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

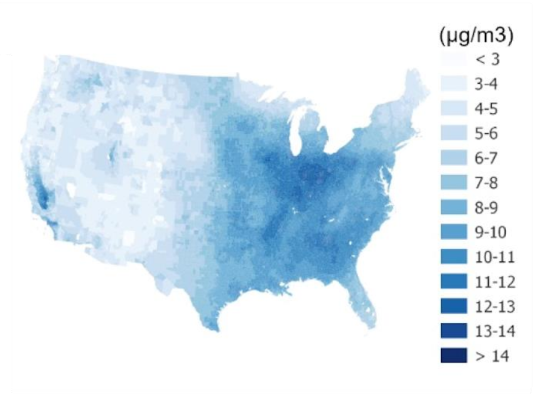
[8] Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA

[9] Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115, USA

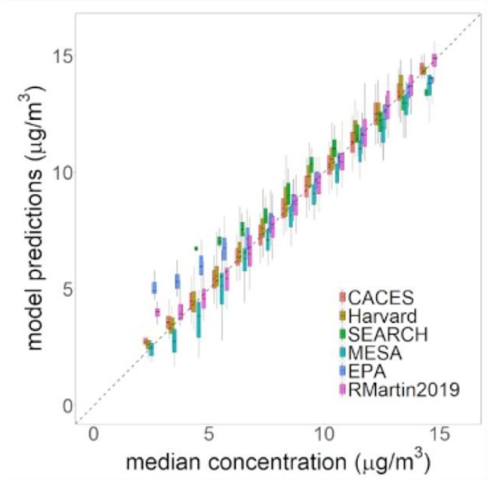
[*] Corresponding author. Email: jdmarsh@uw.edu.

TOC Art

Six research groups generated national empirical-models for $PM_{2.5}$, using different methods.



Predictions from the models show relative agreement.



Abstract

Empirical models, previously called land-use regression (LUR), are used to understand and predict spatial variability in levels of outdoor air pollution at unmeasured locations, for example, to conduct health risk assessment, environmental epidemiology, or environmental justice analysis. Many methods are used to generate empirical models, yet almost no research compares models generated by separate research groups. We intercompare six national-scale empirical models for year-2010 concentrations of $PM_{2.5}$ in the US, each generated by a different research group. Despite substantial differences in the statistical methods and input data used to build the models, our main finding is a relatively high degree of agreement among model predictions. For example, in pairwise intercomparisons, the average Pearson correlation coefficient is 0.87 (range: 0.84 to 0.92); the RMSD (root-mean-square-difference; units: μg/m^3) is 1.1 on average (range: 0.8 to 1.4), or ~12% of the average concentration; and many best-fit lines are near the 1:1 line. The underlying reason for this agreement is likely that, while the methods and the independent variables differ among the models, in all cases the models are built using, and are calibrated to, the same information: publicly available measurement at US EPA regulatory monitoring stations. Findings here suggest that future improvements to national empirical models will come not from further refinements to the methods (e.g., more-advanced models) but from employing a fundamentally different set of observations, in addition to regulatory monitoring data.

Key Words

Land use regression, exposure assessment, air quality models, empirical model comparison, point-based models, gridded models

Synopsis

Model predictions of six national-scale empirical models in the contiguous US have a high degree of agreement.

1. Introduction

Empirical models can be used to understand and predict levels of outdoor air pollution, including at unmeasured locations. 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) (Brauer et al, 2003; Jerrett et al., 2005; Hoek et al., 2008; Marshall et al., 2008; Su et al., 2008; Eeftens et al., 2012). Subsequent developments include (1) adding many more independent variables, including microscale point-of-interest sources (Wu et al., 2017; Lu et al., 2019), satellite-derived estimates for pollution (e.g., atmospheric column totals) or land-cover (Knibbs et al., 2018; Bechle et al., 2015; de Hoogh et al., 2016), and predictions from chemical transport models (Bechle et al., 2015; Goldberg et al., 2019), (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 (Hong et al., 2019; Weichenthal et al., 2019; Ganji et al., 2020; Lu et al., 2021; Qi & Hankey, 2021; Qi et al., 2022), (3) employing more-advanced mathematics rather than linear regression (Beckerman et al., 2013; Weichenthal et al., 2016; Di et al., 2019; Lautenschlager et al., 2020; Wong et al., 2021), (4) quantifying temporal variability (Wu et al., 2018; Masiol et al., 2019), and (5) using a national or international, rather than urban, spatial domain (Hoek et al., 2015; Hystad et al., 2011; Novotny et al., 2011; Knibbs et al., 2014; Larkin et al., 2017; Saha et al., 2021). For the dependent variable, early models often employed purposefully-placed passive NO_2 samplers (Aguilera et al., 2008; Wang et al., 2013; Lee et al., 2017); subsequent developments include using regulatory monitoring data (Hystad et al., 2011; Novotny et al., 2011; Larkin et al., 2017), mobile monitoring (Larsen et al., 2007; Thai et al., 2008; Hankey & Marshall, 2015; Weichenthal et al., 2016; Messier et al., 2018; Minet et al., 2018; Hankey et al., 2019), and freely-available data from ubiquitous low-cost sensors already deployed by the public (Bi et al., 2020; Lu et al., 2022).

Studies to intercompare empirical models are scarce, especially for large geographies. Some studies have compared empirical models with mechanistic models (e.g., CMAQ) (Marshall et al., 2008; Samoli et al., 2020), satellite-based models (e.g., aerosol optical depth, AOD) (Yu et al., 2018; Cowie et al., 2019), or hybrid models (Michanowicz et al., 2016; Zhang et al., 2021). Other studies have compared results using different methods for model-building (e.g., LUR vs. machine learning vs. kriging vs. hybrid empirical models) (Adam-Poupart et al., 2014; Jain et al., 2021; Dharmalingam et al., 2022). 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 (Lu et al., 2021).

