Environmental prediction for diverse contexts

Insights for contemporary $PM_{2.5}$ research and policy

Andrew Dickinson *with Ed Rubin*

adickin3@uoregon.edu

University of Oregon | TWEEDS 2025

How "good" are pollution predictions?

Introduction

Motivation

Joint advancements in machine learning + satellite imagery has led to an emergence of predictions of *environmental quality*.

Data source increasingly applied in causal inference settings. *Why?*

- **High coverage.** Satellite imagery is spatially continuous
- **Fine resolution.** Raster data at 1km pixels and daily frequency

Features allowing researchers to answer previously unanswerable questions

Introduction

Fine particulate matter

One particular literature where prediction estimates are growing in empirical applications is predicted fine particulate matter, or $\mathbf{PM_{2.5}}$

- Daily/monthly predictions of $\mathbf{PM_{2.5}}$ concentrations across space
- Increasingly popular data in public health and economics literature

Learn relationship between *in situ* monitors and remotely-sensed features

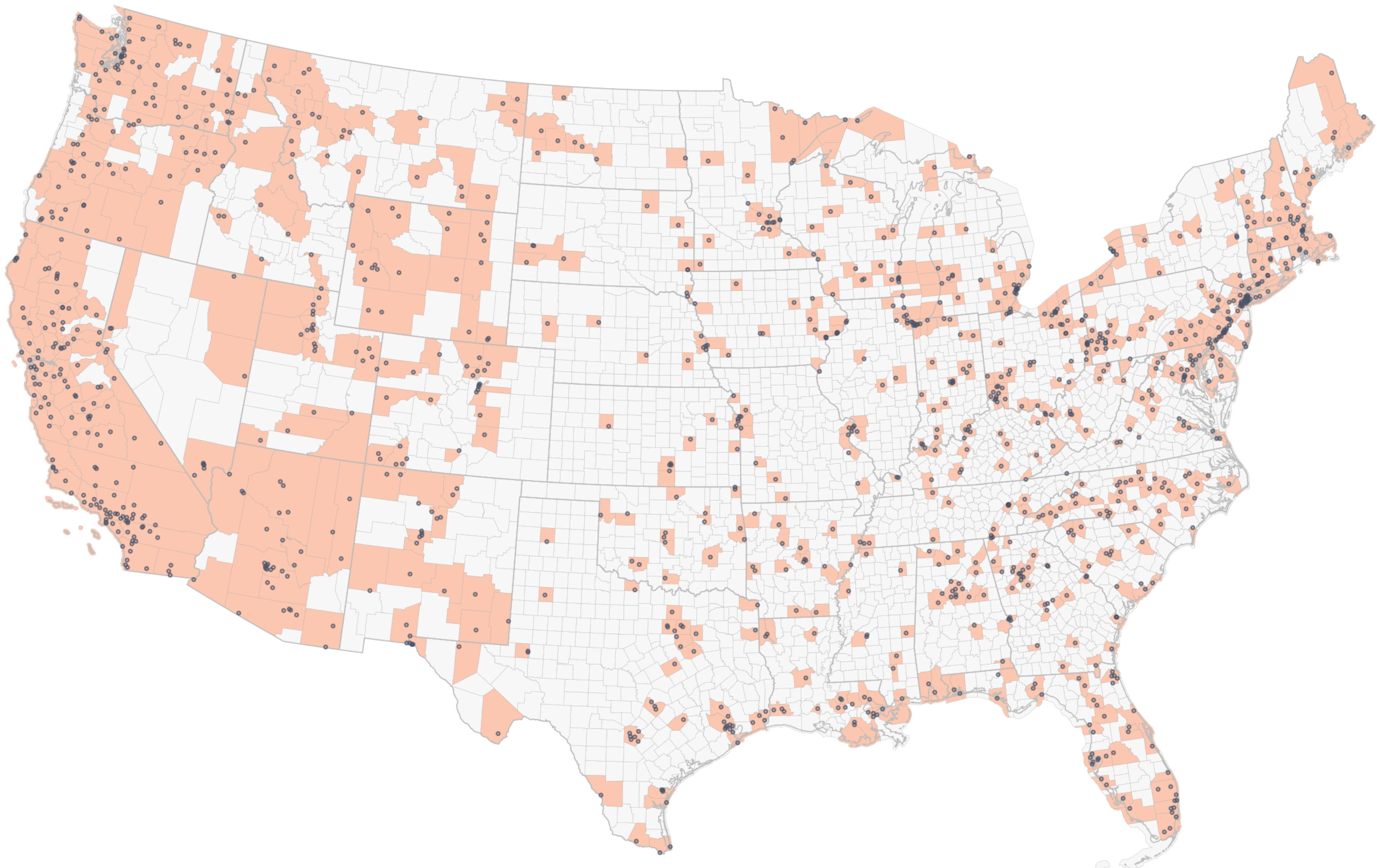
- Monitor observations as "**ground truth**"
- Validate predictions using cross-validation to prevent overfitting
- Predict $PM_{2.5}$ concentrations at unobserved locations/times

Introduction

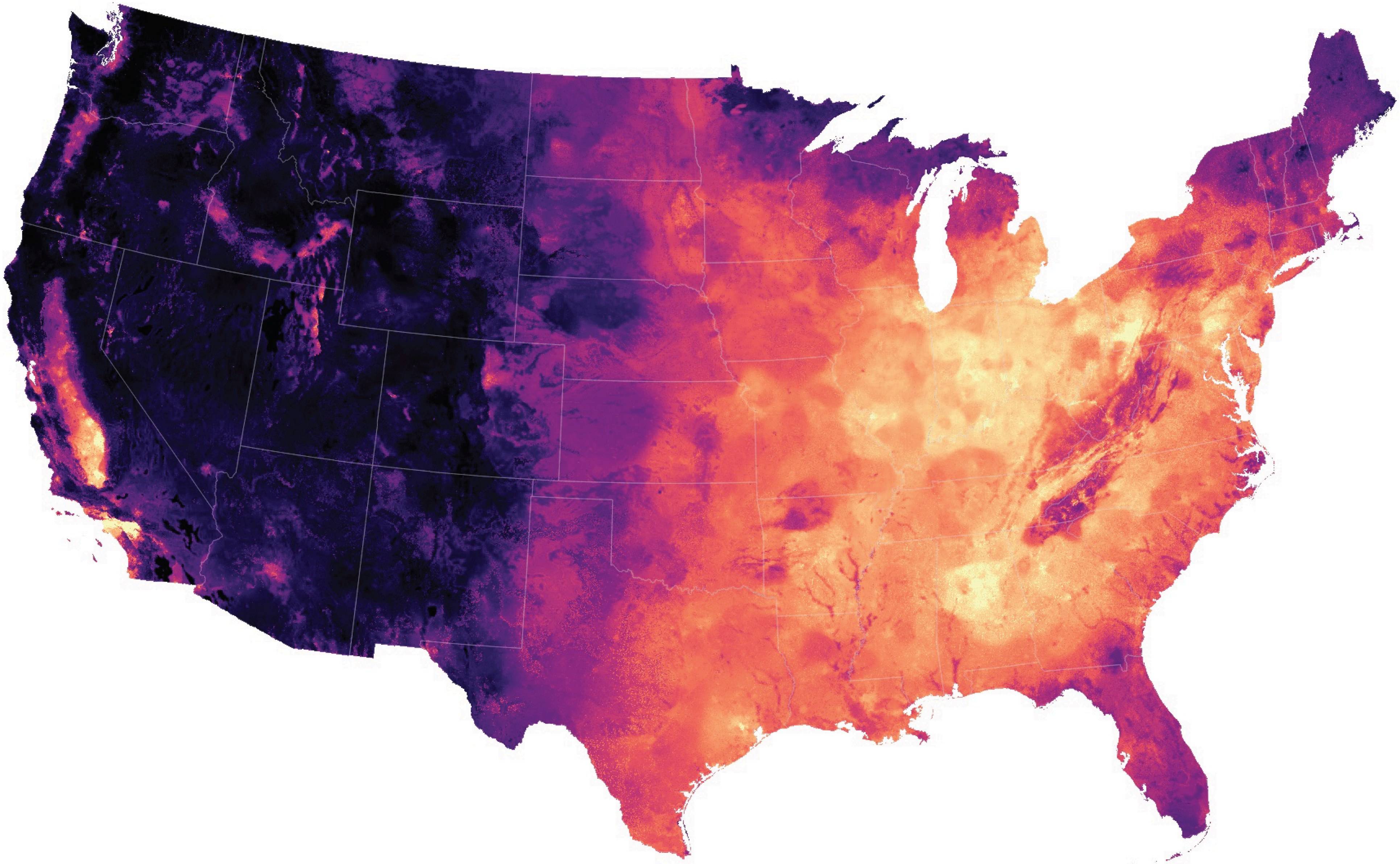
Regulatory monitors

Between 2002-2019, CONUS monitored by an array of **2,920**[1] *in situ* monitors

