Original Research Article

A Machine Learning Framework for Spatially Honest Benchmarking and Environmental Justice Auditing of Hyperlocal PM2.5 Predictions

Nathan Tan and Saketh Chebrolu

St. Mark's School of Texas, 10600 Preston Rd, Dallas, TX 75230, United States

ABSTRACT

Machine learning is increasingly used to map fine particulate matter (PM2.5) between sparse monitors, but models are commonly validated with random splits that place records from the same sensor in both partitions, inflating apparent accuracy. This study assembles a year of quality-assured daily observations from 226 PurpleAir sensors across four Texas metropolitan areas (54,325 sensor-day records) and evaluates four regression algorithms and a tree ensemble under three protocols: a random 80/20 split; leave-one-sensor-out cross-validation (LOSO-CV) with the withheld sensor's own recent PM2.5 history available (transfer); and LOSO-CV without any site history (cold start), the setting that corresponds to an unmonitored location. When site history is available, random and LOSO scores nearly coincide. Removing site history exposes the true gaps: the ensemble falls from $R^2 = 0.825$ (transfer) to 0.662 (cold start; RMSE = 2.37 μg/m³), the linear baseline collapses from 0.469 to 0.037, and random splitting overstates cold-start ensemble skill by 0.176 R^2. Environmental-justice (EJ) indicators are excluded from the deployed cold-start model so the equity audit is independent of the audited variables. Deployed across 3,758 census tracts with daily 2025 meteorology and averaged to true annual means, predicted annual PM2.5 correlates modestly with the EJScreen EJ index ($r = 0.129$, $p < 10^{-14}$), far below the $r = 0.875$ obtained when EJ variables were model inputs; tract exceedance of the 9 μg/m³ standard rises from 0.2% to 6.0% across EJ quartiles. Spatially honest validation and audit-independent models are prerequisites for credible exposure and equity mapping.

Keywords: air quality; PM2.5; machine learning; spatial cross-validation; environmental justice; low-cost sensors

INTRODUCTION

Fine particulate matter with an aerodynamic diameter at or below 2.5 μm penetrates deep into the respiratory system and enters the bloodstream, and long-term exposure is causally associated with cardiovascular disease, respiratory illness, lung cancer, and premature death (1–3). The global burden is large: ambient and household air pollution together are linked to roughly seven million premature deaths each year, and tens of millions of people in the United States live in areas that fail national air-quality standards (4, 5). In 2024 the U.S. Environmental Protection Agency tightened the primary annual PM2.5 standard to 9 μg/m³, increasing the number of communities classified as out of attainment and raising the value of accurate, fine-scale exposure estimates (6).

The central measurement problem is spatial.

Corresponding author: Nathan Tan, E-mail: nathantan2027@gmail.com.
Copyright: © 2026 Nathan Tan *et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Accepted August 5, 2026
https://doi.org/10.70251/HYJR2348.4410241039

Monitoring networks are accurate but sparse, and a single station is often treated as representative of an entire county even though PM2.5 can change substantially from block to block with proximity to highways, rail, ports, industrial facilities, and green space (7, 8). This mismatch matters most for the populations monitoring is meant to protect, because communities of color and lower-income neighborhoods in the United States are exposed on average to higher PM2.5 than the rest of the population, a pattern that persists across emission sectors and income levels (9–11). Coarse monitoring averages these gradients away, so the neighborhoods with the highest burden are also the least well resolved.

Low-cost sensor networks such as PurpleAir partially close this gap by providing dense, continuous measurements that, after correction, agree well with reference instruments (12, 13). Sensors alone do not produce a continuous exposure surface and cannot report concentrations where no sensor exists. Supervised machine learning offers a route from scattered point measurements to a complete map, and tree-based and deep models have estimated PM2.5 from meteorological, land-use, and satellite predictors at regional and national scale (14–19). The accuracy of such a map, and any conclusion drawn from it, depends on whether the model has been evaluated as it will be used: prediction at locations absent from the training data.

The dominant practice holds out a random fraction of all observations for testing, but when a single sensor contributes many daily records, random splitting places records from that same sensor on both sides of the partition. A flexible model can then memorize a sensor's local mean and recover it at test time, so the reported score measures interpolation in time at known locations rather than prediction at new locations (20–22). Spatial dependence is known to inflate random-split estimates, yet much applied air-quality work still reports them. A second, subtler issue arises when such models are used for equity analysis: if sociodemographic indicators are included as predictors, any subsequent correlation between the model's output and those same indicators is partly built in by construction. This study therefore enforces a strict separation between variables used for exposure prediction and variables reserved for independent equity evaluation.

LITERATURE REVIEW

Machine learning for PM2.5 estimation

Tree-based ensembles are common for ground-level PM2.5 because they capture nonlinear interactions among meteorology, land use, and emission proxies. Hu *et al*. estimated daily PM2.5 across the conterminous United States with a random forest combining satellite aerosol optical depth, meteorology, and land-use variables, reaching cross-validated $R^2 \approx 0.80$ (14). Satellite-driven approaches have since been refined with explicit attention to measurement error and to honest evaluation of spatiotemporal models over large regions (15, 16). Machine learning has produced daily high-resolution PM2.5 estimates within Texas (17, 18), and reviews report that nonparametric models generally outperform linear regression while warning that headline accuracy depends on the evaluation design (19).

Low-cost sensing and mobile monitoring

Dense observation is the second ingredient of a hyperlocal map. Surveys of low-cost air-quality sensors judge them fit for many community and supplementary monitoring purposes (12), and a United-States-wide correction brings PurpleAir measurements into close agreement with federal reference monitors (13). Mobile monitoring offers an alternative, mapping block-level pollution with instrumented vehicles and comparing strategies for converting such campaigns into continuous surfaces (7, 8).

Spatial validation

The risk of optimistic evaluation under spatial dependence is well documented outside the air-quality literature. Cross-validation strategies for temporally, spatially, or hierarchically structured data show that ignoring such structure inflates performance estimates (20), and including location descriptors lets flexible models reproduce training data while failing to predict at new locations, so spatial variable selection and spatial cross-validation are needed (21). Spatial k-fold cross-validation formalizes estimation of true prediction performance (22); leave-one-sensor-out validation is its natural instance for a sensor network. An important distinction is that a withheld sensor can be predicted either with or without access to its own recent measurement history. The two settings answer different questions, transfer to a new sensor site versus prediction at a location with no sensor at all, and only the second corresponds to deployment at unmonitored census tracts.

Environmental justice in exposure

A large body of evidence documents systematic

disparities in PM2.5 exposure in the United States. Nearly all major emission sectors contribute to a systemic exposure disparity experienced by people of color (9), exposure gaps across racial and income groups have been quantified nationally (10), and the most polluted areas have remained disproportionately populated by lower-income and minority residents over time (11). These studies motivate the deployment analysis below, which asks whether a spatially honest model, trained without any sociodemographic input, independently reproduces such disparities at neighborhood scale, and whether its errors are equitably distributed.

This study makes three contributions. First, it benchmarks regression algorithms for hyperlocal PM2.5 prediction across four Texas metropolitan areas under random validation, spatial-transfer LOSO-CV, and cold-start LOSO-CV, quantifying the generalization gaps among the three. Second, it identifies which feature families drive predictive skill using feature-importance analysis and targeted ablations, including an ablation that quantifies the marginal predictive value of EJScreen variables while keeping them out of the deployed model. Third, it deploys the cold-start model as a true annual exposure surface built from daily predictions and audits the result against an established environmental justice index that the model never saw, examining whether prediction errors are uniform across the burden gradient.

