An Ensemble-based Model of $PM_{2.5}$ Concentration across the Contiguous United States with High Spatiotemporal Resolution

Qian Di[a], Heresh Amini[a], Itai Kloog[b], Rachel Silvern[c], James Kelly[d]; M. Benjamin Sabath[e], Christine Choirat[e], Petros Koutrakis[a], Alexei Lyapustin[f], Yujie Wang[g], Francesca Dominici[e], Joel Schwartz[a]

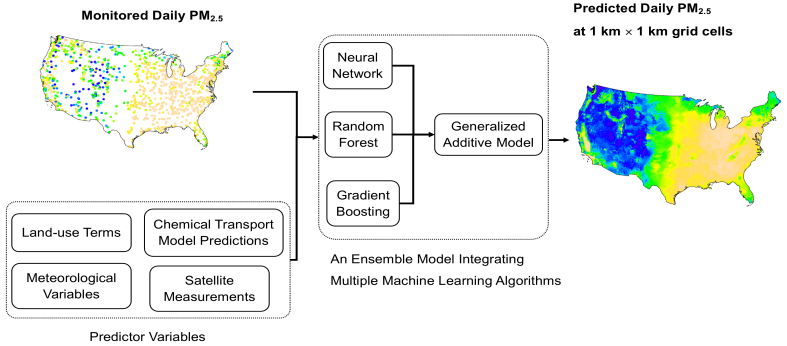
a. Department of Environmental Health, Harvard T.H. Chan School of Public Heath, Boston, Massachusetts, U.S.; b. Department of Geography and Environmental Development, Ben-Gurion University of the Negev, Beer Sheva, Israel; c. Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts, U.S.; d. U.S. Environmental Protection Agency, Office of Air Quality Planning & Standards, Research Triangle Park, North Carolina, U.S.; e. Department of Biostatistics, Harvard T.H. Chan School of Public Heath, Boston, Massachusetts, U.S.; f. NASA Goddard Space Flight Center, Greenbelt, Maryland, U.S.; g. NASA Goddard Space Flight Center, Greenbelt, Maryland, U.S.

Highlights

- Developed an ensemble model which integrates three machine learning algorithms and estimates $PM_{2.5}$
- Satellite measurements, land-use terms, and many variables as predictors
- Model predicts **daily $PM_{2.5}$** at 1 km × 1 km grid cells for the entire United States
- Model predictions were also downscaled to 100 m × 100 m
- Monthly uncertainty level of prediction was estimated

Background

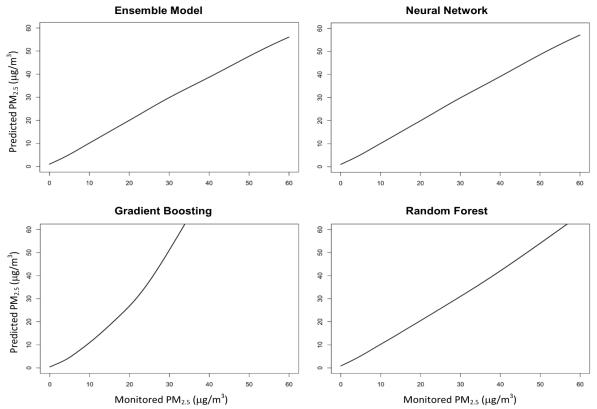
- Various approaches have been proposed to model $PM_{2.5}$ using satellite-derived aerosol optical depth, land-use variables, chemical transport model predictions, and several meteorological variables
- Training methods have been evolving from simple linear regressions to more complex machine learning algorithms
- **Goal:** Use an ensemble model which integrates multiple machine learning algorithms and predictor variables to estimate **daily $PM_{2.5}$** at a resolution of 1 km × 1 km across the contiguous United States.