- High accuracy, *at a particular location*
- High costs, *limited number of monitors*



Regulatory $PM_{2.5}$ monitors and counties in the US, June 2012



Predicted $PM_{2.5}$ concentrations in the US in 2005

Sources: Fowlie, Rubin and Walker (2019) and Di *et al.* (2016)

Introduction

Problem

While these estimates are exciting and promising, there are no oracles

Despite this, some applications have treated these estimates as *"truth"*

- **Measurement error** underestimated, treated as classical
- **Uncertainty** ignored

Predictions treated as a "one-size-fits-all" dataset

$PM_{2.5}$ products

Authors & Year	Years	Frequency	Extent	R^2	Citations
van Donkelaar *et al.* (2016)	1998-2014	Yearly	Global	[0.78, 0.81]	1,015
Wei *et al.* (2021)	2000-2018	Monthly	China	[0.80, 0.90]	531
Di *et al.* (2016)	2000-2012	Daily	CONUS	[0.74, 0.88]	413
Hu *et al.* (2017)	2011	Daily	CONUS	[0.64, 0.83]	404
Di *et al.* (2019)	2000-2016	Daily	CONUS	[0.73, 0.91]	382
Wei *et al.* (2020)	2018	Daily	China	[0.88, 0.89]	373
Reid *et al.* (2015)	2008	Daily	Northern CA	0.80	252
Van Donkelaar *et al.* (2021)	1998-2019	Monthly	Global	[0.51, 0.86]	73
van Donkelaar *et al.* (2019)	1998-2019	Monthly	Global	[0.75, 0.95]	68
Meng *et al.* (2019)	1981–2016	Yearly	North America	[0.60, 0.85]	59
Requia *et al.* (2020)	2000-2016	Daily	CONUS	[0.86, 0.93]	56
Reid *et al.* (2021)	2008-2018	Daily	Western US	[0.58, 0.73]	30

Introduction

Research questions

We hope to elucidate these issues by answering the following:

1. How does predictive accuracy change across uses?
2. How much uncertainty lies behind predictions?
3. How does non-randomness of monitor sites affect generalizability?

Introduction

To answer these questions

Produce monthly **$PM_{2.5}$** (*1km x 1km*) predictions for the **CONUS** (*2002-2019*)

We follow the approach and feature set of two highly cited papers:

- Di *et al.* (2016); Di *et al.* (2019)

Why take this approach?

- Raw data and gridded output are publicly available
- Missing is the intermediate steps used to generate the gridded output

Modeling

Predicting $PM_{2.5}$

To estimate monthly $\mathbf{PM_{2.5}}$ (1km × 1km) using a LightGBM learner:

$$\widehat{PM}_{it} = f_{GBM}\left(\mathbf{X}_{it}, \mathbf{Z}_i, \mathbf{S}_i\right)$$

- $\mathbf{X}_{it}$: Time-varying features (e.g., AOD, weather, CTM outputs)
- $\mathbf{Z}_i$: Time-invariant features (e.g., land use, elevation, NDVI)
- $\mathbf{S}_i$: Spatial lag features (IDW monitor readings)

Trained via *nested cross-validation* to minimize **MSE**

Modeling

Measuring uncertainty in $PM_{2.5}$ predictions

We quantify predictive uncertainty using LightGBM quantile regression:

- Separate models for 2.5th and 97.5th percentiles
- Trained using the *pinball loss function*

$$L(\tau, x, y) = \begin{cases} \tau(x - y), & \text{if } x \geq y \\ (1 - \tau)(y - x), & \text{if } x < y \end{cases}$$

Two quantile regressions are differenced to produce a 95% prediction intervals

$$\widehat{PM}_{0.975} - \widehat{PM}_{0.025}$$

How does predictive accuracy change across uses?

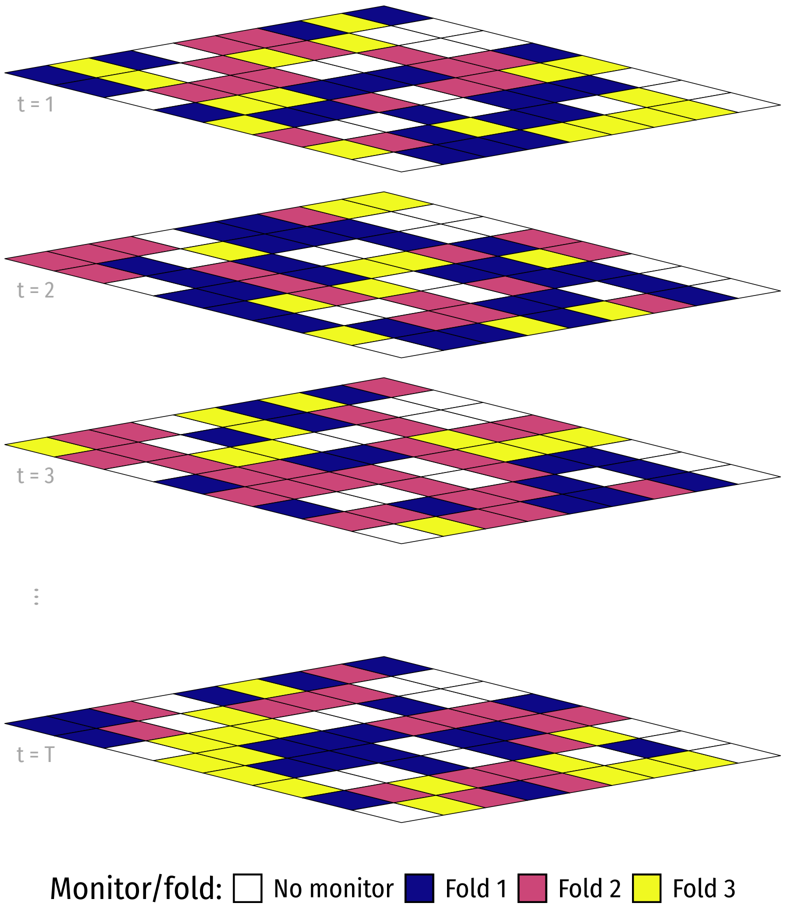
Model Evaluation

How does predictive accuracy change across uses?

The standard CV approach is independent identically distributed (**IID**) CV

- Randomly samples monitor-month observations, unclustered

IID CV



Cross validation in temporally repeated grids. Standard IID CV using 3-fold cross-validation. Each layer of pixel describes the sample across different points in time, and the color of each pixel describes the fold that the observation is assigned to. White folds indicate areas without monitors.

Model Evaluation

How does predictive accuracy change across uses?