This paper adds to the literature by comparing concentration predictions from six annual-average PM2.5 empirical models for the contiguous US. Each model was generated by a different research group; they differ in their approaches. Our analysis compares the predictions from these models at three spatial scales: nationally, regionally, and urban/rural.

2. Methods

2.1. General

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.

2.2 Input data

We obtained year-2010 predicted PM$_{2.5}$ concentrations for six empirical models (see Table 1) via data download or direct request from researchers. Three models are “point based” (their predictions apply to a specific spatial location): (1) The CACES EPA ACE center model is based on universal kriging with partial least squares data-reduction (PLS-UK) (Kim et al., 2020). (2) The EPA downscaler provides a

Bayesian space-time "fuse" of monitoring data and 12 km CMAQ model outputs (US EPA, 2022b). (3) The MESA-Air models use space-time PLS and expectation-maximization to fill in missing observations (Keller et al., 2015). The other three models are "gridded models" (predictions are for grid locations, reflecting the average concentration in that region [e.g., in a ~ 1 km^2 area]): (4) The Harvard/MIT EPA ACE center model employs a generalized additive model to integrate multiple machine-learning algorithms (Di et al., 2019). (5) The SEARCH EPA ACE center model is based on a fusion of WRF-Chem, satellite data (MAIAC AOD), and a kriging of EPA monitor data (Goldberg et al., 2019). (6) The model by van Donkelaar et al. (2019) statistically "fuses" a chemical transport model (GEOS-Chem), satellite observations of aerosol optical depth, and ground-based observations using a geographically weighted regression.

To make direct intercomparisons, we aligned spatiotemporal aspects of the models to be annual-average and by Census Tract (n ~ 74,000). When sub-annual (e.g., monthly) predictions were provided, we calculated annual averages. When sub-tract (e.g., block) predictions were provided, we calculated Tract means. When predictions were gridded, we converted to Census geographies by extracting values at block locations and then population weighting to the tract level.

One of the models (SEARCH model) is only available for the eastern half of the contiguous US (90° W longitude), which includes US cities as far west as Chicago. The other five models are available for the entire contiguous US.

2.3 Analysis

We conducted three pairwise comparisons of the model-predictions: (1) scatterplot matrices, (2) Pearson's *r*, and (3) root mean square difference (RMSD) between predictions. We also generated boxplots showing distribution of predictions, and calculated the two values in each tract to indicate the range of model predictions: range (i.e., max minus min), and trimmed range (second-highest value minus second-lowest value).

To assess factors that may modify model agreement, we conducted comparisons for the following geographies: (1) all locations, (2) urban vs. rural (urban defined as all tracts intersecting with Census urbanized areas, all remaining tracts are considered rural), (3) by region (using the 9 NOAA climate regions), and (4) stratified by population density (using the 2010 tract-level population density).

3. Results and discussion

Pairwise scatterplots of model predictions (Figure 1) indicate a relatively high degree of agreement. The average Pearson correlation coefficient ("*r*") is 0.87 (range: 0.84 to 0.92), RMSD (units: μg/m^3) is 1.1 on average (range: 0.8 to 1.4), and many best-fit lines are near the 1:1 line. The population average concentration of $PM_{2.5}$ in 2010 was ~9.3 μg/m^3 (mean), ~9.5 μg/m^3 (median), so the RMSD (1.1 μg/m^3) represents ~12% of the average concentration. Thus, at the national level, the models agree well.

Model-model comparisons by geography (Figure 2) suggests modest differences among most regions, and minor differences between urban/rural locations. Pearson correlation coefficients indicate that model-model agreement is slightly lower in the Midwest and South than in other regions. RMSDs indicate agreement is slightly lower in the West.

Figure 3 shows the prediction variability by concentration and location. Two aspects stand out: first, the relative agreement among the models, across the range of concentrations (Figure 3D). In locations for which the median predicted concentration is comparatively low (less than 6 μg/m^3), EPA predictions tend to be slightly higher than the other models. For the very lowest-concentration locations, with median predicted concentrations less than 3 μg/m^3, the Martin2019 predictions too tend to be slightly larger than the other models. The SEARCH model is only available for the eastern half of the contiguous US and so therefore excludes lower-population-density lower-concentration regions found in the western half of the contiguous US. The CACES and Harvard models tend to agree with each other and to be near the median prediction, for each concentration range (Figure 3D).

Second, the range of model predictions (a measure of between-model disagreement) exhibits a potentially surprising relationship with concentration level (see Figure 3E, 3F). We might have expected that the range of predictions would be wider for higher-concentration locations (e.g., consistent with the models having a certain percent-error in their predictions). Instead, the range of model predictions is approximately constant across levels of pollution (Figure 3E, 3F). This is consistent with an additive rather than multiplicative error model. To the extent that there is a pattern (more so for Figure 3E than Figure 3F), the range of predictions is greater in lower- than in higher-concentration locations. The finding reflects the patterns mentioned in the previous paragraph: below 5 or 6 μg/m^3, the EPA predictions (and, below 3 μg/m^3, the Martin2019 predictions too) are larger than the other models' predictions; it suggests that predicting concentrations in low-concentration locations might be more challenging (greater model-model difference) than in medium- or high-concentration locations.

Overall, while some model-model differences are revealed in Figure 3, the main finding is relative agreement. Model-model comparisons can identify the level of model agreement/disagreement, but not of accuracy or error. In cases where the models agree (or disagree), it's possible all of the models are incorrect. Thus, a useful step for future research would be to compare against held-out measurements --- either via a coordinated effort by the researchers to hold out a consistent set of measurements, or via an independent dataset of concentrations that none of the researchers employed in model-building.

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.

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.

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 above (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 above: 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.

