Using Statistical and Machine Learning Models with Remotely Sensed Data to Estimate PM2.5 in the San Francisco Bay Area

Anna Boser[1] , Mohammad Al-Hamdan[2], Christian White[2]

[1]Bren School of Environmental Science and Management, University of California Santa Barbara
[2]Universities Space Research Association, NASA Marshall Space and Flight Center, Hunstville, AL

Abstract

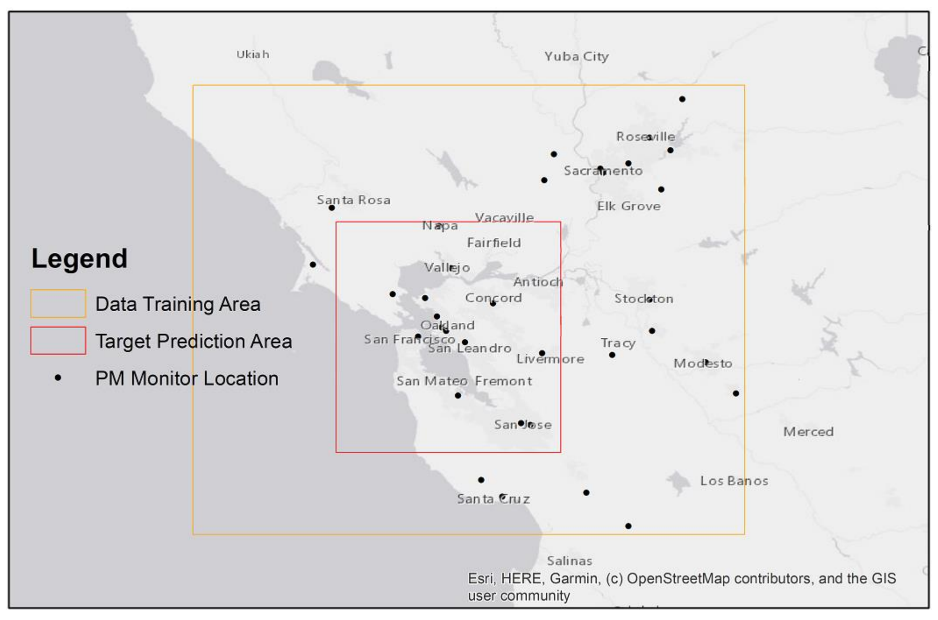
Ambient fine particulate matter ($PM_{2.5}$) is associated with significant adverse health impacts. Continuous, high quality and high resolution $PM_{2.5}$ data has the potential to be greatly useful in public health research and mitigation efforts, but $PM_{2.5}$ monitors are few and unevenly distributed over the landscape. In California, this is of particular concern because catastrophic wildfires have caused and are projected to continue causing episodes of very high levels of $PM_{2.5}$. Previous studies have shown the potential for Aerosol Optical Depth (AOD), meteorological data, and land cover/land use (LCLU) data to estimate $PM_{2.5}$ using a variety of models. However, the most recent research has yet to be applied in the San Francisco Bay Area, where high density episodes of $PM_{2.5}$ were observed in 2017 and 2018. In addition, few studies have taken advantage of flexible and powerful machine learning algorithms to estimate $PM_{2.5}$ levels, especially considering the variety of parameters known to improve such models. This study aims to apply the state of the art $PM_{2.5}$ estimation techniques, including a proven two-stage model trained on AOD, meteorological, and LCLU data, and compare it to promising ML algorithms including random forests and gradient boosted decision trees. We envision that this approach will lead to greatly improved estimation of $PM_{2.5}$ in California, and that more flexible ML techniques will allow for improved results when predicting extreme $PM_{2.5}$ events, such as resulting from a wildfire, which are particularly important for public health research.

Objective

Create an accurate and **high resolution (1km²) dataset** of $PM_{2.5}$ estimates that covers the **San Francisco Bay Area** by training and comparing comprehensive **machine learning and statistical models** using **pertinent parameters**.