The standard CV approach is independent identically distributed (**IID**) CV

- Randomly samples monitor-month observations, unclustered If we wanted to interpolate missing data at monitors, **IID CV** is reasonable

If the goal is to estimate $PM_{2.5}$ in unmonitored areas, **IID CV** is not appropriate

- Ignores the **spatial** and **temporal** (panel) structure of the data
  - ↳ Overestimates model performance

Model Evaluation

How does predictive accuracy change across uses?

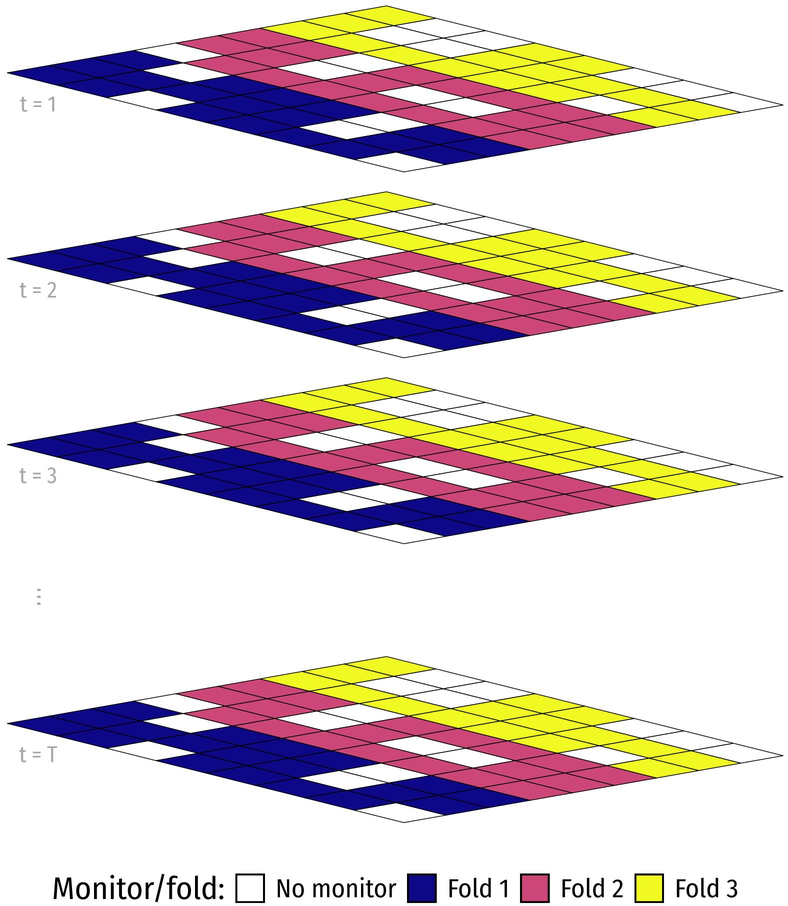
There is no *one-size-fits-all* cross-validation approach

- Training and validation should match the downstream use case

To learn out-of-sample, **spatial cross-validation** (SPCV) is better suited

- Clusters monitor-months by spatial proximity
- Evaluation is done outside each cluster, mimicking unmonitored space

Spatial cross-validation



Cross validation in temporally repeated grids. Spatial resampling approach, where the data is clustered into 3 distinct spatial clusters. Each cluster is then used as a fold in the cross-validation process, effectively limiting the model to only learn from observations in the same cluster.

Model Evaluation

Nested cross-validation

Additionally, we incorporate a nested cross-validation approach

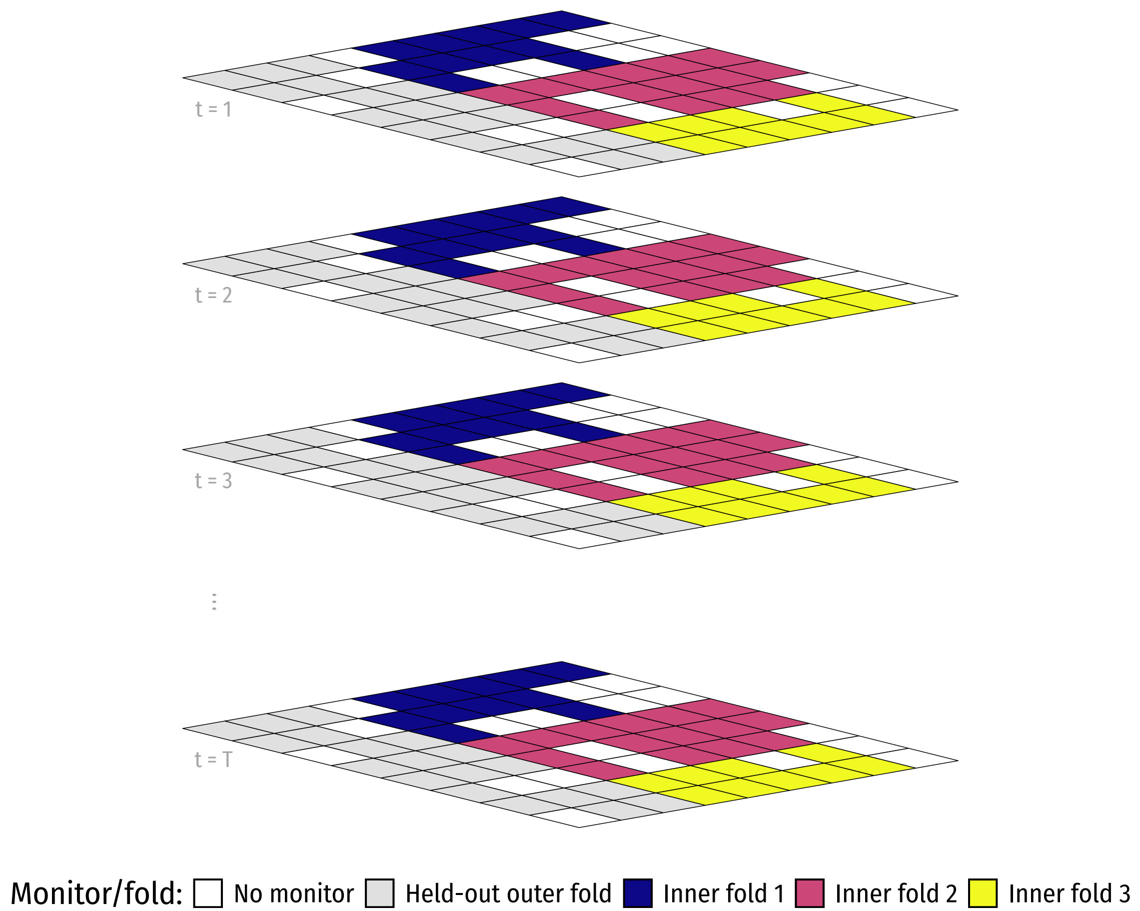
- Inner loop: hyperparameter tuning
- Outer loop: model evaluation