Acknowledgements

We gratefully acknowledge the funders. This publication was developed as part of the Center for Air, Climate, and Energy Solutions (CACES), which was supported under Assistance Agreement No. R835873 awarded by the U.S. Environmental Protection Agency (EPA). Additional funding from the EPA for the SEARCH Center (RD83587101) and the Harvard-MIT ACE center (RD83479801). This manuscript has not been formally reviewed by EPA. The views expressed here are solely those of authors and do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this publication.

References

Adam-Poupart, A., A Brand, M Fournier, M Jerrett, A Smargiassi (2014). Spatiotemporal modeling of ozone levels in Quebec (Canada): a comparison of kriging, land-use regression (LUR), and combined Bayesian maximum entropy–LUR approaches. *Environmental Health Perspectives*, *122*(9), 970-976.

Aguilera, I., J Sunyer, R Fernández-Patier, G Hoek, A Aguirre-Alfaro, K Meliefste, K., MT Bomboi-Mingarro, MJ Nieuwenhuijsen, D Herce-Garraleta, B Brunekreef (2008). Estimation of outdoor NOx, NO2, and BTEX exposure in a cohort of pregnant women using land use regression modeling. *Environmental Science & Technology*, *42*(3), 815-821.

Bechle, M.J., DB Millet, JD Marshall (2015). National spatiotemporal exposure surface for NO2: monthly scaling of a satellite-derived land-use regression, 2000–2010. *Environmental Science & Technology*, *49*(20), 12297-12305.

Beckerman, B.S., M Jerrett, M Serre, RV Martin, S-J Lee, A Van Donkelaar, Z Ross, J Su, RT Burnett (2013). A hybrid approach to estimating national scale spatiotemporal variability of PM2.5 in the contiguous United States. *Environmental Science & Technology*, *47*(13), 7233-7241.

Bi, J., A Wildani, HH Chang, Y Liu (2020). Incorporating low-cost sensor measurements into high-resolution PM2.5 modeling at a large spatial scale. *Environmental Science & Technology*, *54*(4), 2152-2162.

Brauer, M., G Hoek, P van Vliet, K Meliefste, P Fischer, U Gehring, J Heinrich, J Cyrus, T Bellander, M Leone, B Brunekreef (2003). Estimating long-term average particulate air pollution concentrations: Application of traffic indicators and geographic information systems. *Epidemiology,* 14 (2). 228-239.

Cowie, C.T., F Garden, E Jegasothy, LD Knibbs, I Hanigan, D Morley, A Hansell, G Hoek, GB Marks (2019). Comparison of model estimates from an intra-city land use regression model with a national satellite-LUR and a regional Bayesian Maximum Entropy model, in estimating NO2 for a birth cohort in Sydney, Australia. *Environmental Research*, *174*(March), 24–34.

de Hoogh, K., J Gulliver, A van Donkelaar, RV Martin, JD Marshall, MJ Bechle, ... & Hoek, G. (2016). Development of West-European PM2.5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data. *Environmental Research*, *151*, 1-10.

Dharmalingam, S., N Senthilkumar, RR D'Souza, Y Hu, HH Chang, S Ebelt, H Yu, CS Kim, A Rohr (2022). Developing air pollution concentration fields for health studies using multiple methods: Cross-comparison and evaluation. *Environmental Research*, *207*(September 2021), 112207.

Di, Q., H Amini, L Shi, I Kloog, R Silvern, J Kelly, …, J Schwartz (2019). An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment International*, *130*, 104909.

Di, Q., H Amini, L Shi, I Kloog, R Silvern, J Kelly, …, J Schwartz (2019). Assessing NO2 concentration and model uncertainty with high spatiotemporal resolution across the contiguous United States using ensemble model averaging. *Environmental Science & Technology*, *54*(3), 1372-1384.

Eeftens, M., R Beelen, K de Hoogh, T Bellander, G Cesaroni, M Cirach, …, G Hoek (2012). Development of land use regression models for PM2.5, PM2.5 absorbance, PM10 and PMcoarse in 20 European study areas; results of the ESCAPE project. *Environmental Science & Technology*, *46*(20), 11195-11205.

EPA Downscaler Model for predicting daily air pollution. https://www.epa.gov/air-research/downscaler-model-predicting-daily-air-pollution

US EPA (2022a). *Nonattainment Area Summary with History (Green Book).* Retrieved November 30, 2022, from https://www3.epa.gov/airquality/greenbook/knsum2.html

US EPA (2022b). Fused Air Quality Surface Using Downscaling (FAQSD) Files (Updated: October 24, 2022). https://www.epa.gov/hesc/rsig-related-downloadable-data-files.

Ganji, A., L Minet, S Weichenthal, M Hatzopoulou (2020). Predicting traffic-related air pollution using feature extraction from built environment images. *Environmental Science & Technology*, *54*(17), 10688-10699.

Goldberg, D.L., P Gupta, K Wang, C Jena, Y Zhang, Z Lu, DG Streets (2019). Using gap-filled MAIAC AOD and WRF-Chem to estimate daily PM2.5 concentrations at 1 km resolution in the Eastern United States. *Atmospheric Environment*, *199*, 443-452.

Hankey, S., JD Marshall (2015). Land use regression models of on-road particulate air pollution (particle number, black carbon, PM2.5, particle size) using mobile monitoring. *Environmental Science & Technology*, *49*(15), 9194-9202.

Hankey, S., P Sforza, M Pierson (2019). Using mobile monitoring to develop hourly empirical models of particulate air pollution in a rural Appalachian community. *Environmental Science & Technology*, *53*(8), 4305-4315.

Hoek, G., R Beelen, K de Hoogh, D Vienneau, J Gulliver, P Fischer, D Briggs (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric environment*, *42*(33), 7561-7578.

Hoek, G., M Eeftens, R Beelen, P Fischer, B Brunekreef, KF Boersma, P Veefkind (2015). Satellite NO2 data improve national land use regression models for ambient NO2 in a small densely populated country. *Atmospheric Environment*, *105*, 173-180.