METHODS AND MATERIALS

Problem formulation

Let $s \in \{1, \dots, S\}$ index the $S = 226$ sensors and let t index calendar days within the study year. For sensor s on day t, the daily mean concentration $y_{s,t}$ (in μg/m³) is observed together with a predictor vector $x_{s,t}$ that collects meteorological, temporal, autoregressive, and spatial features defined below. The goal is a regression function (Equation 1) that minimizes prediction error at locations not seen during training, since the intended use is estimation at unmonitored census tracts.

$$\hat{y}_{s,t} = f(x_{s,t}) \qquad \text{(Eq. 1)}$$

Validation protocols

Three ways of forming training and test sets are contrasted. In random validation the full set of records is shuffled and partitioned into 80% training and 20% test records without regard to sensor identity, so records from a given sensor generally appear in both partitions. In leave-one-sensor-out cross-validation the data are partitioned by sensor: for each held-out sensor s' the model is trained on the remaining sensors (Equation 2) and evaluated only on the held-out records. This is repeated for every sensor in the network, so each of the 226 sensors is withheld in turn and every record receives exactly one out-of-fold prediction from a model that never saw its sensor. The reported LOSO score aggregates these predictions across all 226 folds.

$$D_{train}^{(s')} = \{(x_{s,t}, y_{s,t}) : s \neq s'\} \qquad \text{(Eq. 2)}$$

LOSO-CV is further evaluated in two settings that answer different questions. In the transfer setting, the held-out sensor's autoregressive predictors (its own one- and seven-day PM2.5 history) remain available at test time; this measures accuracy at a newly installed sensor whose recent readings exist but which contributed nothing to model fitting. In the cold-start setting, all autoregressive predictors are removed from the model, so the held-out sensor is predicted from meteorology, calendar, and location alone; this is the only setting that corresponds to a census tract with no sensor history, and it is therefore the setting used for deployment. Random validation measures interpolation at known locations and is an optimistic upper bound; transfer LOSO-CV measures prediction at a new site with local history; cold-start LOSO-CV measures prediction at a genuinely unmonitored location.

Evaluation metrics

Predictions are scored with the coefficient of determination, the root-mean-square error, and the mean absolute error (Equation 3), where $\bar{y}$ is the mean observed concentration and n is the number of scored records. RMSE is emphasized as the primary error metric because it penalizes the large errors that occur during pollution episodes more heavily than MAE.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$
$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad \text{(Eq. 3)}$$

Ensemble

The ensemble used for deployment is the unweighted mean of the three gradient-boosted and bagged tree models (Equation 4), which reduces variance relative to any single tree model while requiring no additional parameters to estimate.

$$\hat{y}_{s,t}^{\text{ens}} = 1/3\left(\hat{y}_{s,t}^{\text{RF}} + \hat{y}_{s,t}^{\text{LGBM}} + \hat{y}_{s,t}^{\text{XGB}}\right) \qquad \text{(Eq. 4)}$$

Data and study design

Study area and observations

The study covers the four largest metropolitan areas in Texas: Austin, Dallas–Fort Worth (DFW), Houston, and San Antonio, defined as fixed sets of counties (26 counties in total; the full county list is provided in the study repository). These regions differ in industrial composition, traffic density, and climate, providing variation in concentration and its spatial structure.

PurpleAir data acquisition and quality assurance

Candidate sensors were the 240 outdoor PurpleAir sensors within the study counties identified at the original data pull. For the revision, the complete hourly history of every candidate sensor was re-downloaded from the PurpleAir API (25) on July 30, 2026 (fields: pm2.5_cf_1_a, pm2.5_cf_1_b, pm2.5_atm_a, pm2.5_atm_b, humidity, temperature, pressure; 60-minute averages; December 18, 2024 through January 2, 2026, America/Chicago time). Sensor locations and installation environments were verified against API metadata, with only outdoor sensors (location_type = 0) eligible for retention; verification confirmed that all 240 candidates are outdoor units with valid locations, so no sensor was excluded at this step.

Quality assurance followed the channel-level protocol of Barkjohn *et al.* (13). Each PurpleAir unit carries two independent laser counters (channels A and B). A channel-hour was considered valid if its cf_1 concentration was non-negative and below 500 µg/m³, which removes stuck and physically implausible readings at the channel level. Daily channel means were computed from valid hours in local (America/Chicago) days, and a sensor-day was retained only if both channels had at least 18 of 24 valid hourly means (75% completeness). Days on which the two daily channel means disagreed by more than 5 µg/m³ and more than 61% of their mean were excluded. The retained daily concentration is the mean of the two channel means, corrected to reference-equivalent PM2.5 with the United States–wide equation of Barkjohn *et al.* (13), using the sensor's own daily-mean relative humidity (RH):

$$PM_{2.5} = 0.524 \times PA_{cf1} - 0.0862 \times RH + 5.75 \qquad \text{(Eq. 5)}$$

for $PA_{cf1} \leq 343$ µg/m³, with the high-concentration extension of the same reference applied above that level; corrected values are floored at zero. Of 72,208 candidate sensor-days in the download window, 66,711 passed the completeness screen and 62,173 additionally passed the channel-agreement screen. After restriction to calendar year 2025 and to sensor-days with complete predictors (below), the final analysis table contains 54,325 sensor-day records from 226 sensors (53 in Austin, 56 in DFW, 71 in Houston, and 46 in San Antonio; Figure 1), falling in 188 unique census tracts. The median absolute A–B disagreement among retained days was 0.91 µg/m³.

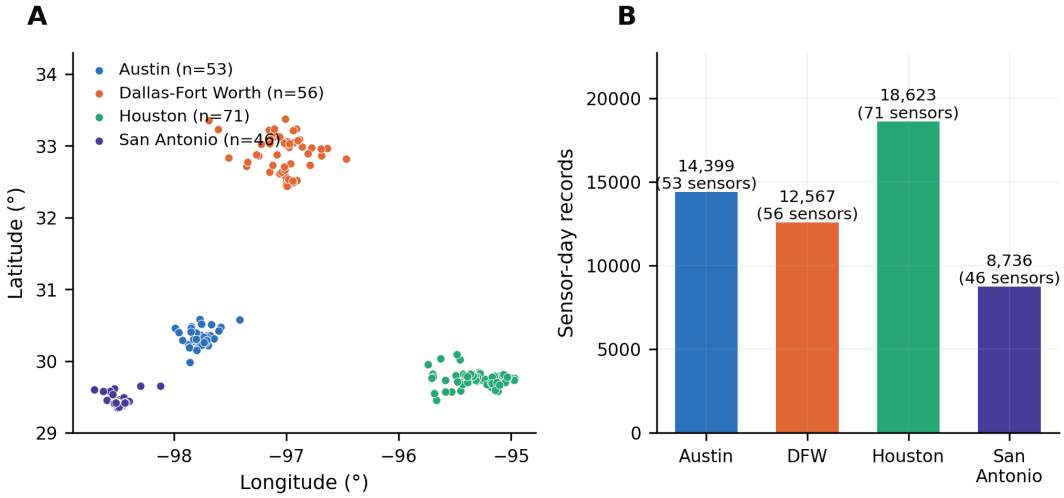


Figure 1. *Study area and sensor coverage. (A) Locations of the 226 quality-assured PurpleAir sensors across the four Texas metropolitan areas. (B) Sensor-day records contributed by each metropolitan area, with sensor counts; the four metros contribute 54,325 records in total.*

The sensor-level attrition from 240 candidates to 226 decomposes as follows (per-sensor accounting in the repository file results/sensor_attrition.csv). 7 sensors had no 2025 day that passed the completeness and channel-agreement screens (2 in Austin, 4 in DFW, 1 in Houston). 6 sensors, although labeled DFW in the original data pull, are located in Collin or Johnson County, outside the fixed study-area county definition, and are excluded by the study-area restriction that the deployment also uses; the original submission had included them inconsistently. 1 Houston sensor retained only three QA-valid days, none with the complete seven-day history required for the autoregressive predictors. No sensor was excluded for being indoor or unlocatable.