Study Area

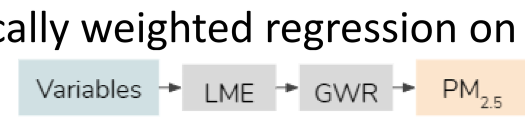
A small target prediction area encompassing the San Francisco Bay Area was selected. This area included 14 EPA $PM_{2.5}$ monitors. In order to increase the training dataset, we designated a larger training area surrounding the target site, which included a total of 34 EPA $PM_{2.5}$ monitors. 327 days in 2017 were included.



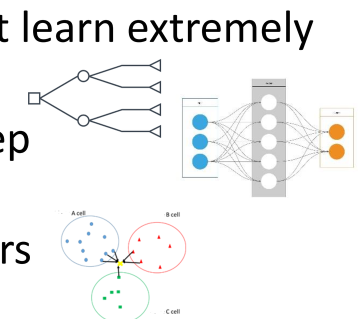
Methodology

Parameters were processed from various publicly available AOD, meteorological, LCLU, and emissions datasets. These data were resampled to a 1 km² grid, and the grid cells containing an EPA Air Quality System (AQS) $PM_{2.5}$ monitor were selected to train and validate the various types of models:

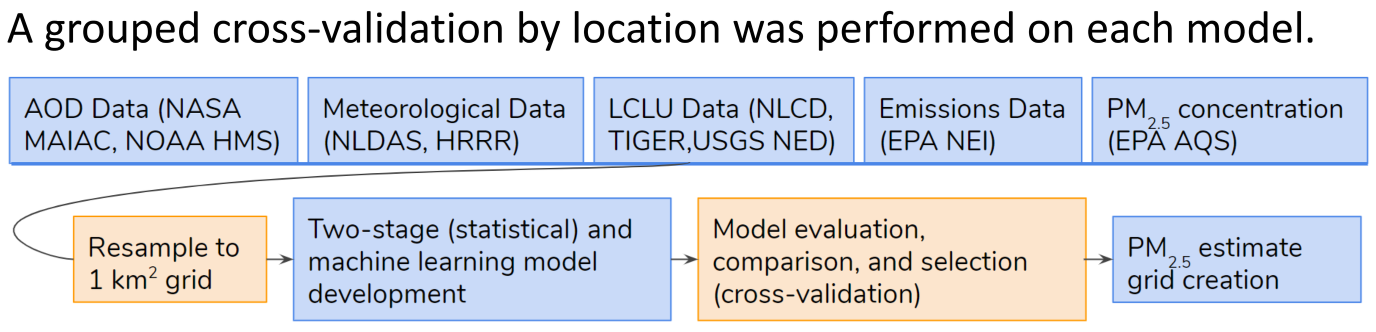
- **Statistical two-stage model.** Stage 1: linear mixed effects model with random effect "Day." Stage 2: geographically weighted regression on aerosol optical depth.



- **Random forests and decision trees.** Popular python libraries were used: CatBoost boosted trees. LightGBM boosted trees (regular and custom), Scikit learn random forest classifier. Scikit learn extremely randomized trees
- **Artificial Neural Networks.** Autogluon tabular deep neural networks (with and without tuning)
- **K Nearest Neighbors.** Scikit learn nearest neighbors (uniform and distance weighted)



A grouped cross-validation by location was performed on each model.