Hong, K.Y., PO Pinheiro, L Minet, M Hatzopoulou, S Weichenthal (2019). Extending the spatial scale of land use regression models for ambient ultrafine particles using satellite images and deep convolutional neural networks. *Environmental Research*, 176, 108513.

Hystad, P., S Eleanor, A Cervantes, K Poplawski, S Deschenes, M Brauer, A van Donkelaar, L Lamsal, R Martin, M Jerrett, P Demers (2011). Creating National Air Pollution Models for Population Exposure Assessment in Canada. *Environmental Health Perspectives*, 119(8), 1123-1129.

Jain, S., AA Presto, N Zimmerman (2021). Spatial modeling of daily PM2.5, NO2, and CO concentrations measured by a low-cost sensor network: Comparison of linear, machine learning, and hybrid land use models. *Environmental Science & Technology*, *55*(13), 8631–8641.

Jerrett, M., A Arain, P Kanaroglou, B Beckerman, D Potoglou, T Sahsuvaroglu, J Morrison, C Giovis (2005). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Science & Environmental Epidemiology*, *15*(2), 185-204.

Keller, J.P., C Olives, SY Kim, L Sheppard, PD Sampson, AA Szpiro, AP Oron, J Lindström, S Vedal, JD Kaufman (2015). A unified spatiotemporal modeling approach for predicting concentrations of multiple air pollutants in the Multi-Ethnic Study of Atherosclerosis and Air Pollution. *Environmental Health Perspectives*, 123, 301–309.

Kim, S.Y., MJ Bechle, S Hankey, L Sheppard, AA Szpiro, JD Marshall (2020). Concentrations of criteria pollutants in the contiguous US, 1979–2015: role of prediction model parsimony in integrated empirical geographic regression. *PLoS ONE*, *15*(2), e0228535.

Knibbs, L.D., MG Hewson, MJ Bechle, JD Marshall, AG Barnett (2014). A national satellite-based land-use regression model for air pollution exposure assessment in Australia. *Environmental Research*, *135*, 204-211.

Knibbs, L.D., A van Donkelaar, RV Martin, MJ Bechle, M Brauer, DD Cohen, …, AG Barnett (2018). Satellite-based land-use regression for continental-scale long-term ambient PM2.5 exposure assessment in Australia. *Environmental Science & Technology*, *52*(21), 12445-12455.

Larkin, A., JA Geddes, RV Martin, Q Xiao, Y Liu, JD Marshall, M Brauer, P Hystad (2017). Global Land Use Regression Model for Nitrogen Dioxide Air Pollution. *Environmental Science & Technology*, 51(12), 6957–6964.

Larson, T., J Su, AM Baribeau, M Buzzelli, E Setton, M Brauer (2007). A spatial model of urban winter woodsmoke concentrations. *Environmental Science & Technology*, *41*(7), 2429-2436.

Lautenschlager, F., M Becker, K Kobs, M Steininger, P Davidson, A Krause, A Hotho (2020). OpenLUR: Off-the-shelf air pollution modeling with open features and machine learning. *Atmospheric Environment*, *233*, 117535.

Lee, M., M Brauer, P Wong, R Tang, TH Tsui, C Choi, …, B Barratt (2017). Land use regression modelling of air pollution in high density high rise cities: A case study in Hong Kong. *Science of the Total Environment*, *592*, 306-315.

Lu, T., J Lansing, W Zhang, MJ Bechle, S Hankey (2019). Land Use Regression models for 60 volatile organic compounds: Comparing Google Point of Interest (POI) and city permit data. *Science of the Total Environment*, *677*, 131-141.

Lu, T., MJ Bechle, Y Wan, AA Presto, S Hankey (2022). Using crowd-sourced low-cost sensors in a land use regression of PM2.5 in 6 US cities. *Air Quality, Atmosphere & Health*, 15(4), 667-678.

Lu, T., JD Marshall, W Zhang, P Hystad, S Kim, MJ Bechle, M Demuzere, S Hankey (2021). National empirical models of air pollution using microscale measures of the urban environment. *Environmental Science and Technology*, *55*, 15519–15530.

Marshall, J.D., E Nethery, M Brauer (2008). Within-urban variability in ambient air pollution: Comparison of estimation methods. *Atmospheric Environment*, *42*(6), 1359–1369.

Masiol, M., S Squizzato, D Chalupa, DQ Rich, PK Hopke (2019). Spatial-temporal variations of summertime ozone concentrations across a metropolitan area using a network of low-cost monitors to develop 24 hourly land-use regression models. *Science of The Total Environment*, *654*, 1167-1178.

Messier, K.P., SE Chambliss, S Gani, R Alvarez, M Brauer, JJ Choi, SP Hamburg, J Kerckhoffs, B LaFranchi, MM Lunden, JD Marshall, CJ Portier, A Roy, AA Szpiro, RCH Vermeulen, JS Apte (2018). Mapping Air Pollution with Google Street View Cars: Efficient Approaches with Mobile Monitoring and Land Use Regression. *Environmental Science & Technology*, 52(21), 12563–12572.

Michanowicz, D.R., JLC Shmool, BJ Tunno, S Tripathy, S Gillooly, E Kinnee, JE Clougherty (2016). A hybrid land use regression/AERMOD model for predicting intra-urban variation in PM2.5. *Atmospheric Environment*, *131*, 307–315.

Minet, L., R Liu, M-F Valois, J Xu, S Weichenthal, M Hatzopoulou (2018). Development and comparison of air pollution exposure surfaces derived from on-road mobile monitoring and short-term stationary sidewalk measurements. *Environmental Science & Technology*, 52(6), 3512-3519.

Novotny, E.V., MJ Bechle, DB Millet, JD Marshall (2011). National satellite-based land-use regression: NO2 in the United States. *Environmental Science & Technology*, *45*(10), 4407-4414.

