Supporting Information for

Intercomparison of six national empirical models for PM$_{2.5}$ air pollution in the contiguous US

Matthew J. Bechle$^1$, Michelle L. Bell$^2$, Daniel L. Goldberg$^3$, Steve Hankey$^4$, Tianjun Lu$^5$, Albert A. Presto$^6$, Allen L. Robinson$^6$, Joel Schwartz$^7$, Liuhua Shi$^{7,8}$, Yang Zhang$^9$, Julian D. Marshall$^{1,*}$

$^1$ Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, USA
$^2$ Yale School of Forestry & Environmental Studies, Yale University, New Haven, CT, USA
$^3$ Department of Environmental and Occupational Health, George Washington University, Washington, DC, USA
$^4$ School of Public and International Affairs, Virginia Tech, Blacksburg, VA, USA
$^5$ Department of Earth Science and Geography, California State University, Dominguez Hills, Carson, CA, USA
$^6$ Department of Mechanical Engineering and Center for Atmospheric Particle Studies, Carnegie Mellon University, Pittsburgh, PA, USA
$^7$ Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA
$^8$ Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA
$^9$ Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115, USA
$^*$ Corresponding author. Email: jdmash@uw.edu.

Correspondence to: jdmash@uw.edu

The supporting information include:
1. Introduction and background
2. Additional considerations regarding the methods, including strengths and weaknesses of the paper
3. Sensitivity Analyses, including Figures S1 to S9
4. References
Introduction and background

Empirical models are used to understand and predict levels of outdoor air pollution, including at unmeasured locations and typically with good spatial precision. The name (“empirical”) emphasizes that the models reflect empirical measurements. Such model results have been used, for example, in health risk assessment, environmental epidemiology, and environmental justice analysis.

Generating empirical-model results typically involves three steps: (1) Model building: generating an empirical model to predict measured concentrations (i.e., the dependent variable; the model is calibrated to and attempts to predict these), using several parameters that might correlate with concentrations (i.e., the potential independent variables). (2) Model testing, to quantify parameters such as uncertainty, robustness, error, and bias. If multiple models were built by a research group, the model-testing phase could involve a model-selection process. Hold-out cross-validation typically occurs in this step. (3) Model application, wherein the final selected model(s) is used to estimate concentrations throughout the domain of interest (e.g., at all Census Block centroids in the continuous US).

Early empirical models were developed at the urban-scale, using land-use variables (e.g., road locations, industrial locations) and linear regression, and hence were called “land-use regression” (LUR).\(^1\)–\(^6\)

Subsequent developments include (1) adding many more independent variables, including microscale point-of-interest sources,\(^7\),\(^8\) satellite-derived estimates for pollution (e.g., atmospheric column totals) or land-cover,\(^9\)–\(^11\) and predictions from chemical transport models,\(^10\),\(^12\) (2) deriving independent variables from imagery (Google Street View images or satellite images) or using images directly via machine learning rather than first obtaining specific independent variables,\(^13\)–\(^18\) (3) employing more-advanced mathematics rather than linear regression,\(^19\)–\(^23\) (4) quantifying temporal variability,\(^24\),\(^25\) and (5) using a national or international, rather than urban, spatial domain.\(^26\)–\(^31\) For the dependent variable, early models often employed purposefully-placed passive NO\(_2\) samplers,\(^32\)–\(^34\) subsequent developments include using regulatory monitoring data,\(^27\),\(^28\),\(^30\) mobile monitoring,\(^20\),\(^35\)–\(^40\) and freely-available data from ubiquitous low-cost sensors already deployed by the public.\(^41\),\(^42\) Strengths of empirical models include excellent spatial
precision in model-predictions, and being based in observations and datasets (e.g., land-use datasets; satellite observations); weaknesses include the long list of input data required (which is unavailable for the future and for the distant past), inability to allow “what if?” questions (e.g., “what if emissions were reduced?”), and potential error/bias in the measurements and in the models. Empirical models reflect typical correlations with specific land uses, so they are weaker at detecting atypical conditions (e.g., an unexpected hotspot) or conditions not well correlated with a land-use in the model.