Ensures an unbiased estimate of the model's generalization error

We assess the model's ability across different **four** validation approaches

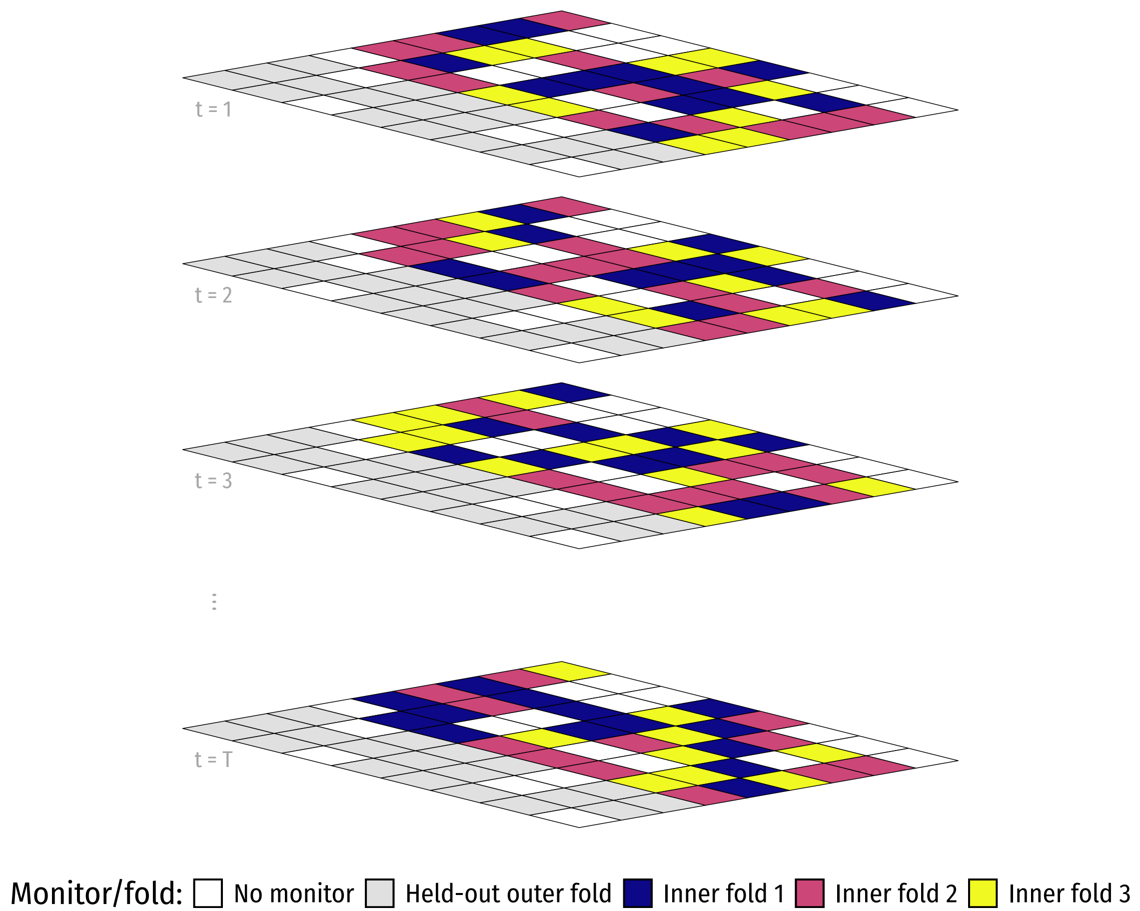
- IID-IID, IID-SPCV, SPCV-IID, SPCV-SPCV

SPCV-SPCV



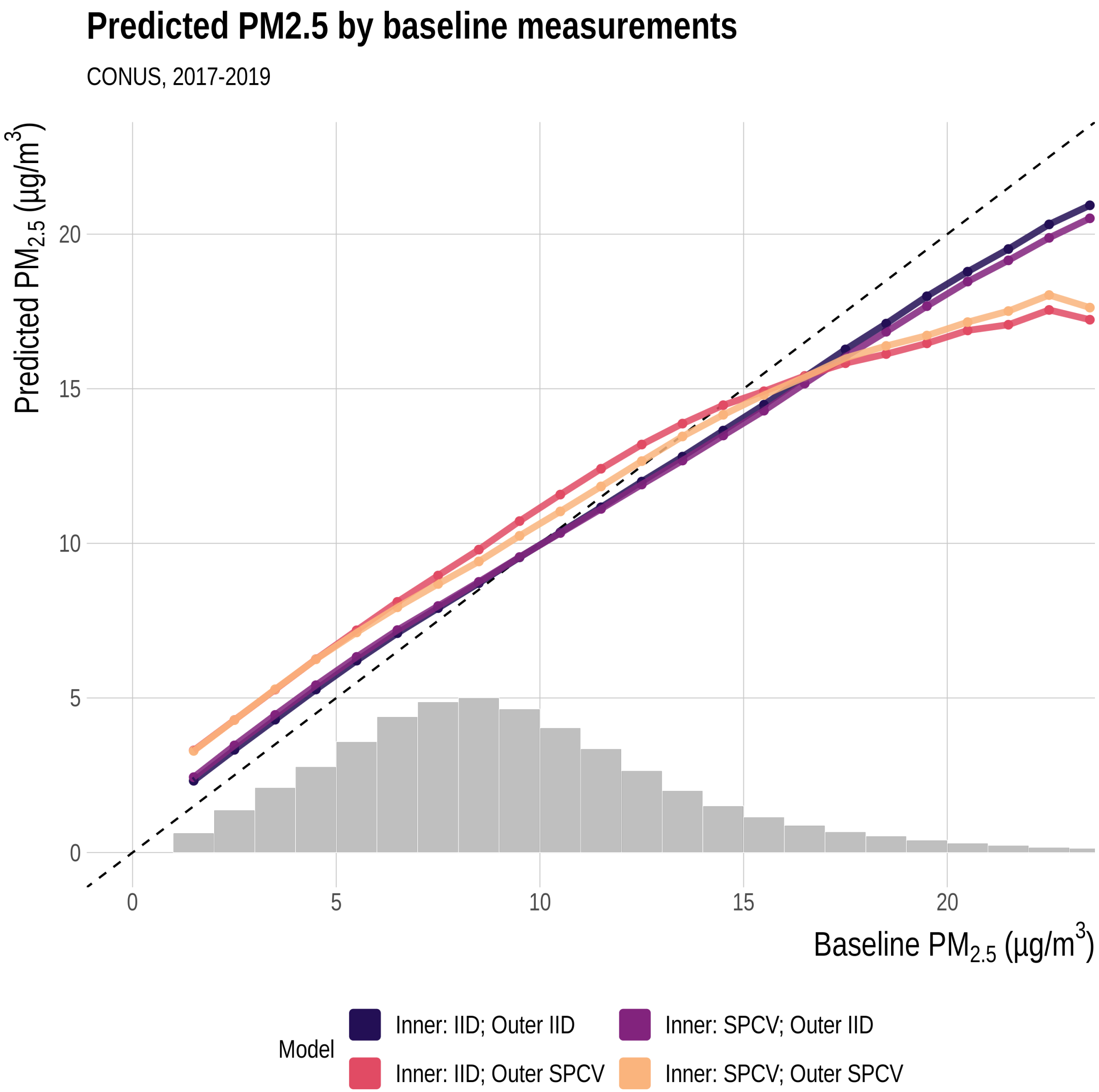
Nested cross validation in temporally repeated grids: Plot illustrates inner SPCV and outer SPCV nested cross-validation in temporally repeated grid. Only one outer fold is shown for clarity, colored in gray, but the process is repeated three times.