Predictor variables

Each sensor-day record was joined to predictors from four families (Table 1). Meteorological predictors, daily mean temperature, relative humidity, and surface pressure, were taken from the ERA5 reanalysis via the Open-Meteo historical archive (23, 24), evaluated at the 0.1° grid cell containing each location. Reanalysis meteorology is used for both training and deployment so that the meteorological inputs at unmonitored tracts are constructed identically to those seen in training; the PurpleAir units' onboard temperature and RH readings are used only inside the sensor correction (Equation 5), never as model predictors. Temporal predictors encode seasonality and weekly structure through day of year, day of week, month, and season. Autoregressive predictors summarize each sensor's recent history through the one-day lag, the seven-day lag, and the mean of the seven preceding days (at least four required); these are present only in the transfer configuration. Spatial predictors are latitude, longitude, and metropolitan area (one-hot). The census-tract identifier is *not* a predictor: a high-cardinality location ID invites the memorization failure documented by Meyer *et al*. (21) and is undefined at tracts absent from training.

EJScreen environmental-justice indicators, three demographic indicators (people of color, low income, linguistically isolated), four pollution-proximity indicators (traffic, diesel PM, Superfund, and Risk Management Plan facility proximity), and the composite five-factor EJ index, all as Texas state percentiles from EJScreen version 2.3 (26), are attached to sensors and tracts by census-tract GEOID. They are reserved for the independent equity audit and are excluded from every deployed model. Their marginal predictive value is quantified once, in a dedicated ablation, and they play no other modeling role.

Models and validation

Models

Four supervised regression algorithms constitute the primary benchmark and were evaluated under the complete protocol suite. Multiple linear regression (MLR, on standardized features) provides an interpretable baseline; random forest (RF) is a bagged tree ensemble; LightGBM (LGBM) and XGBoost (XGB) are gradient-boosted tree ensembles. The deployment ensemble is the unweighted mean of RF, LGBM, and XGB (Equation 4). Hyperparameters are listed in Table 2. Two further models, support vector regression (SVR) with an RBF kernel and a long short-term memory (LSTM) network, are computationally impractical to retrain once per LOSO fold and are therefore reported separately as exploratory models under a defined, reproducible protocol (below); they are excluded from the primary ranking.

Table 1. *Variable families and their roles. PM2.5 is the dependent variable. EJScreen variables are audit variables only, they are not model predictors.*

Family	Variables	Role
Air quality	PM2.5 (μg/m³), QA'd and corrected (13)	Dependent variable
Meteorological	Temperature, relative humidity, surface pressure (ERA5)	Predictor (all configurations)
Temporal	Month, day of year, day of week, season	Predictor (all configurations)
Autoregressive	1-day lag, 7-day lag, 7-day rolling mean	Predictor (transfer only)
Spatial	Latitude, longitude, metropolitan area	Predictor (all configurations)
EJScreen (audit only)	EJ index; % people of color; % low income; % linguistically isolated; traffic, diesel PM, Superfund, RMP proximity	Audit variables, never predictors in deployed models

Table 2. *Model configurations. The ensemble is the unweighted mean of RF, LightGBM, and XGBoost predictions. Seeds fixed at 42 throughout.*

Model	Key hyperparameters
MLR	Standardized features; ordinary least squares
Random forest	n_estimators = 200, max_depth = 12, min_samples_leaf = 4
LightGBM	n_estimators = 500, learning rate = 0.05, max_depth = 8, num_leaves = 31, subsample = 0.8, colsample = 0.8
XGBoost	n_estimators = 500, learning rate = 0.05, max_depth = 6, subsample = 0.8, colsample = 0.8
SVR (exploratory)	RBF kernel, C = 1.0, gamma = scale, epsilon = 0.1; standardized features; training folds subsampled to 20,000 records
LSTM (exploratory)	1 layer, 50 units, dropout 0.2; input = 7-day sequence of [PM2.5, temperature, RH, pressure] plus static current-day covariates; Adam, lr = 0.001, batch 256, ≤25 epochs with early stopping (patience 3) on a grouped validation split of 10% of training sensors
Ensemble	Unweighted mean of RF, LightGBM, XGBoost

Exploratory-model protocol

SVR and the LSTM are evaluated under spatial group 10-fold cross-validation: the 226 sensors are partitioned into ten disjoint folds, and every record is predicted exactly once by a model that never saw its sensor during training. This preserves the spatial-honesty property of LOSO-CV at a fraction of the retraining cost. All remaining details, feature scaling fitted on training folds only, the SVR training-set subsample cap of 20,000 records, the LSTM sequence construction (complete seven-day history required, reducing its usable set to 48,365 records), architecture, optimizer, batch size, early-stopping rule, and random seed (42), are fixed and stated in Table 2, and the implementation is included in the study repository. Because their protocol differs from the 226-fold LOSO of the primary models, SVR and LSTM results are reported in their own table and are not ranked against the primary models.

Experimental setup

All models were implemented in Python 3.14 (scikit-learn 1.8, XGBoost 3.1, LightGBM 4.6, PyTorch 2.10) and trained on identical data with identical preprocessing. The full pipeline, download, quality assurance, feature construction, all validation protocols, deployment, and figures, runs from a single scripted sequence in the study repository. The complete 226-fold LOSO evaluation was computed on the Georgia Tech PACE Phoenix cluster as a Slurm array job (eight tasks of 24 cores each; worker and job scripts in the repository) using the identical version-pinned Python environment, with per-fold checkpointing that makes the computation restartable and byte-reproducible; all remaining stages run in under two hours on a 22-core workstation with 31 GB of memory. Sensitivity of the headline results to hyperparameters was checked with a grid over the gradient-boosting settings (max_depth $\in$ {4, 6, 8, 10}, learning rate $\in$ {0.01, 0.05, 0.1}, n_estimators $\in$ {200, 500, 800}; 36 configurations per algorithm) under spatial three-fold grouped cross-validation on the transfer feature set. Across the full grid, spatial-CV R^2 ranges from 0.493 to 0.862 for XGBoost and from 0.498 to 0.861 for LightGBM, so performance is clearly sensitive to grossly under-fit settings: every configuration below $R^2 \approx 0.66$ combines the smallest learning rate (0.01) with too few boosting rounds. Among adequately fit configurations (learning rate ≥ 0.05), the range narrows to 0.663–0.862 (XGBoost) and 0.660–0.861 (LightGBM). The configurations used throughout this study (Table 2) were adopted a priori from the original submission and were not tuned on this grid; they score 0.829 (XGBoost) and 0.827 (LightGBM) under the same protocol, within 0.033 and 0.034 of the respective grid maxima. The conclusions are therefore robust to any reasonably fit setting, but not to arbitrary ones, and no hyperparameter search was performed on the evaluation folds.

Deployment to census tracts

The deployed model is the cold-start ensemble: RF, LightGBM, and XGBoost trained on all 54,325 sensor-day records using only meteorological, temporal, and spatial predictors (no autoregressive history, no EJScreen

variables, no tract identifier). Deployment covers the 3,758 census tracts of the study-area counties that carry complete EJScreen attributes (2020 Census geography; tract centroids and populations from the EJScreen tract file).

The feature-construction procedure at an unmonitored tract is identical to training, with each predictor assigned as follows. For every tract and every day of calendar year 2025: temperature, relative humidity, and pressure are the ERA5 daily values at the 0.1° cell containing the tract's population centroid, the same source, resolution, and cell mapping used in training; day of year, day of week, month, and season are those of the actual calendar day; latitude and longitude are the tract centroid coordinates; the metropolitan-area indicator follows the tract's county. No variable is fixed, imputed, or borrowed from sensors. Daily predictions from the three models are averaged (Equation 4), floored at zero, and the tract's annual mean is the average of its 365 daily predictions. This is a true annual average over the year's actual meteorology, and it is therefore compared directly with the EPA annual standard of 9 μg/m³ (6). Predicted annual means at the 226 sensor locations, generated by the same procedure, support the sensor-level analyses.

Environmental justice audit and fairness analysis

The audit tests whether the deployed surface, built with no sociodemographic input, reproduces known exposure disparities. Tract-level predicted annual PM2.5 is correlated (Pearson and Spearman) with the EJScreen EJ index and with each of the seven sub-indicators; tracts are sorted into quartiles of the EJ index; and exceedance of the 9 μg/m³ annual standard is summarized by quartile and by metropolitan area, with population counts from the EJScreen tract file. Because the audited variables are excluded from the model, any association found is a property of the predicted concentration field and the audited index, not an echo of the model's inputs.