Results

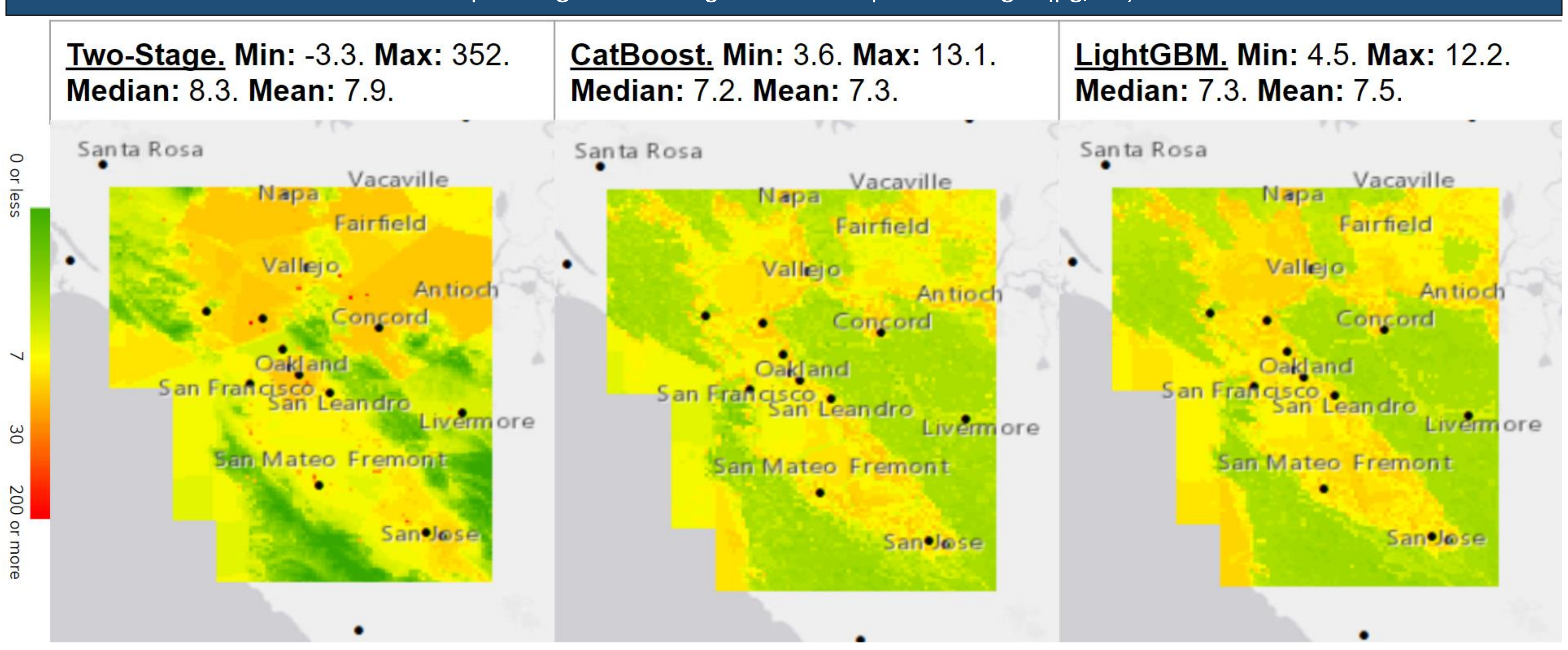
Spatially Grouped Cross-Validated Model Evaluations and Discussion