SPCV-IID



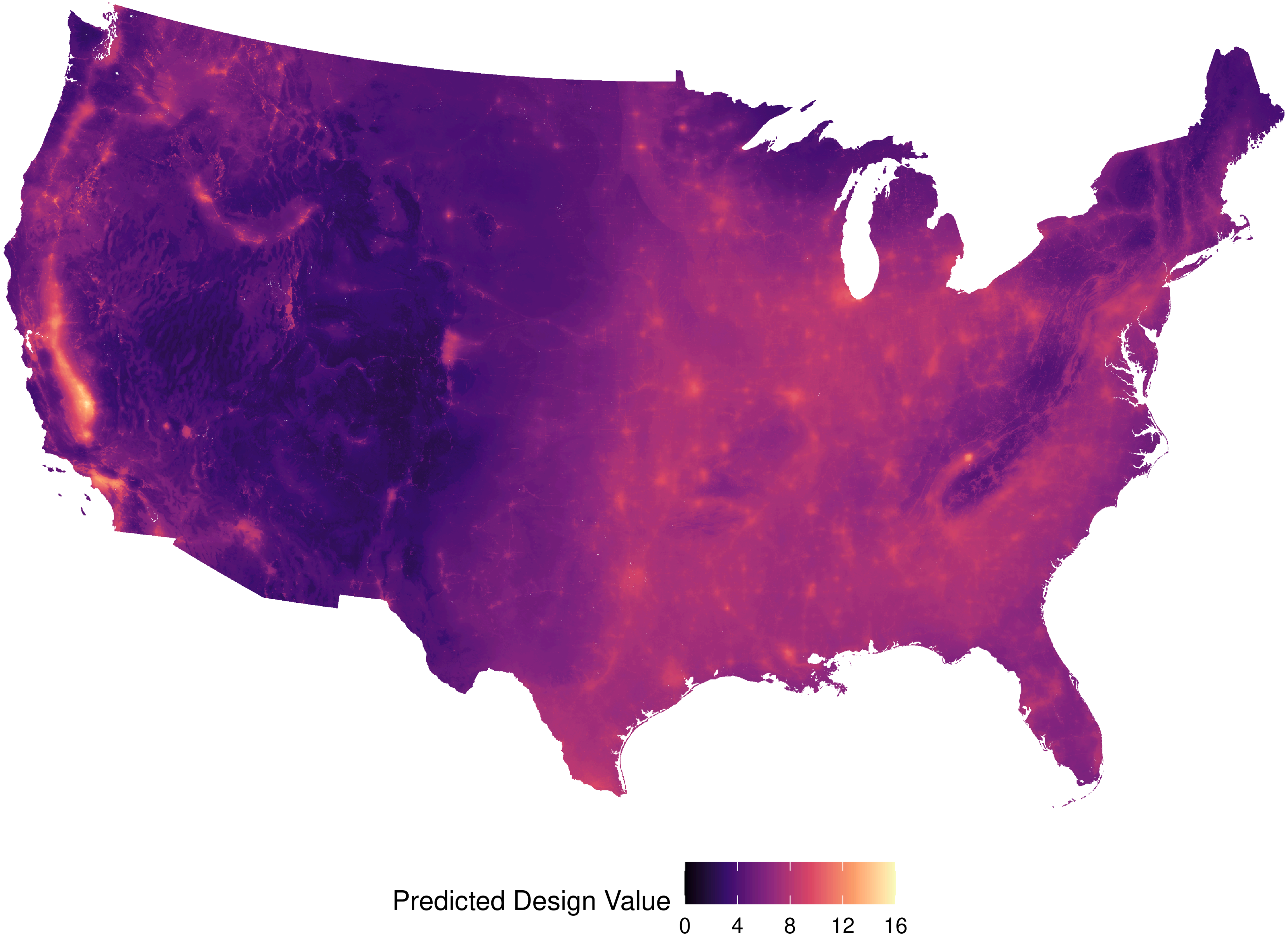
Nested cross validation in temporally repeated grids: Plot illustrates inner IID and outer SPCV nested cross-validation in temporally repeated grid. Only one outer fold is shown for clarity, colored in gray, but the process is repeated three times.

$PM_{2.5}$ prediction accuracy declines steeply when spatially validated and/or restricted from using *close* spatial lags. Matrix cells display (and are filled) by R^2 values from the combination of cross-validation approach (row) and available spatial lags (columns).

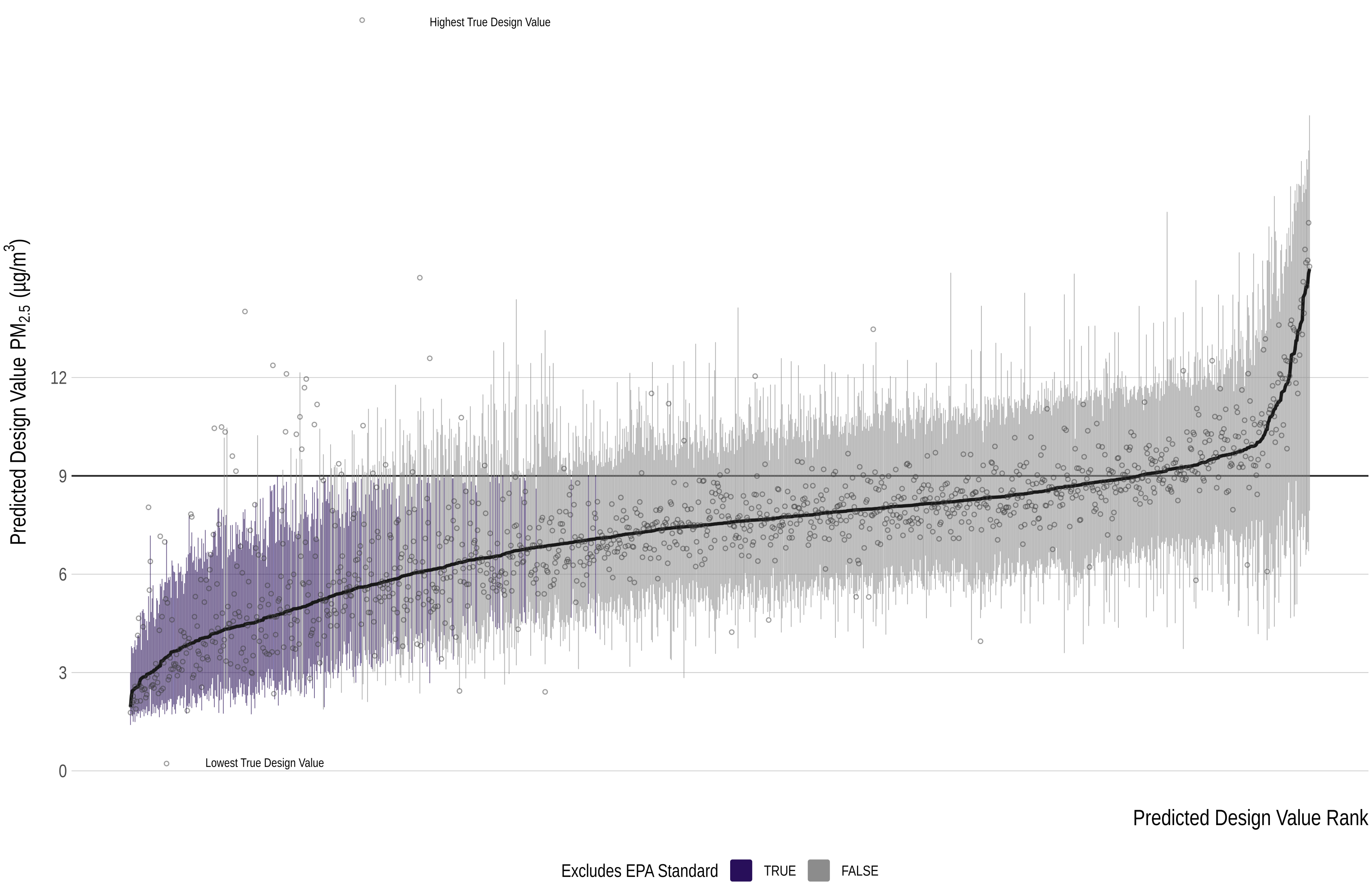


Out-of-sample $PM_{2.5}$ prediction accuracy. Comparison of binned predicted $PM_{2.5}$ values to binned true $PM_{2.5}$ values for pixels with monitors.