Qi, M., K Dixit, JD Marshall, W Zhang, S Hankey (2022). National land use regression model for NO2 using street view imagery and satellite observations. *Environmental Science & Technology*, *56*(18), 13499-13509.

Qi, M., S Hankey (2021). Using street view imagery to predict street-level particulate air pollution. *Environmental Science & Technology*, *55*(4), 2695-2704.

Requia, W.J., Q Di, R Silvern, JT Kelly, P Koutrakis, LJ Mickley, ..., J Schwartz (2020). An ensemble learning approach for estimating high spatiotemporal resolution of ground-level ozone in the contiguous United States. *Environmental Science & Technology*, *54*(18), 11037-11047.

Saha, P.K., S Hankey, JD Marshall, AL Robinson, AA Presto (2021). High-Spatial-Resolution Estimates of Ultrafine Particle Concentrations across the Continental United States. *Environmental Science & Technology*, 55(15), 10320–10331.

Samoli, E., BK Butland, S Rodopoulou, RW Atkinson, B Barratt, SD Beevers, A Beddows, K Dimakopoulou, JD Schwartz, MD Yazdi, K Katsouyanni (2020). The impact of measurement error in modeled ambient particles exposures on health effect estimates in multilevel analysis: A simulation study. *Environmental Epidemiology*, *4*(3).

Sampson, P.D., M Richards, AA Szpiro, S Bergen, L Sheppard, TV Larson, JD Kaufman (2013). A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2.5 concentrations in epidemiology. *Atmospheric Environment*, *75*, 383-392.

Su, J.G., M Buzzelli, M Brauer, T Gould, TV Larson (2008). Modeling spatial variability of airborne levoglucosan in Seattle, Washington. Atmospheric Environment 42 (22), 5519-5525.

Thai, A., I McKendry, M Brauer (2008). Particulate matter exposure along designated bicycle routes in Vancouver, British Columbia. *Science of the Total Environment*, *405*(1-3), 26-35.

Van Donkelaar, A., RV Martin, C Li, RT Burnett (2019). Regional estimates of chemical composition of fine particulate matter using a combined geoscience-statistical method with information from satellites, models, and monitors. *Environmental Science & Technology*, *53*(5), 2595-2611.

Wang, M., R Beelen, X Basagana, T Becker, G Cesaroni, K de Hoogh, …, B Brunekreef (2013). Evaluation of land use regression models for NO2 and particulate matter in 20 European study areas: the ESCAPE project. *Environmental Science & Technology*, *47*(9), 4357-4364.

Weichenthal, S., K Van Ryswyk, A Goldstein, M Shekarrizfard, M Hatzopoulou (2016). Characterizing the spatial distribution of ambient ultrafine particles in Toronto, Canada: A land use regression model. *Environmental Pollution*, *208*, 241-248.

Weichenthal, S., K Van Ryswyk, A Goldstein, S Bagg, M Shekkarizfard, M Hatzopoulou (2016). A land use regression model for ambient ultrafine particles in Montreal, Canada: A comparison of linear regression and a machine learning approach. *Environmental Research*, 146, 65-72

Weichenthal, S., M Hatzopoulou, M Brauer (2019). A picture tells a thousand… exposures: opportunities and challenges of deep learning image analyses in exposure science and environmental epidemiology. *Environment International*, 122, 3-10.

Wong, P.Y., HY Lee, YC Chen, YT Zeng, YR Chern, NT Chen, …, CD Wu (2021). Using a land use regression model with machine learning to estimate ground level PM2.5. *Environmental Pollution*, *277*, 116846.

Wu, C.D., YC Chen, WC Pan, YT Zeng, MJ Chen, YL Guo, SCC Lung (2017). Land-use regression with long-term satellite-based greenness index and culture-specific sources to model PM2.5 spatial-temporal variability. *Environmental Pollution*, *224*, 148-157.

Wu, C.D., YT Zeng, SCC Lung (2018). A hybrid kriging/land-use regression model to assess PM2.5 spatial-temporal variability. *Science of the Total Environment*, *645*, 1456-1464.

Young, M.T., MJ Bechle, PD Sampson, AA Szpiro, JD Marshall, L Sheppard, JD Kaufman (2016). Satellite-based NO2 and model validation in a national prediction model based on universal kriging and land-use regression. *Environmental Science & Technology*, *50*(7), 3686-3694.

Yu, H., A Russell, J Mulholland, T Odman, Y Hu, HH Chang, N Kumar (2018). Cross-comparison and evaluation of air pollution field estimation methods. *Atmospheric Environment*, *179*(January), 49–60.

Zhang, X., AC Just, HHL Hsu, I Kloog, M Woody, Z Mi, J Rush, P Georgopoulos, RO Wright, A Stroustrup (2021). A hybrid approach to predict daily NO2 concentrations at city block scale. *Science of the Total Environment*, *761*(2), 143279.

