of ambient $PM_{2.5}$ concentration. Atmos. Environ 2020, 222, 117130.

46. Jin X; Fiore AM; Civerolo K; Bi J; Liu Y; Van Donkelaar A; Martin RV; Al-Hamdan M; Zhang Y; Insaf TZ, Comparison of multiple $PM_{2.5}$ exposure products for estimating health benefits of emission controls over New York State, USA. Environ. Res. Lett 2019, 14 (8), 084023.
47. Kelly JT; Jang C; Timin B; Di Q; Schwartz J; Liu Y; van Donkelaar A; Martin RV; Berrocal V; Bell ML, Examining $PM_{2.5}$ concentrations and exposure using multiple models. Environ. Res 2021, 196, 110432. [PubMed: 33166538]
48. McGuinn LA; Ward-Caviness C; Neas LM; Schneider A; Di Q; Chudnovsky A; Schwartz J; Koutrakis P; Russell AG; Garcia V, Fine particulate matter and cardiovascular disease: Comparison of assessment methods for long-term exposure. Environ. Res 2017, 159, 16–23. [PubMed: 28763730]
49. Knibbs LD; Coorey CP; Bechle MJ; Marshall JD; Hewson MG; Jalaludin B; Morgan GG; Barnett AG, Long-term nitrogen dioxide exposure assessment using back-extrapolation of satellite-based land-use regression models for Australia. Environ. Res. Lett 2018, 163, 16–25.
50. Wu Y; Di B; Luo Y; Grieneisen ML; Zeng W; Zhang S; Deng X; Tang Y; Shi G; Yang F, A robust approach to deriving long-term daily surface NO_2 levels across China: Correction to substantial estimation bias in back-extrapolation. Environ. Int 2021, 154, 106576. [PubMed: 33901976]
51. Pienkosz BD; Saari RK; Monier E; Garcia-Menendez F, Natural variability in projections of climate change impacts on fine particulate matter pollution. Earths Future 2019, 7 (7), 762–770.
52. Koller D; Friedman N, Probabilistic graphical models: principles and techniques MIT press: 2009; pp 943–1001.
53. Aguilera P; Fernández A; Fernández R; Rumí R; Salmerón A, Bayesian networks in environmental modelling. Environ. Model. Softw 2011, 26 (12), 1376–1388.
54. Zhou Y; Ren X; Li S, Probabilistic weighted copula regression model with adaptive sample selection strategy for complex industrial processes. IEEE Trans. Industr. Inform 2020, 16 (11), 6972–6981.

Synopsis:

We developed a transferable Bayesian Ensemble Machine Learning framework for fine-scale spatiotemporal ozone prediction, applicable to “data-limited” situations that include environmental and climate justice issues.