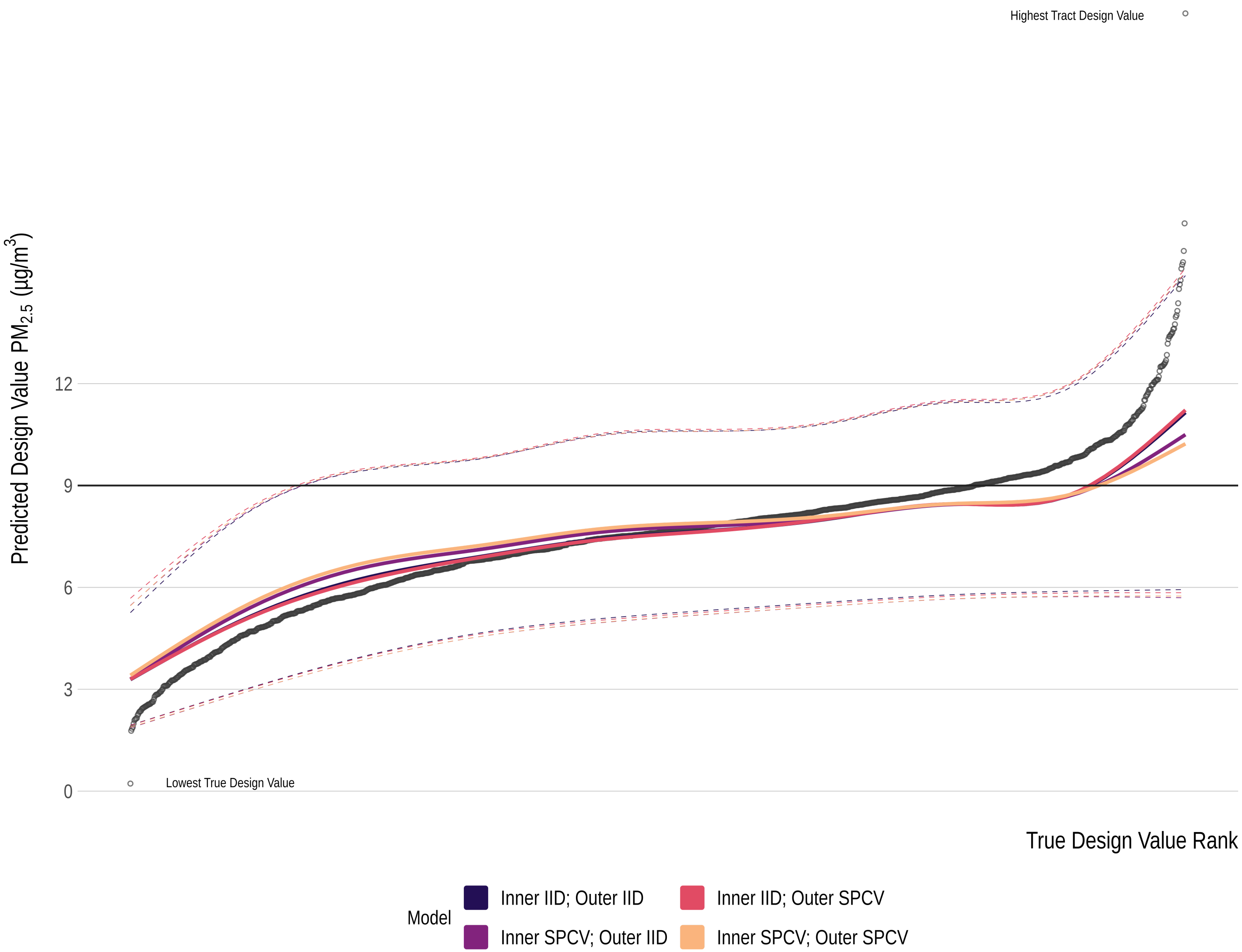
How much uncertainty lies behind predictions?



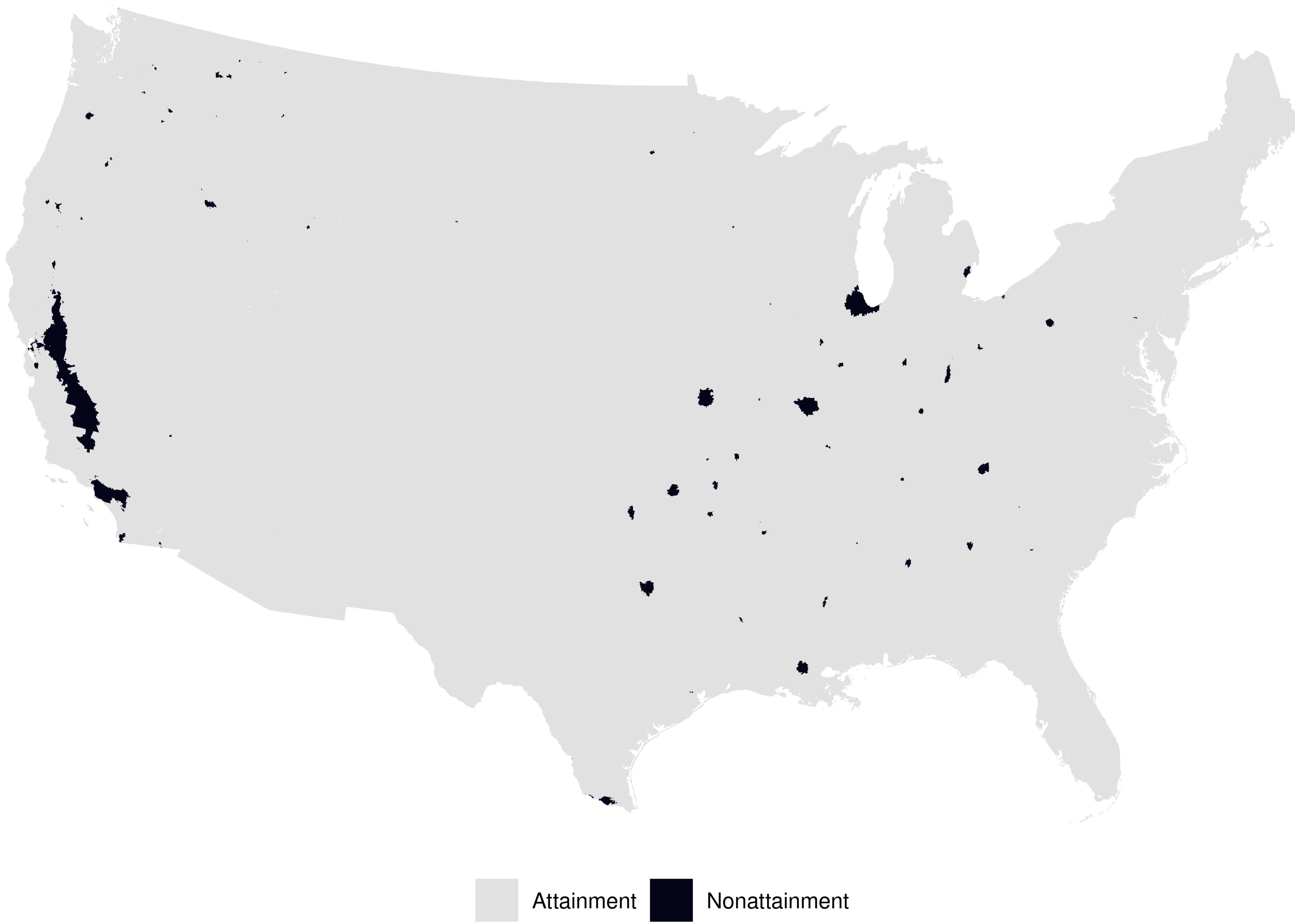
Pixel Design Values: Plot of predicted Design Values for each pixel generated with predictions between 2017-2019