The fairness analysis asks whether predictive skill itself varies across the burden gradient. LOSO out-of-fold errors of the ensemble, under both the transfer and the cold-start configuration, are grouped by the EJ-index quartile of each sensor's census tract, and RMSE, MAE, and mean bias are computed per quartile with 95% confidence intervals from a cluster bootstrap over sensors (1,000 resamples). Because sensors, not tracts, are the sampling units of model evaluation, this analysis characterizes error only where sensors exist; its limitations for unmonitored tracts are taken up in the Discussion.

RESULTS

Predictive performance and the generalization gaps

Table 3 reports R^2 for the four primary models and the ensemble under random splitting and LOSO-CV, in both the transfer and cold-start predictor configurations, and Figure 2 contrasts the protocols graphically. Two distinct comparisons matter, and they behave differently. First, within the transfer configuration, random and LOSO scores nearly coincide for every model (ensemble 0.834 versus 0.825; MLR 0.476 versus 0.469): when the withheld sensor's own recent history is available at test time, the autoregressive predictors carry most of the signal, and little location-specific memorization is possible or needed. Second, within the cold-start configuration, the two protocols diverge sharply: random splitting still awards the ensemble R^2 = 0.838, but LOSO-CV, the setting that corresponds to deployment, yields 0.662, an inflation of 0.176 R^2. This is the random-

Table 3. *Coefficient of determination under random 80/20 splitting and LOSO-CV, for the transfer configuration (withheld sensor's own history available) and the cold-start configuration (no site history; deployment setting). All LOSO values from the complete 226-fold protocol, every sensor withheld in turn, on identical records; RMSE and MAE for the ensemble are given in the text.*

Model	Transfer: random R^2	Transfer: LOSO R^2	Cold start: random R^2	Cold start: LOSO R^2
Multiple linear regression	0.476	0.469	0.051	0.037
Random forest	0.741	0.741	0.738	0.650
LightGBM	0.851	0.837	0.845	0.623
XGBoost	0.859	0.847	0.851	0.629
Ensemble (RF + LGBM + XGB)	0.834	0.825	0.838	0.662

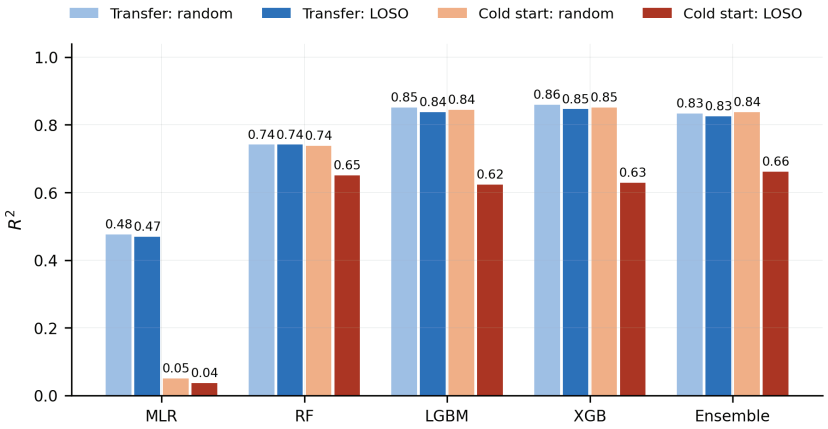


Figure 2. *Predictive performance (R^2) under random 80/20 splitting and LOSO-CV, in the transfer and cold-start predictor configurations. With site history available (left bar pair per model), random and LOSO scores nearly coincide; without site history (right bar pair), random splitting substantially overstates LOSO performance, and the linear baseline collapses.*

split optimism the paper is about, and it appears exactly where spatial generalization is actually demanded. The linear baseline makes the same point in extreme form: its transfer performance (0.469) rests almost entirely on the lag features, and with them removed it collapses to 0.037 under cold-start LOSO-CV.

The cold-start LOSO column measures the deployment-relevant setting: prediction with no site history at all. Relative to transfer, removing autoregressive predictors costs the ensemble 0.163 in R^2 (from 0.825 to 0.662) and raises RMSE from 1.71 to 2.37 μg/m³. This gap is the honest price of predicting at an unmonitored location, and all deployment results below are generated by the cold-start model. Per-metro cold-start LOSO performance is consistent across regions (ensemble R^2 0.58–0.71; Supplementary Table S1), with no metro in which performance collapses.

Exploratory models

Table 4 reports the two exploratory models under their defined protocol (random 80/20 plus spatial group 10-fold CV). Because the LSTM's input sequence contains the site's own seven-day PM2.5 history, it operates in the transfer setting by construction; its performance is nearly identical under record-level random splitting (R^2 = 0.714) and spatial holdout (R^2 = 0.717), and in both cases remains well below the transfer performance of the gradient-boosted ensemble. SVR is weak under both protocols (R^2 = 0.582 random, 0.552 spatial). Neither exploratory model challenges the primary benchmark's conclusions, and because their protocol differs they are not ranked against Table 3.

Table 4. *Exploratory models under the defined reproducible protocol (Methods). Not comparable with Table 3 because of the different cross-validation scheme and, for the LSTM, the restricted record set.*

Model	Random 80/20 R^2	Spatial group 10-fold R^2	Records
SVR (RBF)	0.582	0.552	54,325
LSTM	0.714	0.717	48,365

Feature importance and ablations

Figure 3 shows the share of total gain of the leading predictors in the XGBoost model, for the transfer configuration (panel A) and the cold-start configuration (panel B). In the transfer configuration the leading contributors are the one-day PM2.5 lag (36.9%), day of year (16.4%), the seven-day rolling mean (11.8%), temperature (9.1%); in the cold-start configuration, where no site history is available, the model leans on day of year (36.7%), latitude (13.4%), temperature (11.5%), longitude (10.2%).

Two ablations complete the picture. First, the transfer-versus-cold-start contrast above quantifies the value of site history (0.163 in ensemble R^2). Second, adding the eight EJScreen variables to the predictor set, done

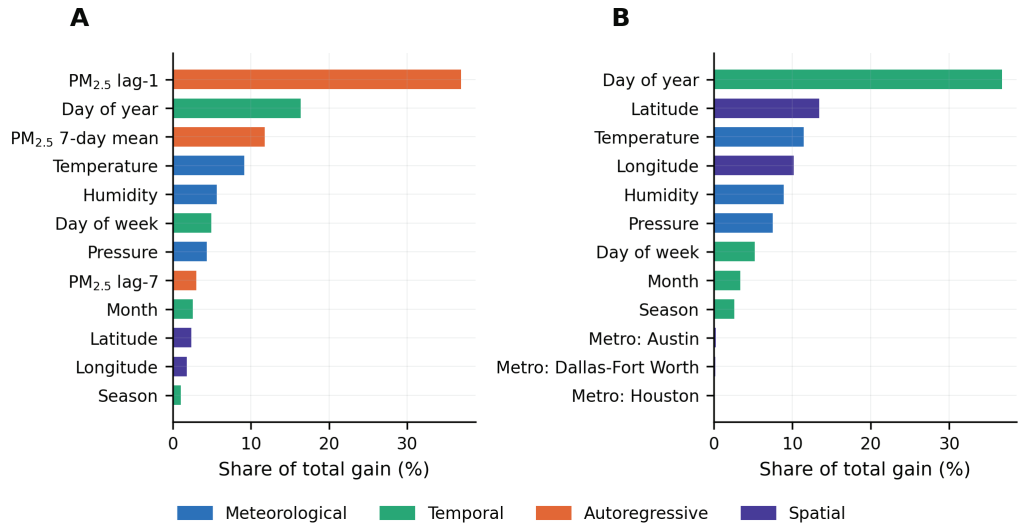


Figure 3. *Leading predictors by share of total gain in the XGBoost model, colored by feature family. (A) Transfer configuration, in which autoregressive site history is available. (B) Cold-start (deployment) configuration, in which meteorological, temporal, and spatial predictors carry all of the signal.*

only in this ablation, never in deployment, changes ensemble LOSO R^2 from 0.825 to 0.817 in the transfer setting and from 0.662 to 0.634 in the cold-start setting. In both settings the EJScreen variables *reduce* spatially honest performance, and markedly so in the cold-start configuration. This is the signature of the location-descriptor failure documented by Meyer *et al*. (21): static tract profiles let a flexible model memorize the identity of training neighborhoods, which inflates random-split scores but does not transfer to withheld locations. Excluding the EJScreen family from the deployed model therefore improves deployment accuracy in addition to guaranteeing that the equity audit is independent of the audited variables.