Studies to intercompare empirical models are scarce, especially for large geographies. Some studies have compared empirical models with mechanistic models (e.g., CMAQ), satellite-based models (e.g., aerosol optical depth, AOD), or hybrid models. Other studies have compared results using different methods for model-building (e.g., LUR vs. machine learning vs. kriging vs. hybrid empirical models). However, most prior comparisons were at the city or region level, and comparisons were generally within a single research team. We identified only one study that compared empirical models nationwide.

### Additional considerations regarding the methods

Our approach is to intercompare a sample of six national empirical models for annual-average ambient PM2.5. We focused on annual-averages for fine particles (PM2.5) for several reasons: PM2.5 is an important criteria-pollutant, regulated by the US EPA through a health-based National Ambient Air Quality Standard (NAAQS); millions of people in the US live in areas that exceed the NAAQS (US EPA, 2022a); and the health effects associated with annual-average PM2.5 are large. Importantly, multiple national empirical models predict annual-average PM2.5 available for this pollutant.

In general, one way to intercompare models would be for all modelers to pre-agree to a set of model-building and model-testing observations. (Or, if there were a set of measurements that no model included in model-building — e.g., a dataset that was unknown or otherwise unused — then the outcome would be
similar: a dataset that could be used to test all of the models.) In this case, it would be possible to compare each method against the held-out cross-validation measurements. However, in the current intercomparison, each research group used their own held-out data, comparison metrics, and approach to investigate model uncertainty. Furthermore, the models incorporate the monitoring data in different ways (e.g., via a kriging component); for that reason, simply comparing the six models against observations (which were used during model-building) may not shed light on model reliability at locations without measurements.

Instead, we directly intercompare the models, without comparing against held-out measurements. We do not have “gold-standard” observations to compare against. Nevertheless, we believe that useful insights can be gained from the intercomparisons conducted.

Limitations of this research include the following. (1) We considered one set of spatiotemporal comparisons (annual-average; national/regional/urban-rural) and one set of metrics (RMSE, correlation), but did not compare all possible comparisons (e.g., did not investigate seasonal or daily models, nor sub-regional or local/community model results) or metrics. Other metrics or spatiotemporal representations of the models too may be useful for health, environmental justice, or risk analysis. (2) We have not specifically investigated the fitness of these models for specific purposes, including epidemiological studies, environmental justice studies, public outreach, regulatory analysis, or risk assessment. (3) As mentioned above, we did not compare against measurements; this paper presents only a model-model comparison. Model-model agreement is not the same as a model being “correct”. (4) We have identified that the empirical models are relatively consistent with each other, but we have not investigated, within the models themselves, why. For example, it may be that the models use the same or similar independent variables; or, it may be that the similarities in model-prediction are despite large differences in independent variables employed.

Strengths of this research include the following. We inter-compared several models, and shed light on similarities and differences nationally, regionally, for urban/rural differences, by pollution level, and by
population density. This is, to our knowledge, the first intercomparison of national empirical models. As noted above, we did not compare against monitors; however, that aspect can partially be viewed as a strength, because the monitoring network is not evenly distributed spatially. Comparisons of models at monitor locations may or may not shed light on concentrations at unmonitored locations; the comparisons here are at Census geographics (Tracts) and so reflect locations where people live. Although PM$_{2.5}$ is generally considered to be a regional pollutant, when considering the whole US there is a wide range of concentrations (here: ~2 to ~15 $\mu$g/m$^3$, or approximately a 7-fold range).

The models employ different techniques for model building. Some are closer to a linear model, some use machine learning or highly complex mathematical relationships that would be difficult for a human to create or understand. They employ a wide variety of independent variables. However, all of the models use EPA monitoring station data as the model-building dataset. Whatever strengths or weaknesses exist in using EPA monitors (and their locations) for empirical models, those likely impact all of the models.

**Sensitivity analyses**

We conducted several sensitivity analyses. First, reflecting that SEARCH results are only available in the eastern half of the US, we generated pairwise scatterplots for only the eastern half of the US (Figure S1). Next, we generated separate scatterplots for urban-only (Figure S2) and urban-only in the eastern half of the US (Figure S3) and for rural-only (Figure S4) and for rural-only in the eastern half of the US (Figure S5). We find, for example, that the maximum RMSD is slightly larger for rural areas than for urban areas, a finding that may differ from expectations but is consistent with results described in the main paper (Figure 3E) and may reflect the lower density of monitors in rural areas or that the correlation between concentrations and land use may be lower in rural than in urban areas.