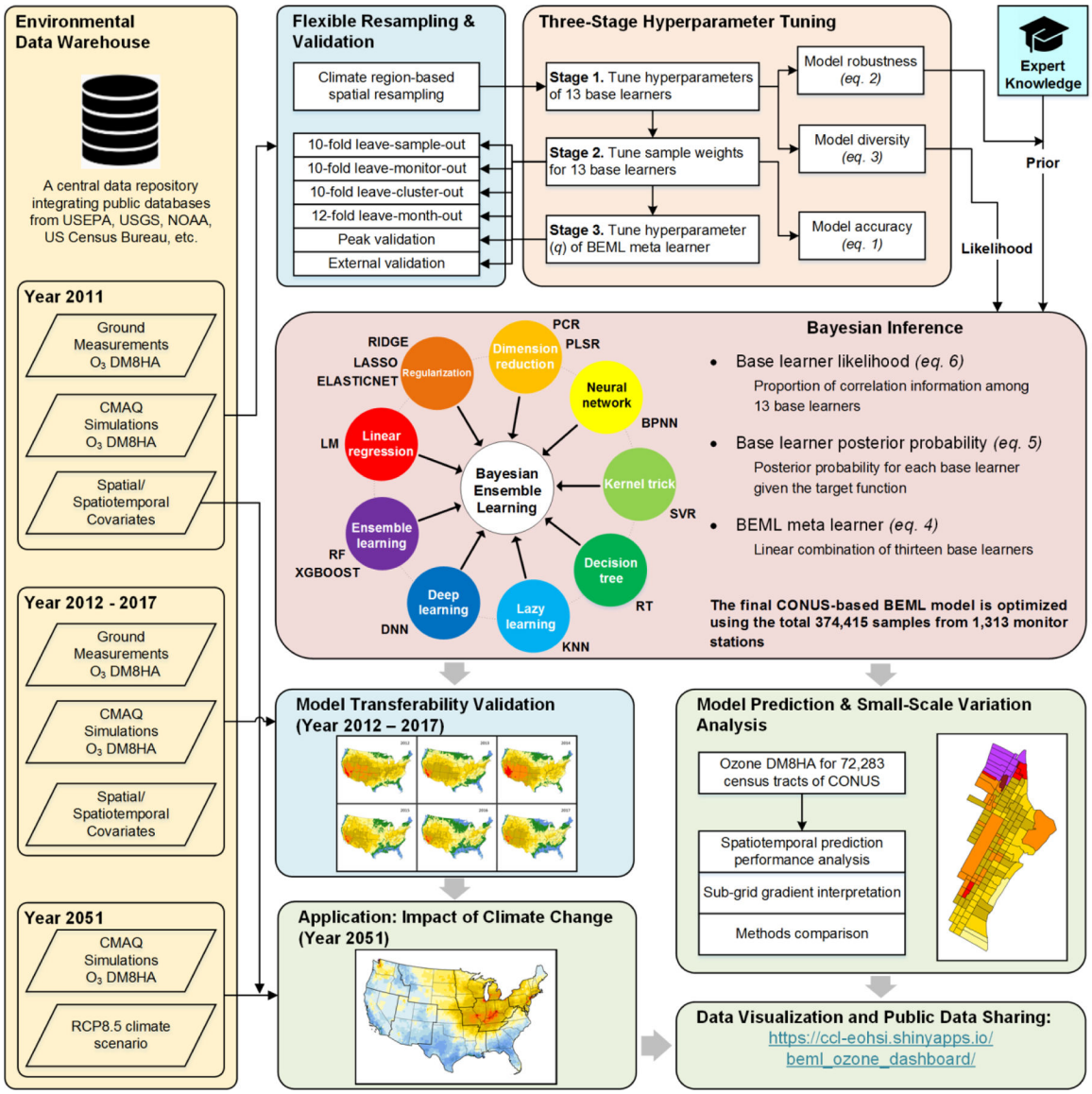


Figure 1.
Flow diagram of the Bayesian Ensemble Machine Learning Downscaler (BEML-DS) framework.

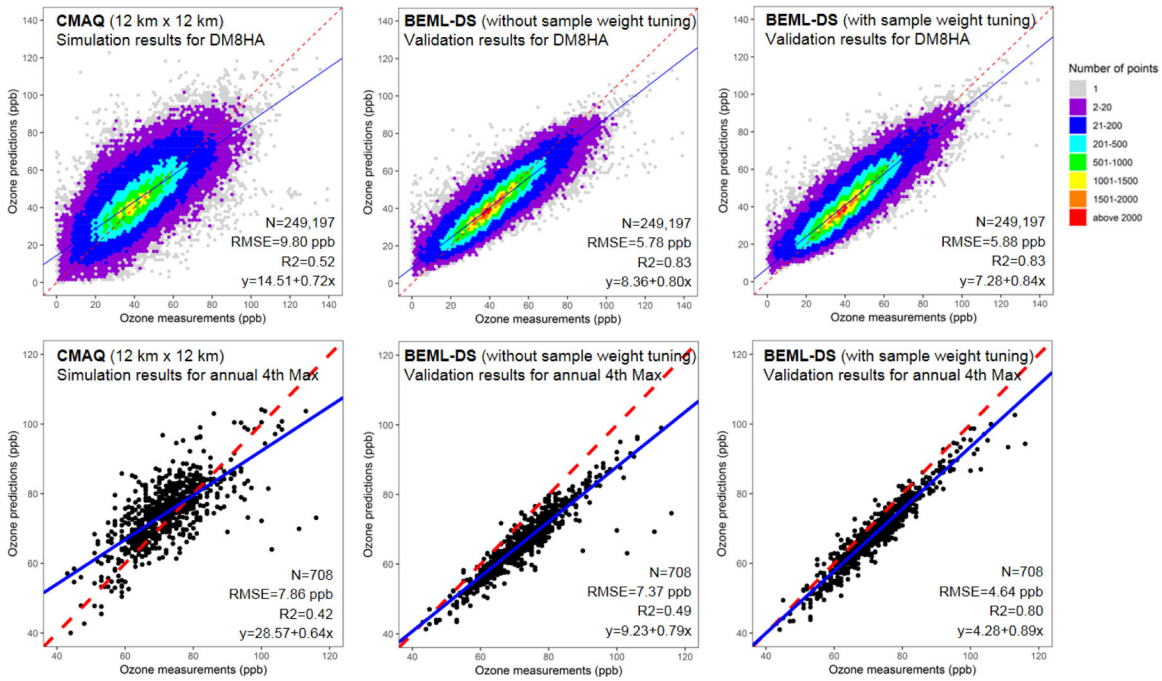


Figure 2.