Spatial deployment

The cold-start ensemble was used to generate daily predictions for every study-area tract and every day of 2025, averaged to true annual means (Methods). Predicted annual PM2.5 ranges from 4.73 to 7.79 µg/m³ across tracts in Austin, 4.46 to 8.49 in DFW, 3.24 to 13.16 in Houston, and 4.19 to 15.00 in San Antonio (Supplementary Figure S4 maps the full surface). The average within-metro spatial spread is 6.96 µg/m³, confirming that intra-urban variation is substantial relative to differences between metro-wide means and that a single monitor cannot represent a metropolitan area.

Environmental justice analysis

Across 3,758 census tracts, predicted annual PM2.5, from a model with no sociodemographic input, is positively but modestly correlated with the EJScreen EJ index (Pearson $r = 0.129$, $p < 10^{-14}$; Spearman $\rho = 0.122$; Figure 4). The association has the same character in the raw observations: observed annual-mean PM2.5 at the 226 sensor sites correlates with the EJ index at $r = 0.212$ ($p = 0.001$), and predicted annual means at those sites at $r = 0.337$. Because the EJ index is not a model input, these associations are read off the predicted concentration field rather than built into it. The contrast with an analysis in which the EJScreen variables are model inputs is itself a central result: when the EJScreen variables were included as predictors and the surface was generated under fixed conditions (a single representative set of meteorological and temporal values applied to every tract, rather than an average of daily predictions over the year's actual meteorology), the same correlation appeared as $r = 0.875$. Removing the audited variables from the model and averaging real daily predictions collapses the association to $r = 0.129$, an order-of-magnitude difference that quantifies how strongly predictor circularity and standardized-condition surfaces can inflate apparent environmental-justice signals.

Of the 3,758 tracts, 95 (2.5%) are predicted to exceed the 9 µg/m³ annual standard; these tracts contain approximately 492,413 residents, 2.9% of the study-

area population. Sorting tracts into quartiles of the EJ index makes the gradient explicit: the tract exceedance rate rises from 0.2% in the lowest-burden quartile through 1.3% and 2.7% to 6.0% in the highest-burden quartile (Figure 5). Across metropolitan areas, predicted exceedance rates are 0.1% of tracts in Houston, 19.3% in San Antonio, 0.0% in DFW, and 0.0% in Austin. The populations carrying the greatest social and pollution-proximity burden are disproportionately the populations predicted to live above the annual standard.

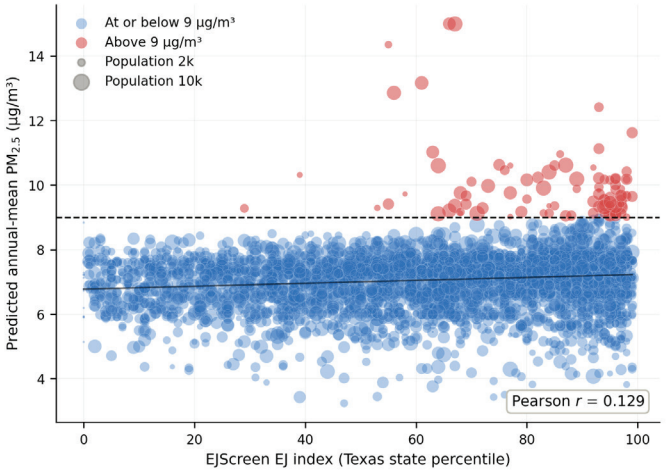


Figure 4. *Predicted annual-mean PM2.5 (true annual average of daily predictions from the EJ-free cold-start ensemble) against the EJScreen EJ index across 3,758 census tracts, with bubble area proportional to tract population. Tracts above the 9 µg/m³ annual standard (red) concentrate at high values of the index.*

Error patterns across the exposure gradient

If the ensemble were systematically less accurate in high-burden neighborhoods, the apparent inequity could partly reflect uneven model skill. Figure 6 shows LOSO out-of-fold RMSE, MAE, and mean bias by EJ quartile for both the transfer and the cold-start ensemble, with cluster-bootstrap 95% confidence intervals. Cold-start RMSE spans 2.10–2.75 µg/m³ across quartiles and MAE spans 1.44–1.89 µg/m³. The variation is moderate and non-monotonic: RMSE is largest in the third quartile (2.75 µg/m³) and smallest in the highest-burden quartile (2.10 µg/m³), and mean bias is small in every quartile (from −0.14 to +0.16 µg/m³), with confidence intervals spanning zero throughout every quartile. Predictive error is therefore not uniform across the gradient, but its pattern works against, rather than for, an artifactual disparity: the highest-burden quartile is the best-predicted of the four, so differential inaccuracy in burdened neighborhoods cannot be generating the positive association reported above. This favorable error pattern remains a necessary rather than sufficient condition for an exposure-based reading of Figure 5, since it is measured only where sensors exist (see Limitations).

DISCUSSION

The results carry a methodological message and a substantive one. Methodologically, the danger of random-split validation is concentrated exactly where deployment lives: when autoregressive site history is present, random and LOSO scores agree, but in the cold-start configuration random splitting overstates ensemble

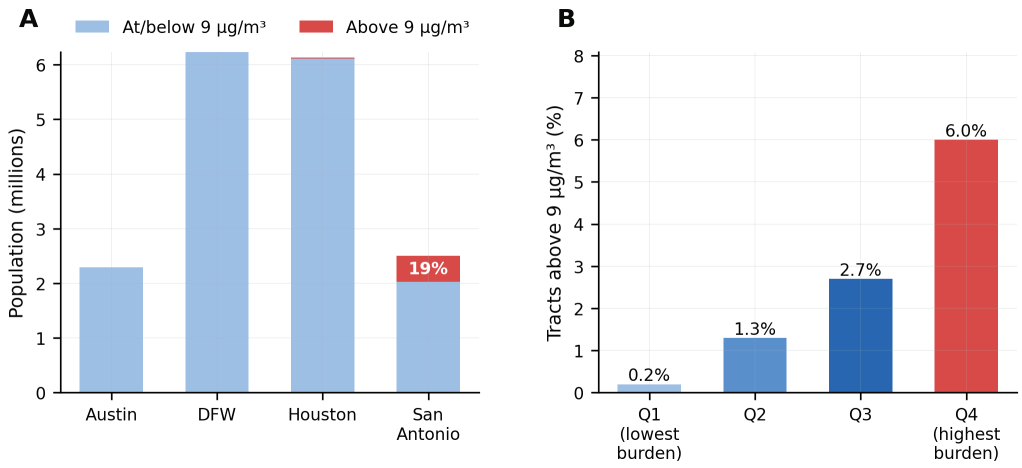


Figure 5. *Predicted population exposure and exceedance burden. (A) Population at or below versus above the 9 µg/m³ annual standard by metropolitan area. (B) Share of tracts above the standard by quartile of the EJScreen EJ index.*