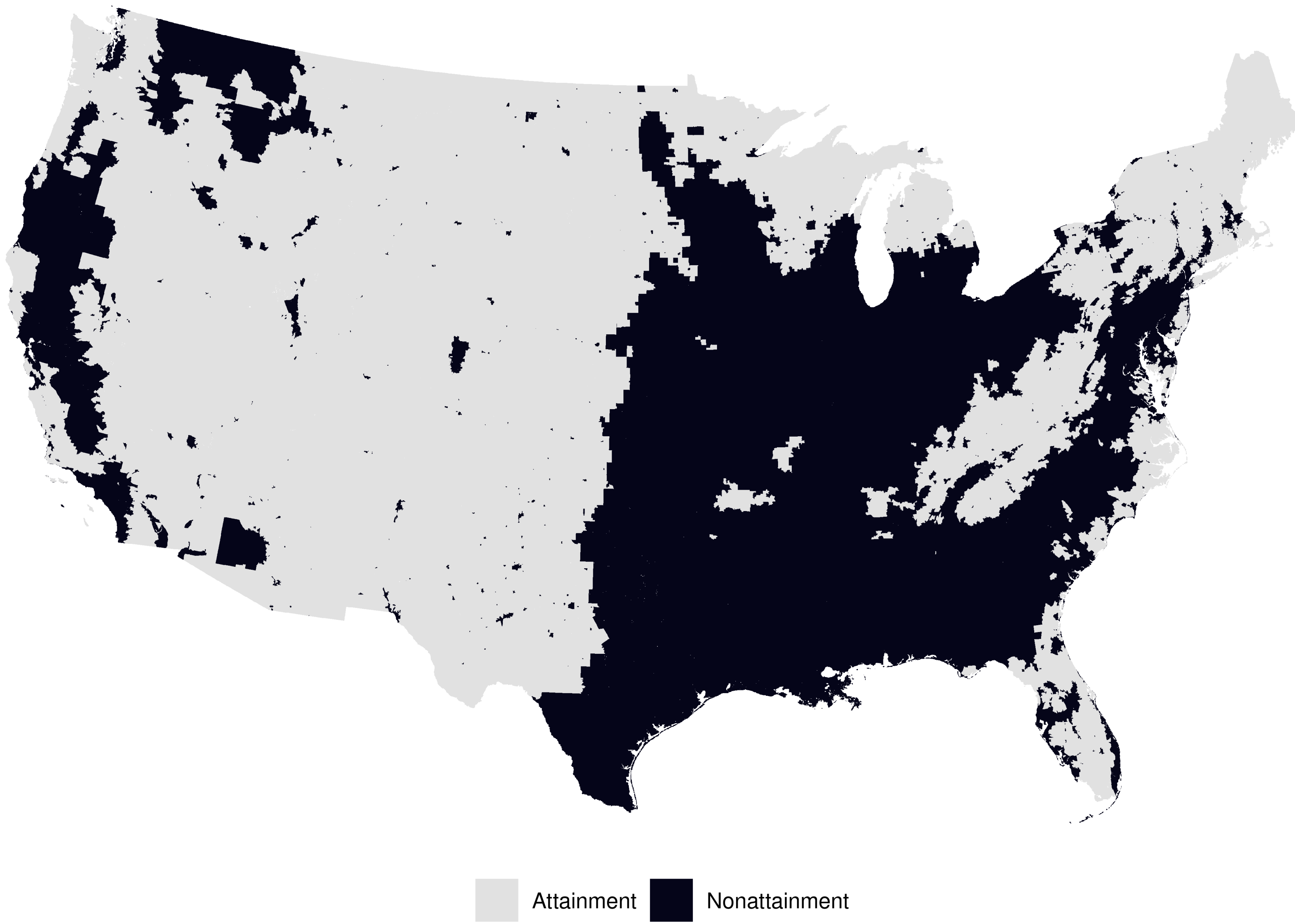
Attainment Status by Design Value Rank. Predicted Design Values against their Predicted Design Value rank-order (from lowest to highest) of Census Tracts with a monitor and associated prediction intervals. Vertical intervals show uncertainty around predicted Design Values, with purple intervals indicating tracts confidently classified as compliant, and grey intervals indicating tracts where compliance status is uncertain.



Attainment Status by Design Value Rank. True Design Values against their True rank-order (from lowest to highest) of Census Tracts with a monitor and associated prediction intervals. Comparison of these results

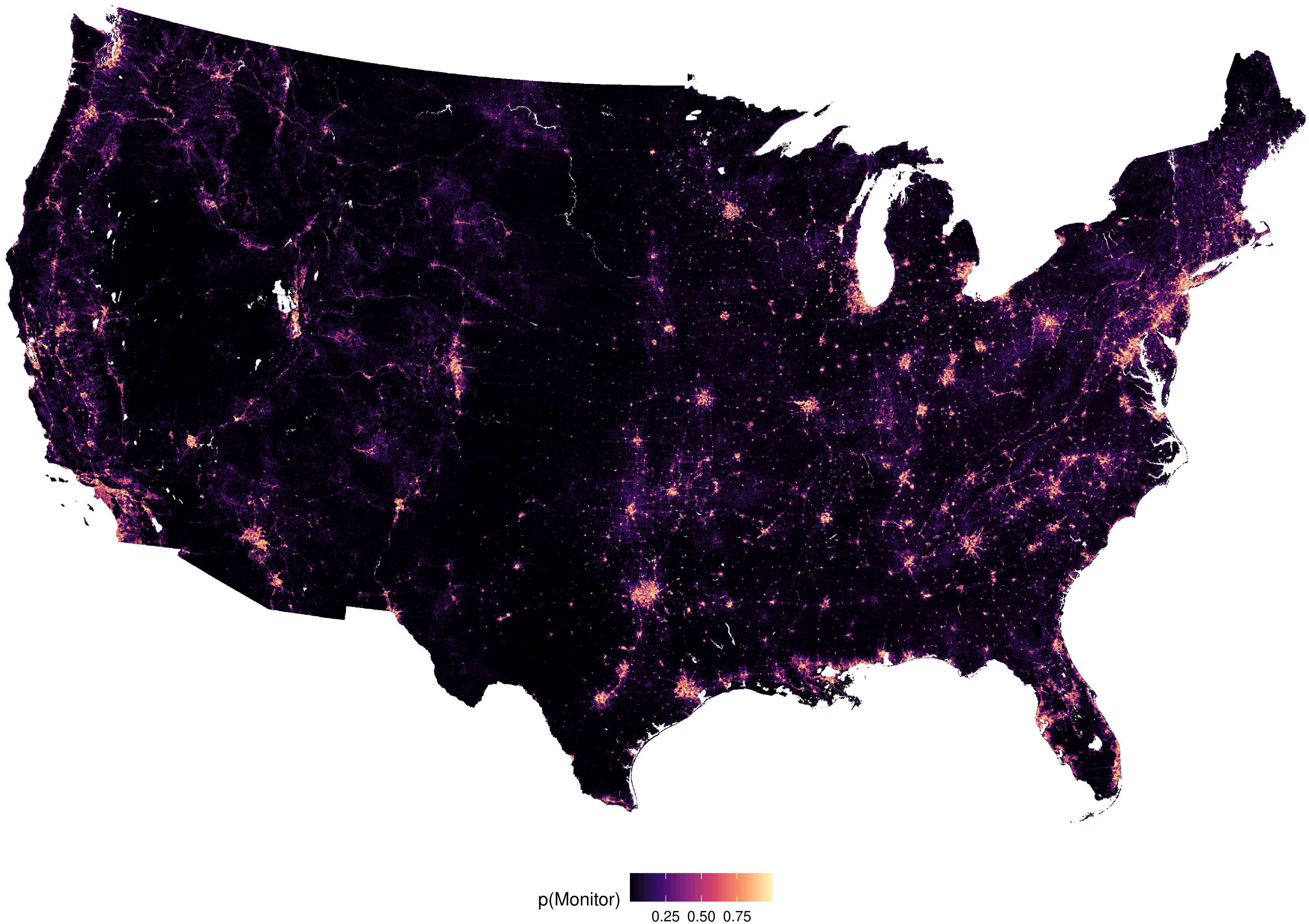


Attainment Status by Design Value. Plot of attainment status by Design Value, aggregated to the tract level. Census tracts that do not meet criteria for attainment are colored dark.