We repeated the analyses in Figure 3 but for the eastern half of the US (Figure S6 and S7). The findings are generally consistent with results in the main paper: the models generally agree with each other. The
range of predictions (a measure of model-model disagreement) is greater at lower-concentration locations than at high-concentration locations.
Figure S1: Scatterplot matrix for 2010 tract-level PM$_{2.5}$ (E-US only). Same as Fig 1, but for locations in E-US only. Scatterplots in the upper right show pairwise tract-level predictions from each model. Dashed line shows 1:1 line, red line shows linear trendline. Corresponding boxes in the bottom left show Pearson’s correlation ($r$; unitless) and root mean squared difference (RMSD; µg/m$^3$) between model predictions.
Figure S2: Scatterplot matrix for 2010 urban tract-level PM$_{2.5}$. Same as Fig 1, but for urban locations only. Scatterplots in the upper right show pairwise tract-level predictions from each model. Dashed line shows 1:1 line, red line shows linear trendline. Corresponding boxes in the bottom left show Pearson’s correlation ($r$; unitless) and root mean squared difference (RMSD; µg/m$^3$) between model predictions.
Figure S3: Scatterplot matrix for 2010 urban tract-level PM$_{2.5}$ (E-US only). Same as Fig 1, but for urban locations in E-US only. Scatterplots in the upper right show pairwise tract-level predictions from each model. Dashed line shows 1:1 line, red line shows linear trendline. Corresponding boxes in the bottom left show Pearson’s correlation ($r$; unitless) and root mean squared difference (RMSD; µg/m$^3$) between model predictions.
Figure S4: Scatterplot matrix for 2010 rural tract-level PM$_{2.5}$. Same as Fig 1, but for rural locations only. Scatterplots in the upper right show pairwise tract-level predictions from each model. Dashed line shows 1:1 line, red line shows linear trendline. Corresponding boxes in the bottom left show Pearson’s correlation ($r$; unitless) and root mean squared difference (RMSD; µg/m$^3$) between model predictions.
Figure S5: Scatterplot matrix for 2010 rural tract-level \( \text{PM}_{2.5} \) (E-US only). Same as Fig 1, but for rural locations in E-US only. Scatterplots in the upper right show pairwise tract-level predictions from each model. Dashed line shows 1:1 line, red line shows linear trendline. Corresponding boxes in the bottom left show Pearson’s correlation \((r; \text{unitless})\) and root mean squared difference \((\text{RMSD}; \mu g/m^3)\) between model predictions.
Figure S6: Predictions by model and concentration level (E-US only). Boxplots show range of tract-level model predictions (y-axis) as a function of the median concentration among model predictions within each tract, binned to 1 µg/m³ bins (x-axis). Horizontal bar shows the median, box shows the interquartile range, and vertical lines show the 5th and 95th percentiles of predicted concentrations for each model within each bin.

Figure S7: Variability by concentration level (E-US only). Boxplots show within-tract variation (y-axis) as a function of the median concentration among model predictions within each tract, binned to 1 µg/m³ bins (x-axis). Within-tract variation is calculated as either max minus the min (left) or 2nd max minus 2nd min (right) of model predictions within each tract. Horizontal bar shows the median, box shows the interquartile range, and vertical lines show the 5th and 95th percentiles of the variability for tracts within each bin.
Figure S8: Predictions by model and population density. Boxplots show range of tract-level model predictions (y-axis) as a function of the tract-level population density (x-axis), binned to equally space linear- (left) and log-space (right) bins (note: last bin contains remaining data). Horizontal bar shows the median, box shows the interquartile range, and vertical lines show the 5th and 95th percentiles of predicted concentrations for each model within each bin.

Figure S9: Predictions by model and population density (E-US only). Boxplots show range of tract-level model predictions (y-axis) as a function of the tract-level population density (x-axis), binned to equally space linear- (left) and log-space (right) bins (note: last bin contains remaining data). Horizontal bar shows the median, box shows the interquartile range, and vertical lines show the 5th and 95th percentiles of predicted concentrations for each model within each bin.
References


Page 16 of 18