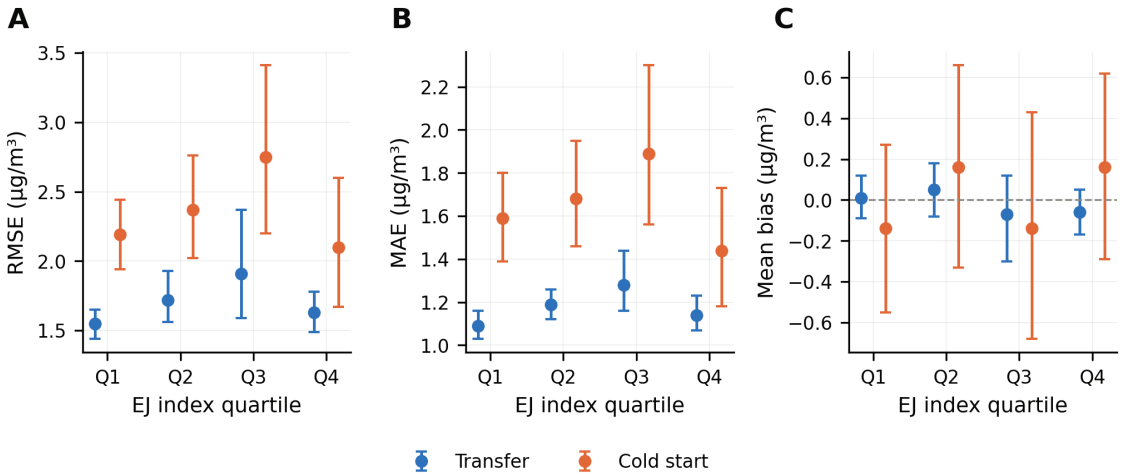


Figure 6. *Ensemble LOSO out-of-fold error within quartiles of the EJ index, for the transfer and cold-start configurations, with cluster-bootstrap 95% confidence intervals over sensors. (A) RMSE. (B) MAE. (C) Mean bias.*

skill by 0.176 R^2, and even a spatially honest LOSO evaluation overstates deployment accuracy if site history silently remains in the predictor set. A linear model that looks moderately capable with site history collapses entirely without it; recurrent sequence modeling of site history offers no advantage over gradient-boosted trees once evaluated under a defined spatial protocol; static sociodemographic predictors inflate random-split scores while degrading spatially honest performance; and the deployment-relevant number for the tree ensemble is the cold-start LOSO R^2 of 0.662, not the more flattering transfer or random-split values. Any air-quality map built for unmonitored places should report the cold-start figure.

Substantively, the independent audit tempers the association obtained when the EJScreen variables are used as predictors and the surface is generated under fixed conditions rather than averaged over daily predictions. A model trained without sociodemographic input still finds a positive, statistically significant association between predicted annual PM2.5 and EJScreen burden, and predicted exceedance of the annual standard remains concentrated in the highest-burden quartile (6.0% of tracts, versus 0.2% in the lowest) and in San Antonio. But the association is modest ($r = 0.129$), an order of magnitude below the $r = 0.875$ obtained when the audited variables were inputs to the model. This contrast is a caution for applied environmental-justice modeling: much of the previously reported signal was contributed by the audited indicators themselves rather than by independent structure in the concentration field. The residual association is consistent in direction with national evidence that PM2.5 exposure is higher in communities of color and lower-income communities (9–11); it cannot be an artifact of predictor circularity, and because the highest-burden quartile is also the best-predicted, uneven model skill is unlikely to be its driver either. The pattern is consistent with a real but geographically concentrated exposure disparity, with the caveats below.

Limitations

Several limitations qualify these conclusions, and the equity findings should be read as strong association rather than settled fact. First, the fairness analysis characterizes error only where sensors exist; in unmonitored tracts, disproportionately located in some high-burden neighborhoods, accuracy is unobserved, and the deployed surface there is a spatial extrapolation whose uncertainty is not fully quantified. Second, PurpleAir sensors are not sited by probability sampling; participation requires purchasing a device and Wi-Fi, so the network may under-represent exactly the neighborhoods of greatest interest, and measurement bias in sensor placement could propagate into the surface (13). Third, although the audited EJ variables are excluded from the model, meteorology and coordinates are spatially smooth, and the analysis remains correlational: it documents association, not causal attribution of exposure to specific sources. Fourth, low-cost sensors are noisier than reference monitors even after correction, and the correction equation (13) was developed against national reference data rather than Texas-specific

collocations. Fifth, the EJScreen indicators are strongly intercorrelated, which limits interpretation of any single sub-indicator (Supplementary Figure S2). Finally, the study covers a single calendar year and cannot speak to multi-year trends or extreme episodes such as wildfire smoke intrusions.

These limitations point to clear extensions. Sensor-network expansion targeted at unmonitored high-burden tracts would convert extrapolation into measurement. Higher-resolution land-use data, traffic counts, satellite aerosol optical depth, or chemical-transport-model output could reduce cold-start uncertainty (16, 17). Multi-year records would enable trend and episode analysis, and pairing the model with causal-inference methods would move the environmental-justice analysis from association toward attribution.

CONCLUSION

This study benchmarked machine learning models for hyperlocal PM2.5 prediction across four Texas metropolitan areas under random, transfer LOSO, and cold-start LOSO validation, and used the cold-start model, trained without sociodemographic inputs, to build a true annual exposure surface and audit exposure inequity. Spatial validation changed the apparent ranking of models, and the transfer/cold-start distinction quantified the additional cost of predicting where no sensor history exists: the deployment ensemble reaches $R^2 = 0.662$ and RMSE = 2.37 μg/m³ at locations absent from training with no site history. Deployed across 3,758 tracts with daily 2025 meteorology, the surface shows a positive but modest independent association between predicted annual PM2.5 and EJScreen burden ($r = 0.129$), far weaker than the $r = 0.875$ obtained when the audited variables were model inputs, with 95 tracts and approximately 492,413 residents above the 9 μg/m³ annual standard and exceedance concentrated in the highest-burden quartile. The audit thus both detects a real, geographically concentrated disparity and quantifies how much of the previously reported signal was circular. LOSO error varies moderately and non-monotonically across the gradient, with the highest-burden quartile predicted most accurately, so uneven skill does not account for the association. The work supports a simple standard for applied air-quality modeling: validate for the deployment setting (new locations, no site history), keep audited variables out of the model, and only then use the map to inform environmental-justice policy.

CONFLICT OF INTEREST

The authors declare no conflicts of interest related to this work.

DATA AND CODE AVAILABILITY

All analysis code, the QA'd modeling table, sensor list with metadata, per-configuration model outputs, deployment surface, and figure-generation scripts are provided in the study repository (https://github.com/nathantantexas/pm25-spatial-ej-paper), together with a data dictionary, environment specification (Python 3.14; package versions pinned), and fixed random seeds (42 throughout). Raw inputs: PM2.5 measurements were downloaded from the PurpleAir API (25) on July 30, 2026 (field list and QA thresholds in Methods; per-sensor raw hourly files are redistributed in the repository within the limits of PurpleAir's data license, with the download script provided for full re-acquisition); meteorology is the ERA5 reanalysis accessed through the Open-Meteo historical weather archive (23, 24) on 2026-07-30; environmental-justice indicators and tract populations are from EPA EJScreen version 2.3 (2024 release) (26); census-tract geography is the 2020 Census tract delineation. The repository README maps each table and figure of this manuscript to the script that generates it.

ARTIFICIAL INTELLIGENCE DISCLOSURE

During the preparation of this work, the authors used generative AI tools (Anthropic Claude, Fable 5) to assist with drafting model-implementation code. Following the use of these tools, the authors carefully reviewed, verified, and edited the content as necessary and take full responsibility for the accuracy, originality, and integrity of the published work.