Simulation/Validation results of Community Multiscale Air Quality (CMAQ) and Bayesian Ensemble Machine Learning Downscaler (BEML-DS). The top row shows Daily Maximum 8-Hour Averages (DM8HA) and the bottom row the annual 4th highest DM8HA. Columns represent CMAQ, BEML-DS without, and BEML-DS with sample weight tuning.

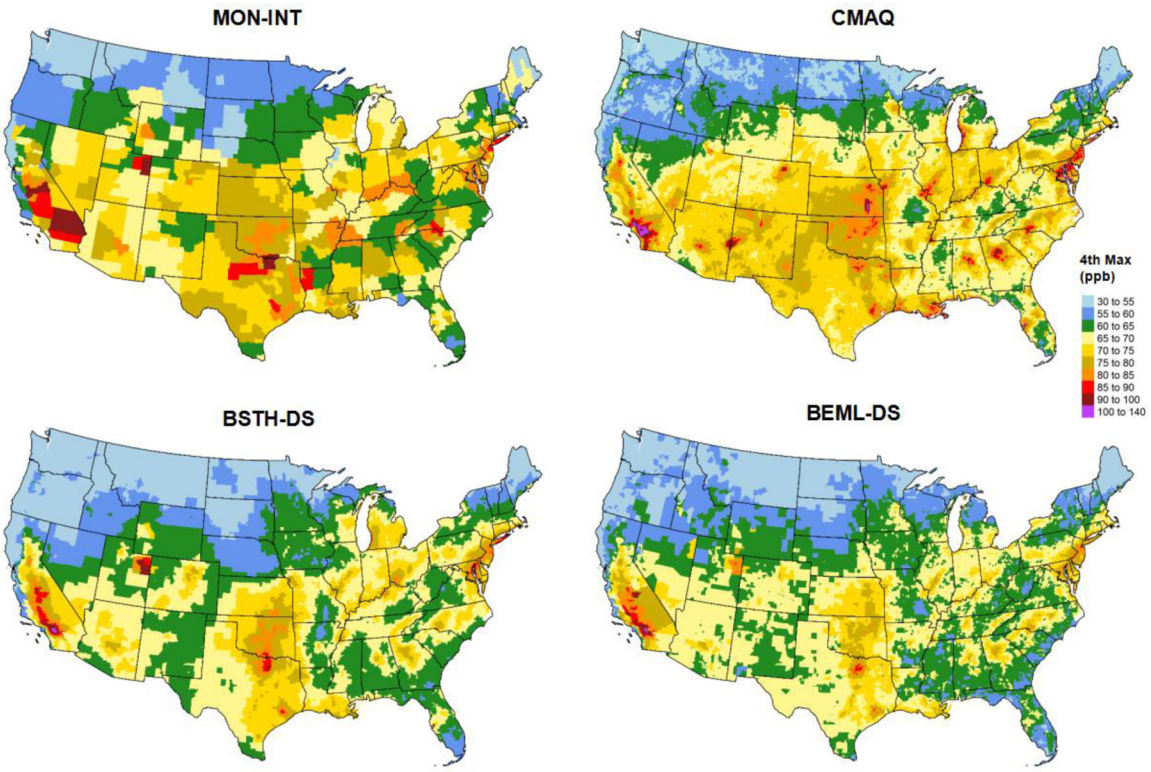


Figure 3.

Spatial distributions of the predicted annual 4th highest DM8HA ozone concentrations from four approaches: Monitor Measurement Interpolation (MON-INT), Community Multiscale Air Quality (CMAQ), Bayesian Spatio-Temporal Hierarchical Downscaler (BSTH-DS), and Bayesian Ensemble Machine Learning Downscaler (BEML-DS).