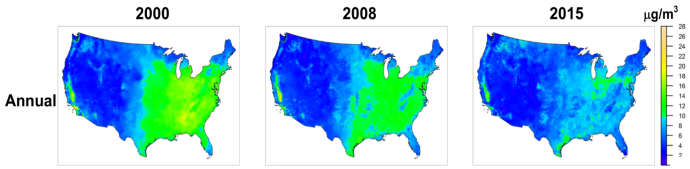
Methods

- Use a generalized additive model accounting for geographic differences to combine $PM_{2.5}$ estimates from neural network, random forest, and gradient boosting.
- Multiple predictor variables: satellite data, meteorological variables, land-use variables, elevation, chemical transport model predictions, several reanalysis datasets, and others.
- Model validation: model was validated with 10-fold cross-validation.
- Predictions: daily $PM_{2.5}$ from 2000 to 2015 at every 1 km × 1 km grid cell in the contiguous United States; used localized land-use variables within 1 km × 1 km grids to downscale predictions to 100 m × 100 m grid cells.
- Uncertainty: used meteorological variables, land-use variables, and elevation to model the monthly standard deviation of the difference between daily monitored and predicted $PM_{2.5}$ for every 1 km × 1 km grid cell.

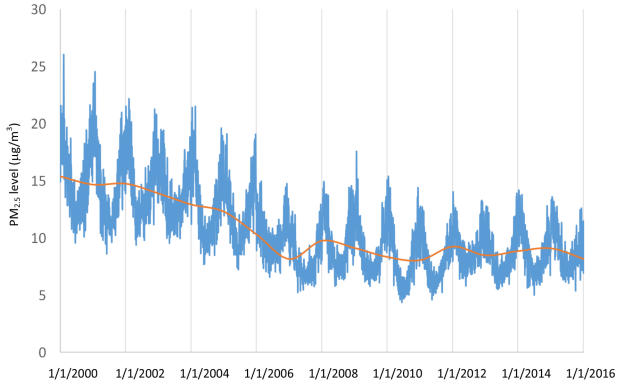
Relationship between Monitored and Predicted $PM_{2.5}$ from the Ensemble Model and Three Machine Learning Algorithms



Spatial Distribution of Predicted $PM_{2.5}$



Temporal Trend of $PM_{2.5}$



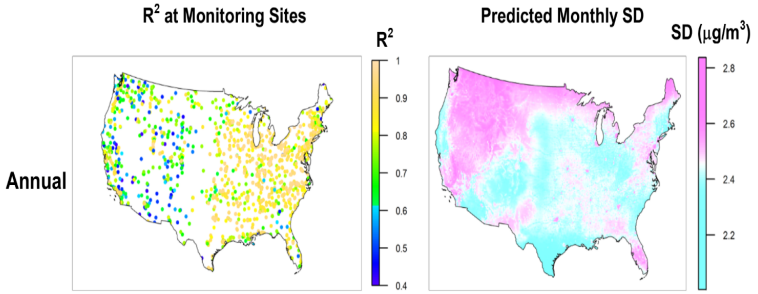
Results

- Model training results from 2000 to 2015 indicated good model performance with a 10-fold cross-validated R^2 of 0.86 for daily $PM_{2.5}$ predictions.
- For annual $PM_{2.5}$ estimates, the cross-validated R^2 was 0.89.
- Model demonstrated good performance up to 60 μg/m^3.
- $PM_{2.5}$ concentrations were higher in populous places, such as Los Angeles, and the entire Eastern United States, excluding the Appalachian Mountains and some remote areas in Northern Maine and Florida. The Central Valley of California also had high concentrations.
- $PM_{2.5}$ concentration dropped significantly after 2008.

Conclusions

- $PM_{2.5}$ predictions including the downscaled and uncertainty predictions will allow epidemiologists to accurately estimate the adverse health effect of $PM_{2.5}$
- Based on the model performance of individual learners, we conclude that the model performance of $PM_{2.5}$ is based on context, and best training algorithm to fit $PM_{2.5}$ globally does not exist.

Cross-Validated R^2 at Monitoring Sites and Predicted Monthly Standard Deviation



Acknowledgements

This publication was made possible by U.S. EPA grant numbers RD-834798, RD-835872, and 83587201; HEI grant 4953-RFA14-3/16-4. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the U.S. EPA. Further, the U.S. EPA does not endorse the purchase of any commercial products or services mentioned in the publication.