Table 1: Model Evaluations

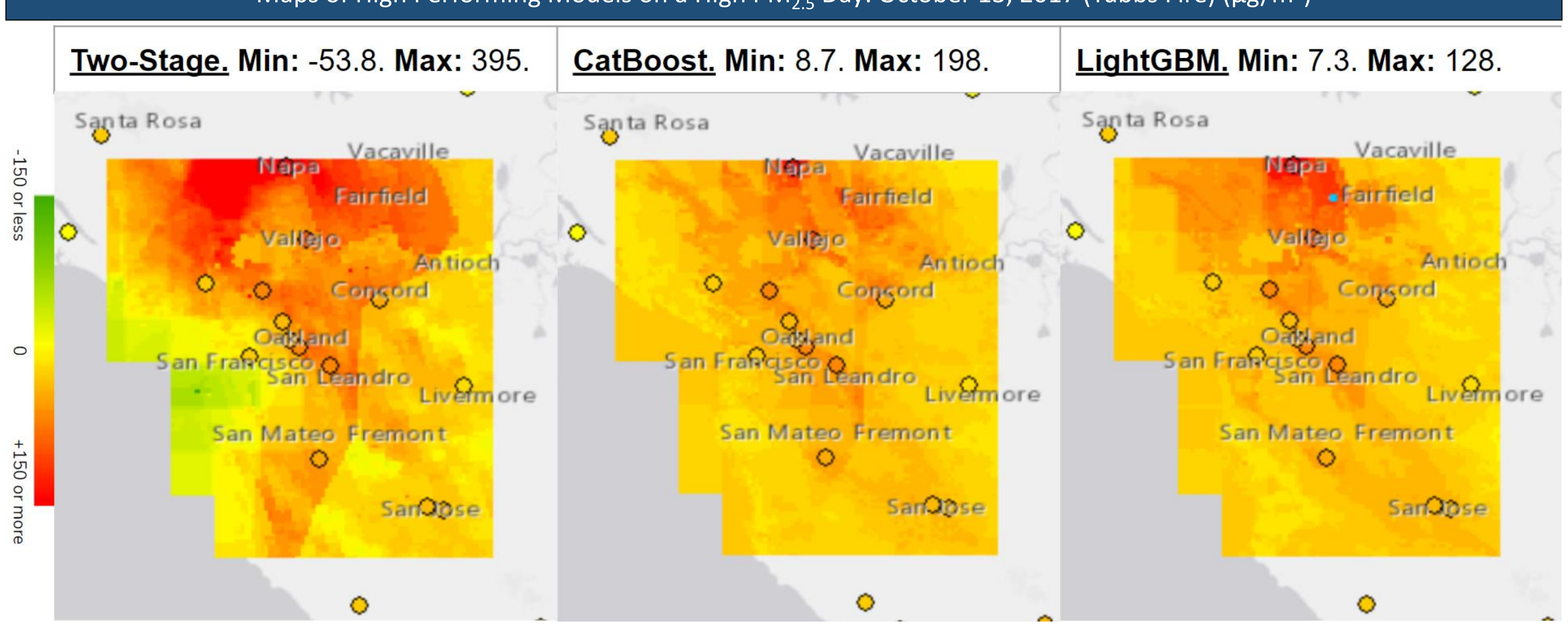
Model	Bias	MAE	MSE	RMSE	NSE	R2
Statistical						
Two-Stage	0.313	2.474	20.939	4.072	0.729	0.827
Random Forests and Boosted Trees						
Boosted trees (CatBoost)	-0.340	2.514	21.025	3.914	0.774	0.856
Boosted trees (LightGBM: custom)	-0.431	2.414	19.752	3.826	0.779	0.850
Boosted trees (LightGBM)	-0.277	2.557	21.004	3.986	0.758	0.835
Extra Randomized Trees (Scikit Learn)	-0.332	2.660	22.005	4.136	0.743	0.819
Random Forest (Scikit Learn)	-0.386	2.636	22.445	4.195	0.732	0.809
Neural Network (Autogluon)						
Neural Network (tuned)	-0.544	3.309	33.098	5.299	0.562	0.662
Neural Network	-0.828	3.252	34.780	5.455	0.546	0.636
Nearest Neighbors (Scikit Learn)						
Nearest Neighbors (distance weighted)	-0.738	4.368	72.553	8.041	0.019	0.160
Nearest Neighbors (uniform)	-0.743	4.391	72.427	8.028	0.025	0.157

- The statistical two-stage model performs well, but certain machine learning methods, especially boosted trees, outperform it.
- Neural networks do not show very good results, but this may be a result of only minimal tuning
- The statistical model extrapolates poorly, both when extreme parameter values (e.g. large point source emissions) and when extreme $PM_{2.5}$ values are present. This is evident when observing the impossibly extreme $PM_{2.5}$ values the model predicts, even when taking the temporal average.
- The boosted trees models appear to handle these non-linear relationships very well.