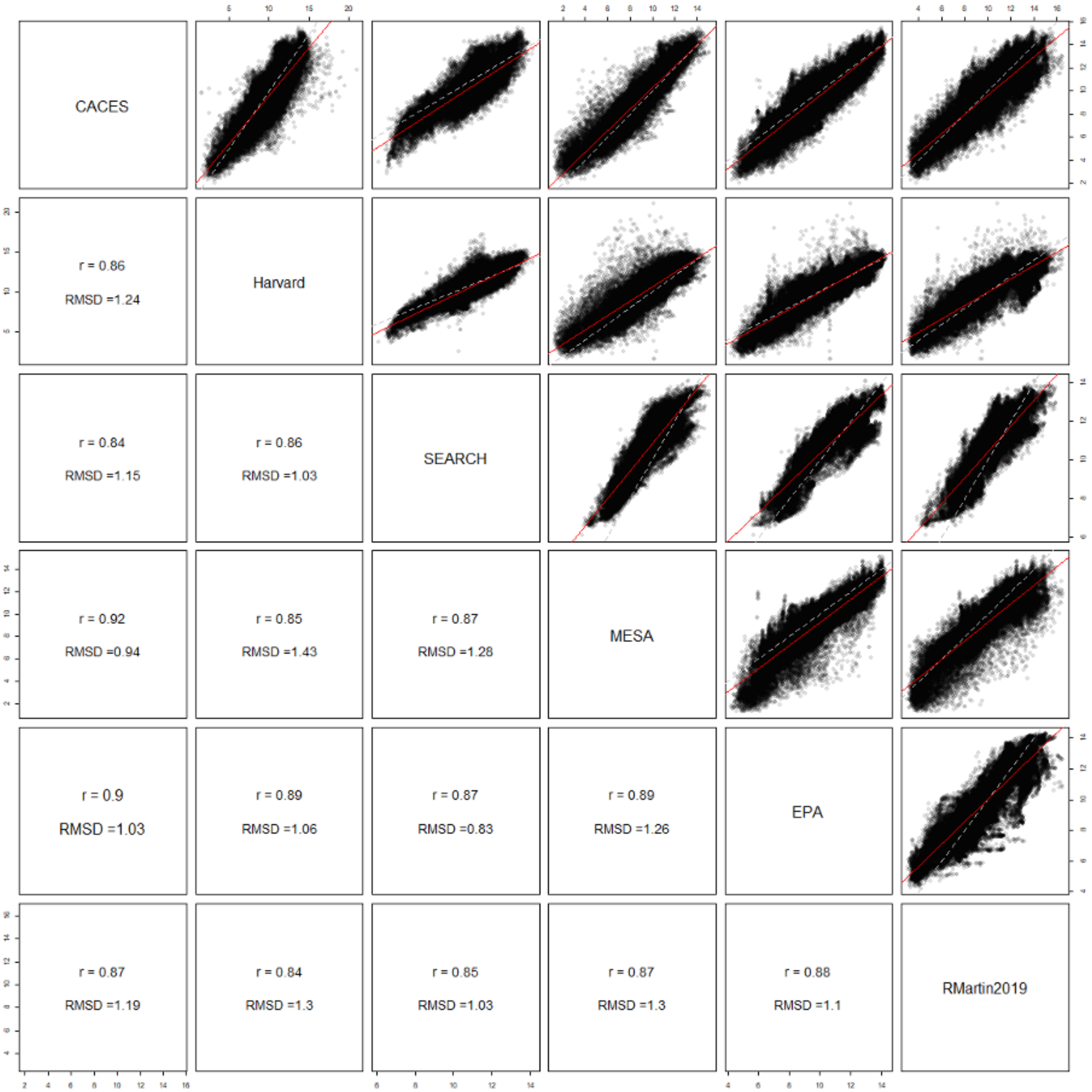


Figure 1: Scatterplot matrix for 2010 tract-level PM$_{2.5}$. Scatterplots in the upper right show pairwise tract-level predictions from each model (µg/m^3). Grey dashed line shows 1:1 line, red solid line shows linear trendline. Corresponding boxes in the bottom left show Pearson’s correlation (*r*; unitless) and root mean squared difference (RMSD; µg/m^3) between model predictions.