REFERENCES

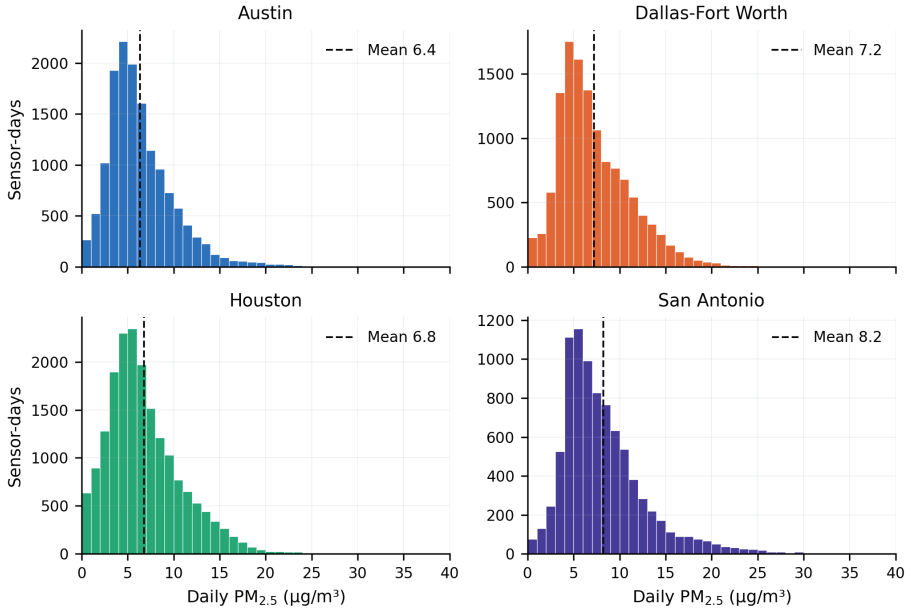
1. Pope CA III, Dockery DW. Health effects of fine particulate air pollution: lines that connect. *J Air Waste Manag Assoc.* 2006; 56 (6): 709-742. https://doi.org/10.1080/10473289.2006.10464485
2. Atkinson RW, Kang S, Anderson HR, Mills IC, Walton HA. Epidemiological time series studies of PM2.5 and daily mortality and hospital admissions: a systematic review and meta-analysis. *Thorax*. 2014; 69 (7): 660-665. https://doi.org/10.1136/thoraxjnl-2013-

204492

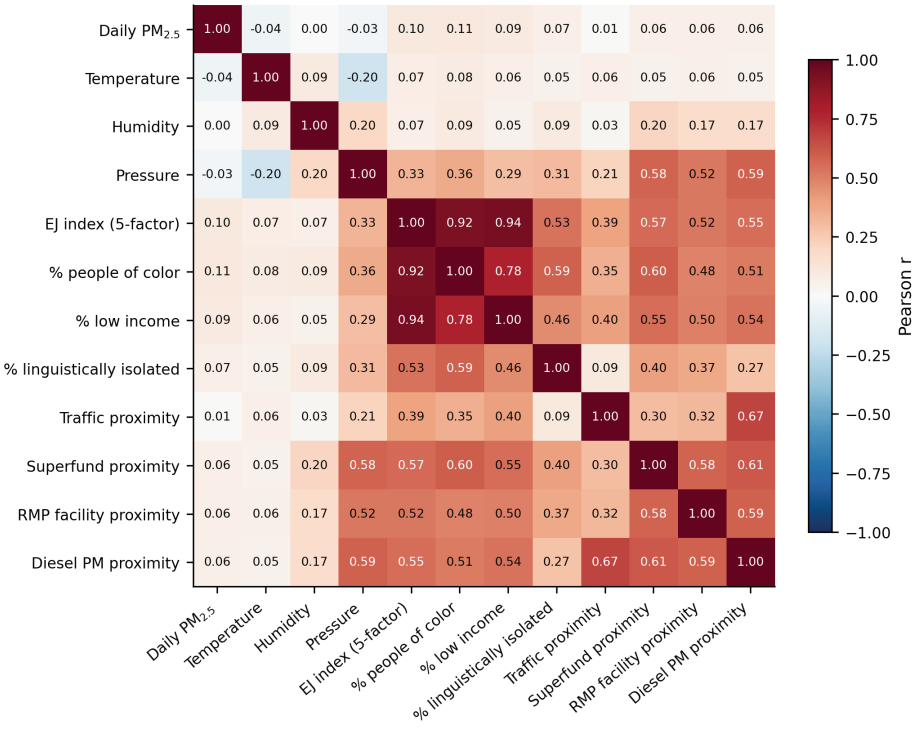
3. Landrigan PJ, Fuller R, Acosta NJR, Adeyi O, Arnold R, Basu NN, *et al*. The Lancet Commission on pollution and health. *Lancet*. 2018; 391 (10119): 462-512. https://doi.org/10.1016/S0140-6736(17)32345-0
4. World Health Organization. WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Geneva: World Health Organization; 2021. ISBN: 978-92-4-003422-8.
5. American Lung Association. State of the Air 2024. 2024. Available from: https://www.lung.org/research/sota (accessed on 2025-09-01).
6. U.S. Environmental Protection Agency. Reconsideration of the National Ambient Air Quality Standards for Particulate Matter. Fed Regist. 2024;89:16202. Available from: https://www.federalregister.gov/documents/2024/03/06/2024-02637/reconsideration-of-the-national-ambient-air-quality-standards-for-particulate-matter (accessed on 2025-09-01).
7. Apte JS, Messier KP, Gani S, Brauer M, Kirchstetter TW, Lunden MM, *et al*. High-resolution air pollution mapping with Google Street View cars: exploiting big data. *Environ Sci Technol*. 2017; 51 (12): 6999-7008. https://doi.org/10.1021/acs.est.7b00891
8. Messier KP, Chambliss SE, Gani S, Alvarez R, Brauer M, Choi JJ, *et al*. Mapping air pollution with Google Street View cars: efficient approaches with mobile monitoring and land use regression. *Environ Sci Technol*. 2018; 52 (21): 12563-12572. https://doi.org/10.1021/acs.est.8b03395
9. Tessum CW, Paolella DA, Chambliss SE, Apte JS, Hill JD, Marshall JD. PM2.5 polluters disproportionately and systemically affect people of color in the United States. *Sci Adv*. 2021; 7 (18): eabf4491. https://doi.org/10.1126/sciadv.abf4491
10. Jbaily A, Zhou X, Liu J, Lee TH, Kamareddine L, Verguet S, *et al*. Air pollution exposure disparities across US population and income groups. *Nature*. 2022; 601 (7892): 228-233. https://doi.org/10.1038/s41586-021-04190-y
11. Colmer J, Hardman I, Shimshack J, Voorheis J. Disparities in PM2.5 air pollution in the United States. *Science*. 2020; 369 (6503): 575-578. https://doi.org/10.1126/science.aaz9353
12. Morawska L, Thai PK, Liu X, Asumadu-Sakyi A, Ayoko G, Bartonova A, *et al*. Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: how far have they gone? *Environ Int*. 2018; 116: 286-299. https://doi.org/10.1016/j.envint.2018.04.018
13. Barkjohn KK, Gantt B, Clements AL. Development and application of a United States-wide correction for PM2.5 data collected with the PurpleAir sensor. *Atmos Meas Tech*. 2021; 14 (6): 4617-4637. https://doi.org/10.5194/amt-14-4617-2021
14. Hu X, Belle JH, Meng X, Wildani A, Waller LA, Strickland MJ, *et al*. Estimating PM2.5 concentrations in the conterminous United States using the random forest approach. *Environ Sci Technol*. 2017; 51 (12): 6936-6944. https://doi.org/10.1021/acs.est.7b01210
15. Just AC, De Carli MM, Shtein A, Dorman M, Lyapustin A, Kloog I. Correcting measurement error in satellite aerosol optical depth with machine learning for modeling PM2.5 in the northeastern USA. *Remote Sens*. 2018; 10 (5): 803. https://doi.org/10.3390/rs10050803
16. Just AC, Arfer KB, Rush J, Dorman M, Shtein A, Lyapustin A, *et al*. Advancing methodologies for applying machine learning and evaluating spatiotemporal models of fine particulate matter (PM2.5) using satellite data over large regions. *Atmos Environ*. 2020; 239: 117649. https://doi.org/10.1016/j.atmosenv.2020.117649
17. Ghahremanloo M, Choi Y, Sayeed A, Salman AK, Pan S, Amani M. Estimating daily high-resolution PM2.5 concentrations over Texas: machine learning approach. *Atmos Environ*. 2021; 247: 118209. https://doi.org/10.1016/j.atmosenv.2021.118209
18. Zhang K, Lin J, Li Y, Sun Y, Tong W, Li F, *et al*. Unmasking the sky: high-resolution PM2.5 prediction in Texas using machine learning techniques. *J Expo Sci Environ Epidemiol*. 2024; 34: 814-820. https://doi.org/10.1038/s41370-024-00659-w
19. Patel P, Patel S, Shah K, Desai K, Patel S, Shah M, *et al*. A systematic study on PM2.5 and PM10 concentration prediction in air pollution using machine learning and deep learning model. *Environ Chem Ecotoxicol*. 2025; 7: 1401-1415. https://doi.org/10.1016/j.enceco.2025.07.001
20. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, *et al*. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*. 2017; 40 (8): 913-929. https://doi.org/10.1111/ecog.02881
21. Meyer H, Reudenbach C, Wöllauer S, Nauss T. Importance of spatial predictor variable selection in machine learning applications: moving from data reproduction to spatial prediction. *Ecol Modell*. 2019; 411: 108815. https://doi.org/10.1016/j.ecolmodel.2019.108815
22. Pohjankukka J, Pahikkala T, Nevalainen P, Heikkonen J. Estimating the prediction performance of spatial models via spatial k-fold cross validation. *Int J Geogr Inf Sci*. 2017; 31 (10): 2001-2019. https://doi.org/10.1080/13658816.2017.1346255

23. Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, *et al*. The ERA5 global reanalysis. *Q J R Meteorol Soc*. 2020; 146 (730): 1999-2049. https://doi.org/10.1002/qj.3803
24. Zippenfenig P. Open-Meteo.com weather API [computer software]. Zenodo; 2023. Available from: https://open-meteo.com (accessed on 2026-07-30).
25. PurpleAir, Inc. PurpleAir API documentation. 2023. Available from: https://api.purpleair.com (accessed on 2026-07-30).
26. U.S. Environmental Protection Agency. EJScreen: environmental justice screening and mapping tool, version 2.3. 2024. Available from: https://www.epa.gov/ejscreen (accessed on 2025-09-01; dataset archived by the Public Environmental Data Project after EJScreen's removal from the EPA website in February 2025).

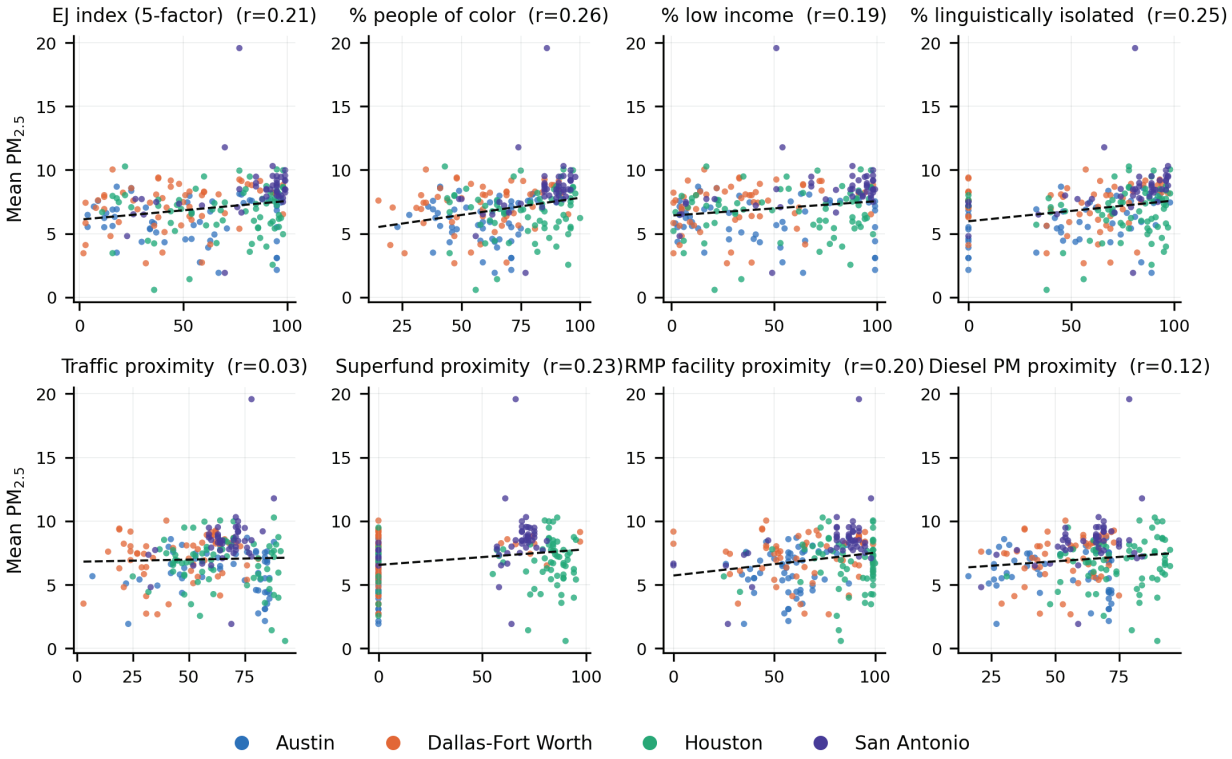
SUPPLEMENTARY MATERIAL



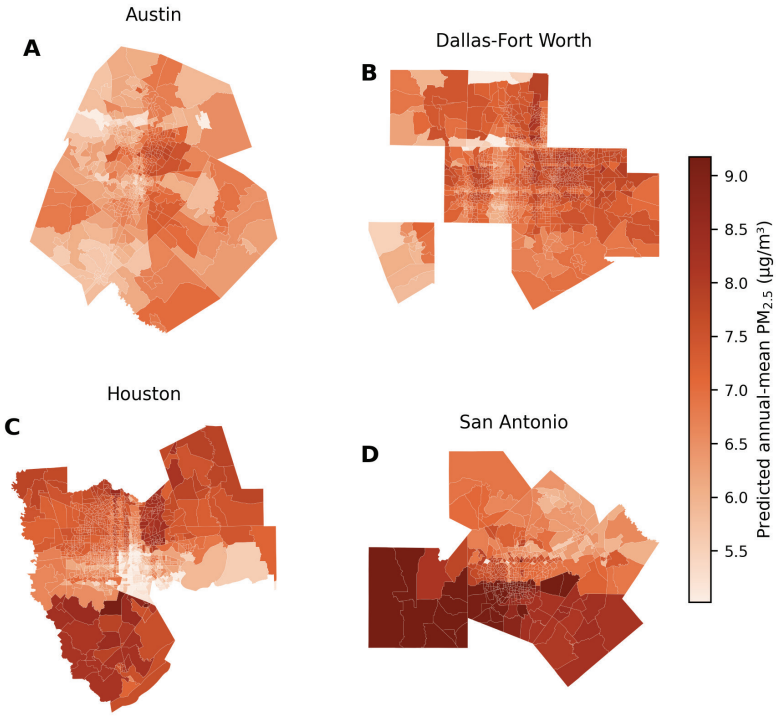
Supplementary Figure S1. *Distribution of quality-assured daily PM2.5 by metropolitan area, with metro means marked. No regulatory threshold is shown: the EPA 9 µg/m³ standard applies to annual means, not to the daily observations displayed here.*



Supplementary Figure S2. *Correlation among daily PM2.5, meteorological predictors, and the EJScreen audit variables. The EJScreen indicators are mutually strongly correlated, which motivates interpreting the family jointly in the audit.*



Supplementary Figure S3. *Sensor-level observed mean PM2.5 against each EJScreen variable across the 226 sensors, the sensor-scale counterpart of the tract-level audit.*



Supplementary Figure S4. *Predicted annual-mean PM2.5 surface by census tract. (A) Austin. (B) Dallas–Fort Worth. (C) Houston. (D) San Antonio. All panels share one color scale.*