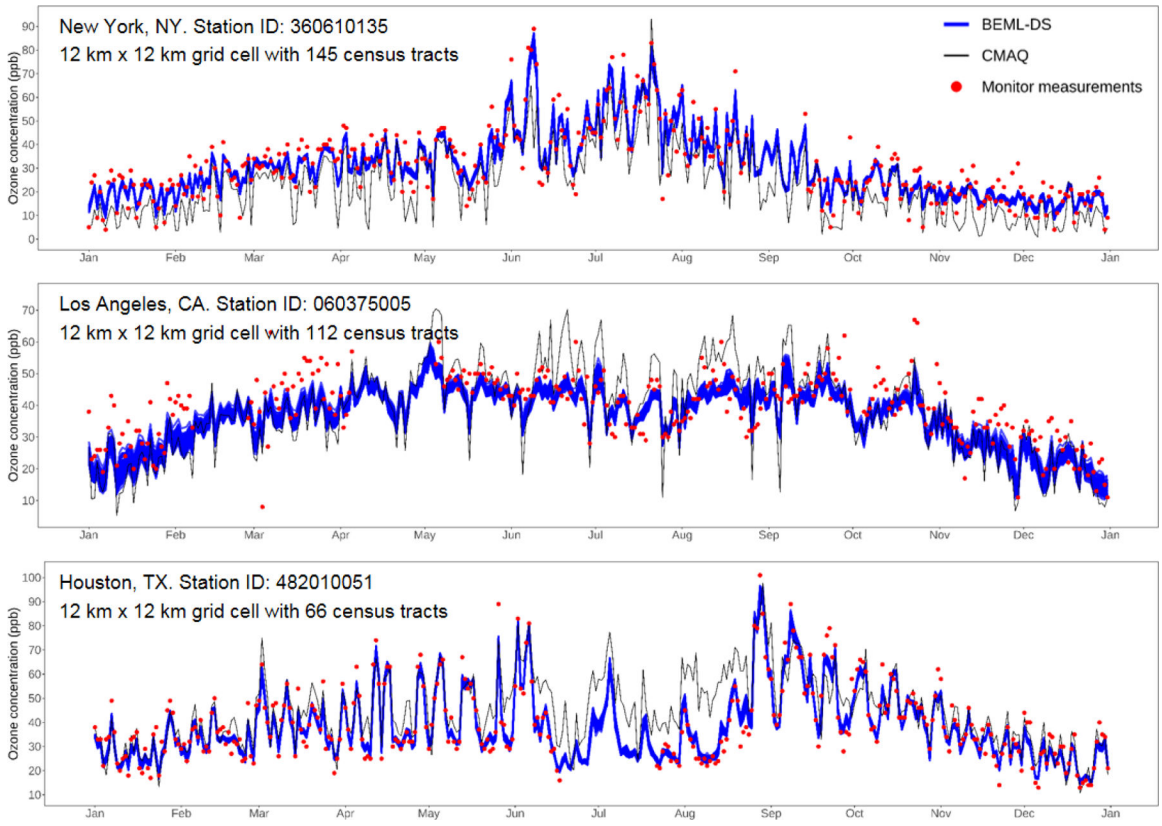


Figure 4.

Concentration series of ozone monitor measurements, 12×12 km^2 grid-based CMAQ model estimates, and the neighboring census tract-based BEML-DS estimates in New York, NY (top), Los Angeles, CA (middle), and Houston, TX (bottom).