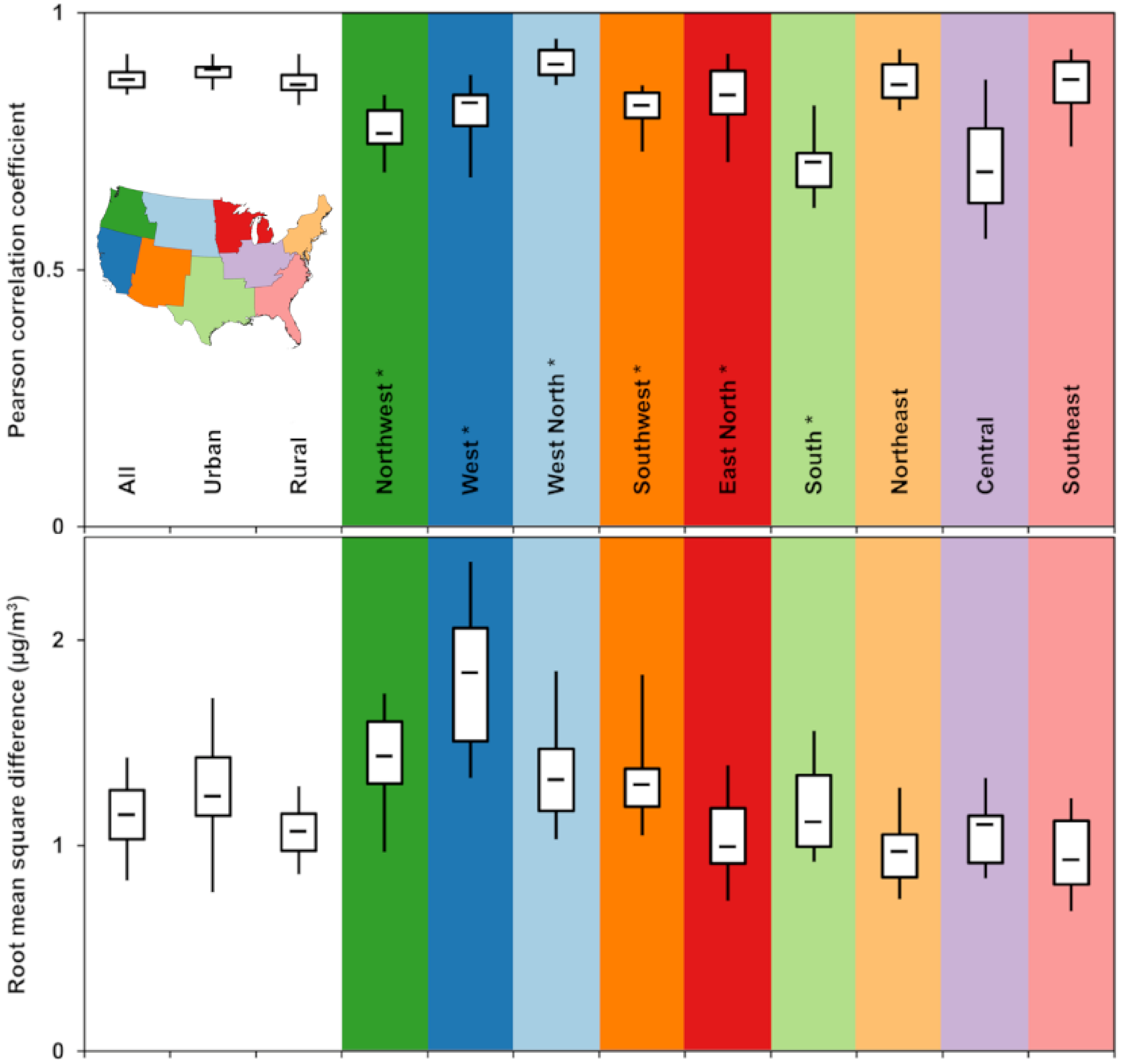


Figure 2: Summary of pairwise Pearson correlation coefficients (top) and root mean square difference (bottom) for all locations, urban and rural locations, and NOAA climate regions. Horizontal bar shows the median, box shows the interquartile range, and vertical lines show max and min values among model comparisons. The six NOAA regions denoted with an asterisk ("*") exclude SEARCH predictions as they were unavailable geographically. The results for those six regions reflect 10 pairwise comparisons of five models; results for the other regions (without an asterisk) reflect 15 pairwise comparisons of six models.

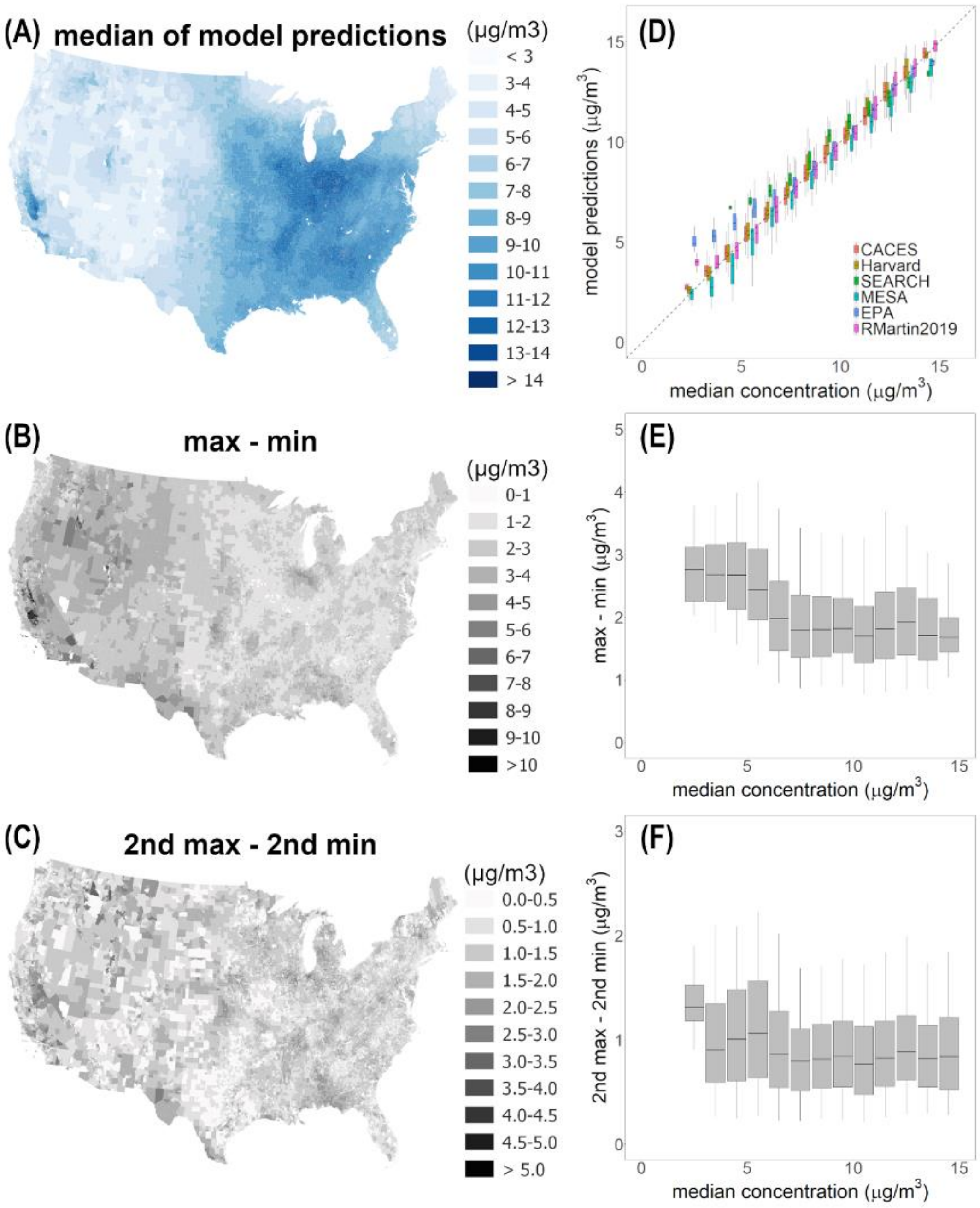
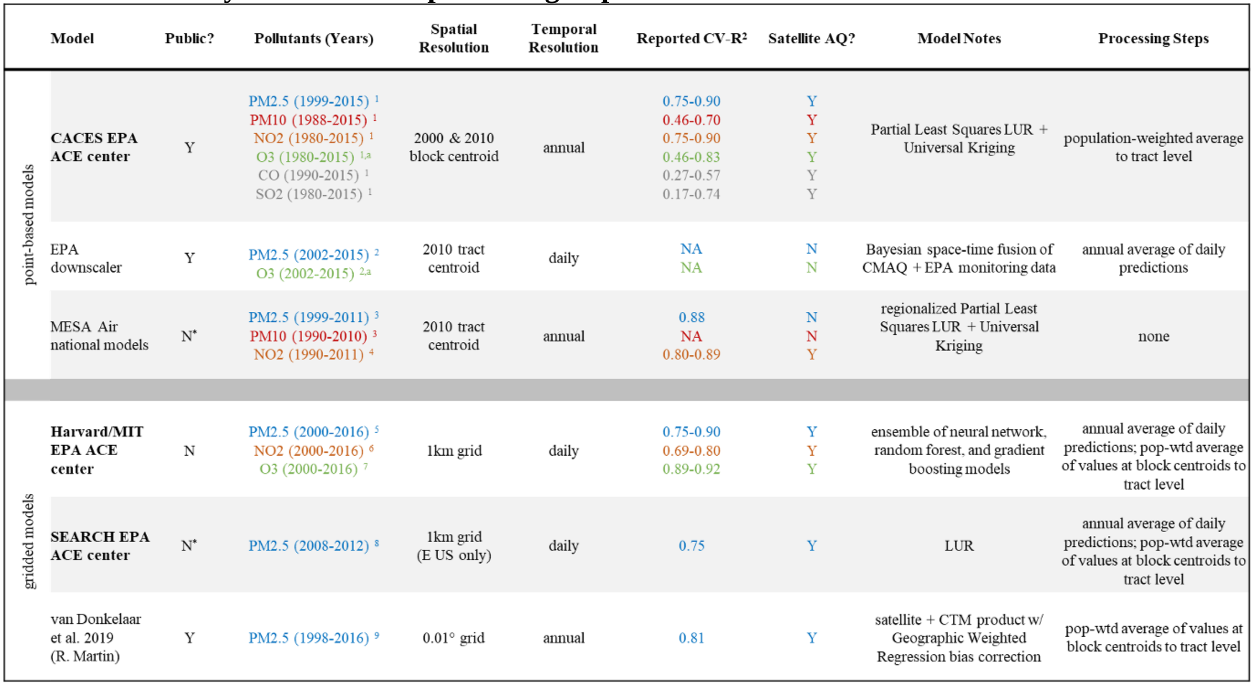