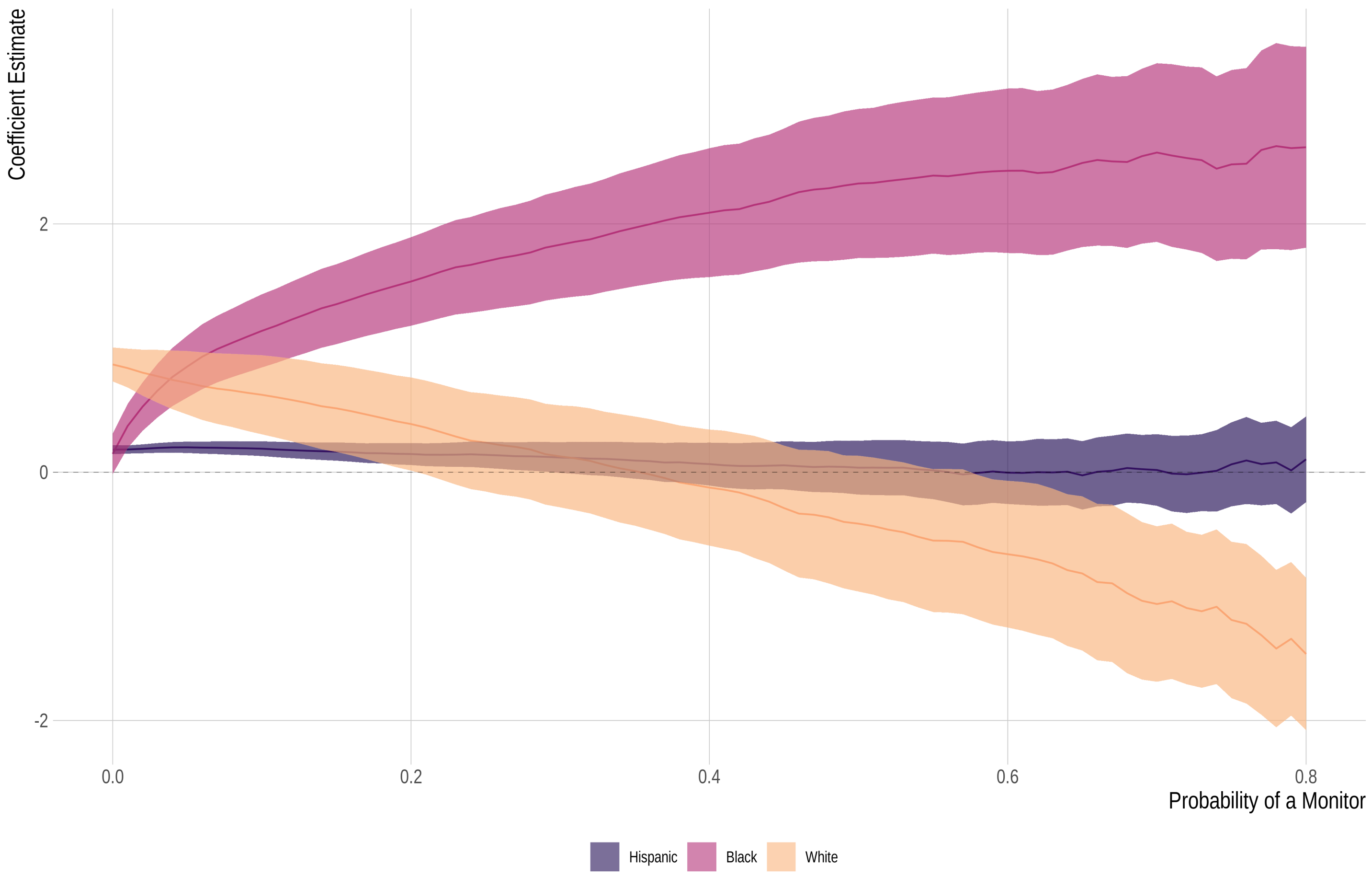
Attainment Status by Upper Bound of Design Value. Plot of attainment status by Design Value, aggregated to the tract level. Tracts colored dark cannot rule out being above the standard given prediction interval.

How does non-randomness of monitor sites affect generalizability?



Probability of monitor presence across the CONUS. Color gradient probability of a pixel containing a monitor. Darker pixels indicate lower probability and greater potential for uncertainty.

$$\widehat{DV}_i = \beta_0 + \beta_1 \text{Hispanic} + \beta_2 \text{Black} + \beta_3 \text{White} + \beta_4 \text{Urban} + \delta_i + \varepsilon_i$$



Regression coefficients for across different monitor-presence probability thresholds: Estimated coefficients of percentile white, black, and hispanic and corresponding confidence intervals of each demographic group against increasing monitor-presence probability thresholds.

Summary

Air quality predictions are a big deal, but the predictions have problems

1a. Accuracy falls sharply with distance from monitors and without spatial lags
1b. Tree-based models are not learning the spatial variation of $PM_{2.5}$

2. Prediction intervals are large, there is a lot of uncertainty, even near monitors

3. Controlling for monitor presence can meaningfully affect OLS regression estimates

References

Di, Q. *et al.* (2016) "Assessing PM2.5 exposures with high spatiotemporal resolution across the continental united states," *Environmental Science & Technology*, 50(9), pp. 4712–4721. Available at: https://doi.org/10.1021/acs.est.5b06121.

Di, Q. *et al.* (2019) "An ensemble-based model of PM2.5 concentration across the contiguous united states with high spatiotemporal resolution," *Environment International*, 130, p. 104909. Available at: https://doi.org/10.1016/j.envint.2019.104909.

Fowlie, M., Rubin, E. and Walker, R. (2019) "Bringing satellite-based air quality estimates down to earth," *AEA Papers and Proceedings*, 109, pp. 283–288. Available at: https://doi.org/10.1257/pandp.20191064.

Hu, X. *et al.* (2017) "Estimating PM $_{2.5}$ Concentrations in the Conterminous United States Using the Random Forest Approach," *Environmental Science & Technology*, 51(12), pp. 6936–6944. Available at: https://doi.org/10.1021/acs.est.7b01210.

Meng, J. *et al.* (2019) "Estimated Long-Term (1981–2016) Concentrations of Ambient Fine Particulate Matter across North America from Chemical Transport Modeling, Satellite Remote Sensing, and Ground-Based Measurements," *Environmental Science & Technology*, 53(9), pp. 5071–5079. Available at: https://doi.org/10.1021/acs.est.8b06875.

Reid, C.E. *et al.* (2015) "Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning," *Environmental Science & Technology*, 49(6), pp. 3887–3896. Available at: https://doi.org/10.1021/es505846r.

Reid, C.E. *et al.* (2021) "Daily PM2.5 concentration estimates by county, ZIP code, and census tract in 11 western states 2008–2018," *Scientific Data*, 8, p. 112. Available at: https://doi.org/10.1038/s41597-021-00891-1.

Requia, W.J. *et al.* (2020) “An ensemble learning approach for estimating high spatiotemporal resolution of ground-level ozone in the contiguous United States,” *Environmental science & technology*, 54(18), pp. 11037–11047. Available at: https://doi.org/10.1021/acs.est.0c01791.

van Donkelaar, A. *et al.* (2016) “Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors,” *Environmental Science & Technology*, 50(7), pp. 3762–3772. Available at: https://doi.org/10.1021/acs.est.5b05833.

van Donkelaar, A. *et al.* (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), pp. 2595–2611. Available at: https://doi.org/10.1021/acs.est.8b06392.

Van Donkelaar, A. *et al.* (2021) “Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty,” *Environmental Science & Technology*, 55(22), pp. 15287–15300. Available at: https://doi.org/10.1021/acs.est.1c05309.

Wei, J. *et al.* (2020) “Improved 1 km resolution $PM_{2.5}$ estimates across China using enhanced space–time extremely randomized trees,” *Atmospheric Chemistry and Physics*, 20(6), pp. 3273–3289. Available at: https://doi.org/10.5194/acp-20-3273-2020.

Wei, J. *et al.* (2021) “Reconstructing 1-km-resolution high-quality PM2.5 data records from 2000 to 2018 in China: Spatiotemporal variations and policy implications,” *Remote Sensing of Environment*, 252, p. 112136. Available at: https://doi.org/10.1016/j.rse.2020.112136.