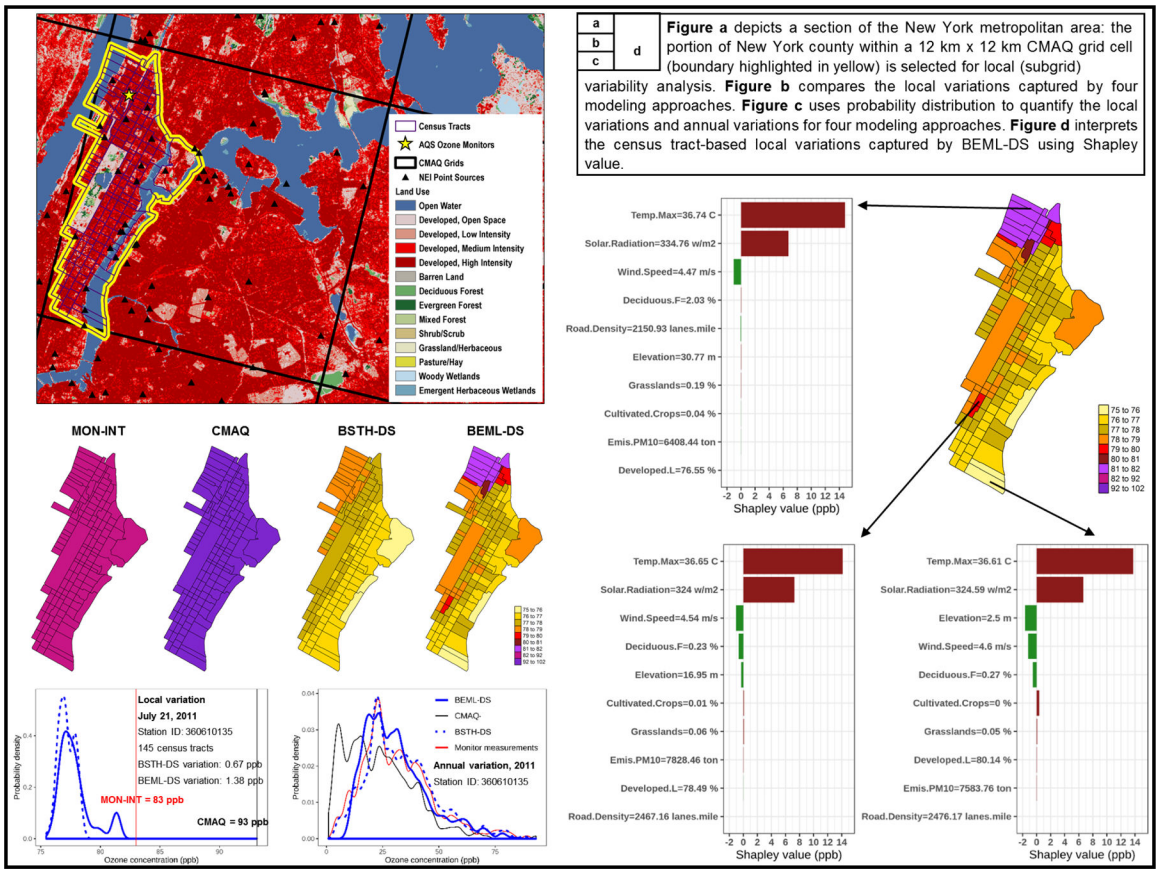


Figure 5.
Subgrid gradients in New York, NY as predicted from four approaches.

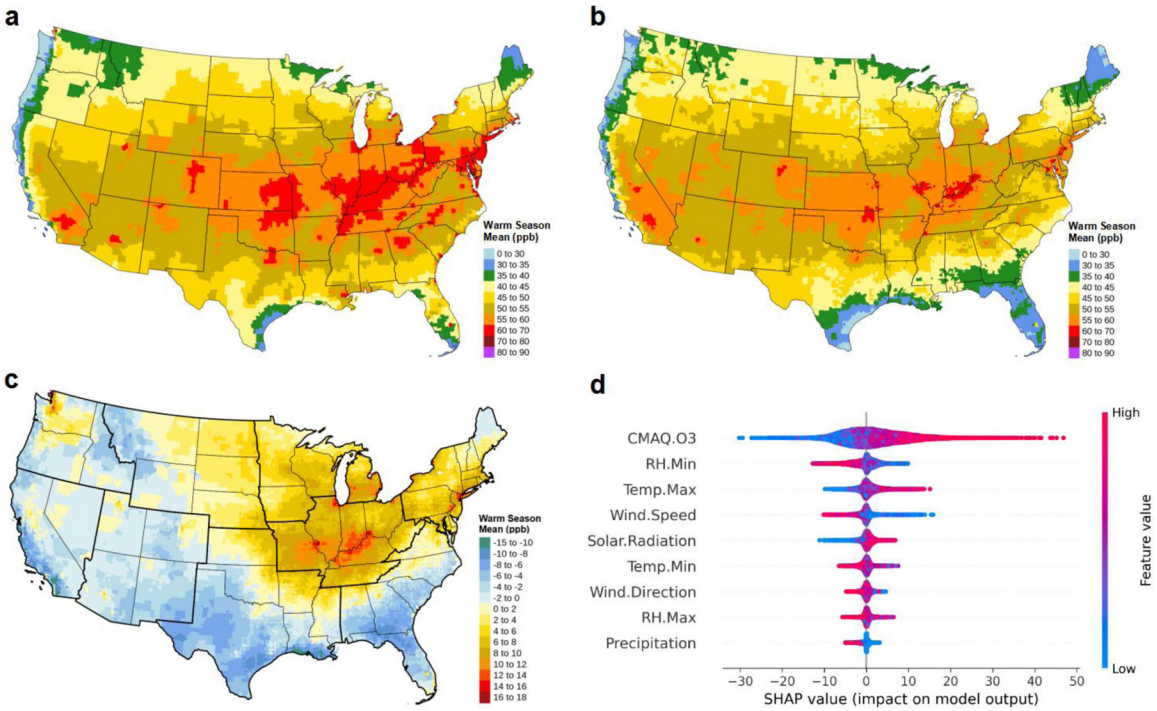


Figure 6.