Figure 3: Variability by concentration and location. Maps show median concentration among model predictions within each tract (A) and within-tract variability of model predictions calculated as the max minus min (B) and 2nd max minus 2nd min (C). Boxplots show (y-axis) range of tract-level model predictions (D) and within-tract variation calculated as either max minus the min (E) or 2nd max minus 2nd min (E) of model predictions within each tract as a function of the median concentration among model predictions within each tract, binned to 1 µg/m^3 bins (x-axis). In the boxplots, 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.

Table 1: Summary of models and processing steps



	Model	Public?	Pollutants (Years)	Spatial Resolution	Temporal Resolution	Reported CV-R²	Satellite AQ?	Model Notes	Processing Steps
point-based models	**CACES EPA ACE center**	Y	PM2.5 (1999-2015) [1] PM10 (1988-2015) [1] NO2 (1980-2015) [1] O3 (1980-2015) [1,a] CO (1990-2015) [1] SO2 (1980-2015) [1]	2000 & 2010 block centroid	annual	0.75-0.90 0.46-0.70 0.75-0.90 0.46-0.83 0.27-0.57 0.17-0.74	Y Y Y Y Y Y	Partial Least Squares LUR + Universal Kriging	population-weighted average to tract level
	EPA downscaler	Y	PM2.5 (2002-2015) [2] O3 (2002-2015) [2,a]	2010 tract centroid	daily	NA NA	N N	Bayesian space-time fusion of CMAQ + EPA monitoring data	annual average of daily predictions
	MESA Air national models	N*	PM2.5 (1999-2011) [3] PM10 (1990-2010) [3] NO2 (1990-2011) [4]	2010 tract centroid	annual	0.88 NA 0.80-0.89	N N Y	regionalized Partial Least Squares LUR + Universal Kriging	none
gridded models	**Harvard/MIT EPA ACE center**	N	PM2.5 (2000-2016) [5] NO2 (2000-2016) [6] O3 (2000-2016) [7]	1km grid	daily	0.75-0.90 0.69-0.80 0.89-0.92	Y Y Y	ensemble of neural network, random forest, and gradient boosting models	annual average of daily predictions; pop-wtd average of values at block centroids to tract level
	SEARCH EPA ACE center	N*	PM2.5 (2008-2012) [8]	1km grid (E US only)	daily	0.75	Y	LUR	annual average of daily predictions; pop-wtd average of values at block centroids to tract level
	van Donkelaar et al. 2019 (R. Martin)	Y	PM2.5 (1998-2016) [9]	0.01° grid	annual	0.81	Y	satellite + CTM product w/ Geographic Weighted Regression bias correction	pop-wtd average of values at block centroids to tract level

[a] CACES and EPA downscaler ozone modeled as 5-month (May-Sept) ozone season average of daily 8-hr max.

*Tract centroid predictions may be publicly released at a later date.

1. Kim et al., 2020.

2. EPA Downscaler Model.

3. Sampson et al., 2013.

4. Young et al., 2016.

5. Di et al., 2019.

6. Di et al., 2020.

7. Requia et al., 2020.

8. Goldberg et al., 2019.

9. van Donkelaar et al., 2019.