Ozone estimation and interpretation of variations between future and historical (reference) years: Spatial distributions of DM8HA ozone concentrations during warm season (May to September) 2051, simulated by (a) CMAQ and (b) BEML-DS; (c) Differences of the BEML-downscaled ozone warm-season mean DM8HA between 2051 and 2011; (d) SHAP (Shapley Additive exPlanation) summary plot for 2051 ozone predictions (199,512 samples) at census tracts with monitor stations across the CONUS, compared to the corresponding 2011 average predictions at each monitor station during warm season.

Table 1.

Model transfer performance for different ozone spatiotemporal models by year.

	2012		2013		2014		2015		2016		2017	
	N=381,020		*N*=381,635		*N*=381,867		*N*=380,026		*N*=383,635		*N*=391,889	
Model	**RMSE (ppb)**	**R^2**	**RMSE (ppb)**	**R^2**	**RMSE (ppb)**	**R^2**	**RMSE (ppb)**	**R^2**	**RMSE (ppb)**	**R^2**	**RMSE (ppb)**	**R^2**
CMAQ	9.10	0.60	9.70	0.42	12.19	0.03*	8.81	0.51	8.14	0.56	9.16	0.45
LM	8.41	0.66	8.38	0.56	9.54	0.41	7.75	0.62	7.49	0.63	8.29	0.55
RIDGE	8.37	0.66	8.36	0.57	9.56	0.40	7.69	0.63	7.42	0.63	8.19	0.56
LASSO	8.34	0.66	8.43	0.56	9.92	0.36	7.66	0.63	7.32	0.64	8.14	0.57
ELASTICNET	8.34	0.66	8.43	0.56	9.92	0.36	7.66	0.63	7.32	0.64	8.14	0.57
PCR	8.34	0.66	7.88	0.62	8.68	0.51	7.37	0.66	7.28	0.65	8.09	0.57
PLSR	8.34	0.66	7.86	0.62	8.66	0.51	7.36	0.66	7.26	0.65	8.06	0.57
KNN	7.85	0.70	7.62	0.64	8.00	0.58	7.45	0.65	7.23	0.65	7.58	0.62
SVR	8.39	0.66	7.80	0.62	8.78	0.50	7.27	0.67	7.17	0.66	8.00	0.58
BPNN	8.03	0.69	8.22	0.58	9.22	0.45	7.64	0.64	7.33	0.64	8.05	0.57
DNN	7.83	0.70	7.66	0.64	8.46	0.53	7.36	0.66	7.03	0.67	7.55	0.63
RT	8.46	0.65	8.47	0.56	9.72	0.38	7.96	0.60	7.62	0.61	8.38	0.54
RF	7.19	0.75	**7.34**	**0.67**	**7.96**	**0.59**	6.86	0.71	6.66	0.70	7.25	0.66
XGBOOST	7.33	0.74	7.74	0.63	8.52	0.53	7.03	0.69	6.82	0.69	7.33	0.65
BEML	**7.14**	**0.75**	7.37	0.67	8.14	0.57	**6.79**	**0.71**	**6.62**	**0.71**	**7.23**	**0.66**

*Note: Here we use a "pseudo-R^2" (that can be either negative or positive), where the term "Pred" (defined in Text S4.2) denotes model outcomes: this is different from the ordinary least squares R^2 (OLS-R^2, shares the same formula with pseudo-R^2), where the term "Pred" corresponds to fitted values from a simple linear regression over model outcomes against monitor observations. Due to higher uncertainties of emissions and other input sources for 2014, the CMAQ outcomes for that year exhibit significantly larger positive bias (Figure S18) than other years, thus resulting in very low pseudo-R^2 (0.03). However, the 2014 CMAQ simulation did capture important spatiotemporal patterns for ozone, with Pearson correlation 0.67 (Table S7) and OLS-R^2 0.45. The OLS-R^2 can be easily calculated as the square of the Pearson correlation r in Table S7.