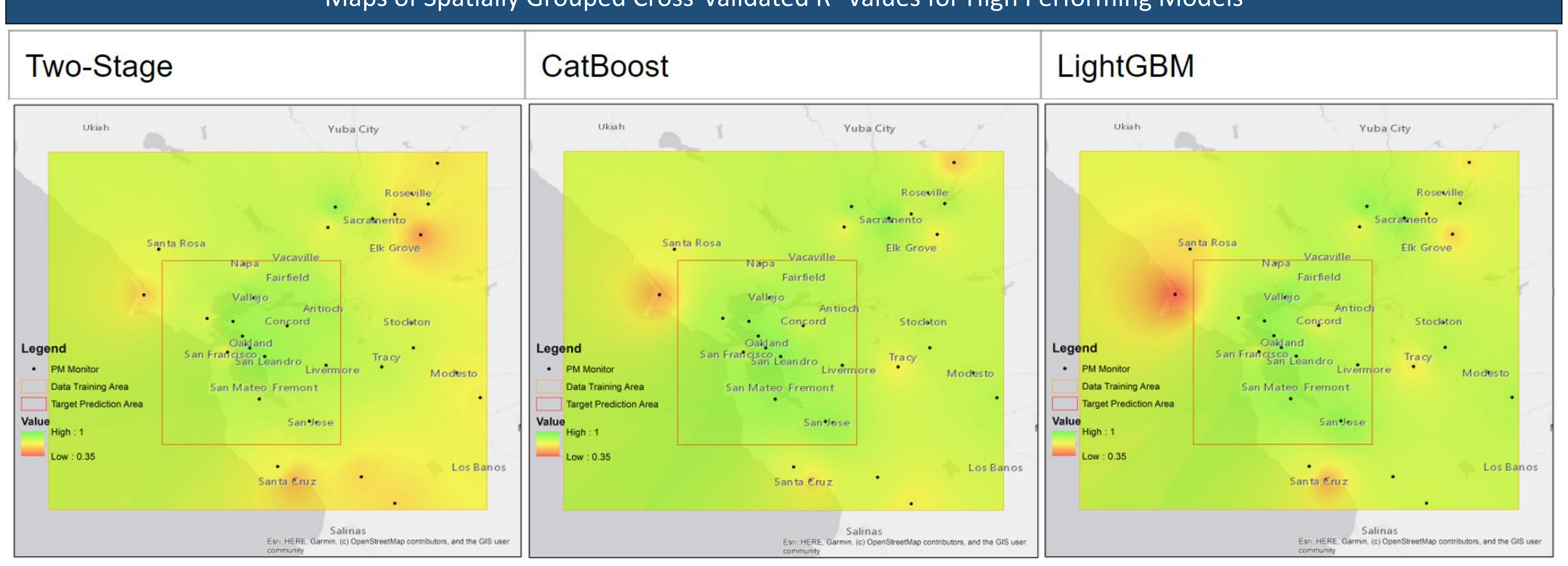
Maps of High Performing Models: Temporal Averages (µg/m³)



Maps of High Performing Models on a High $PM_{2.5}$ Day: October 13, 2017 (Tubbs Fire) (µg/m³)



Maps of Spatially Grouped Cross-Validated R² Values for High Performing Models



Conclusions

- **Machine learning models perform the best.** Certain relationships appear to be strictly non-linear, such as the contribution of point emissions to $PM_{2.5}$ levels.
- **Machine learning models appear more suited to estimate extreme values.** Non-linear models may be better for public health research interested in assessing the effects of a broad range of $PM_{2.5}$ levels in areas without monitors.
- **The buffer zone around the target area is important.** The monitors around the edges of the training area performed worse than those in the middle. This may be an issue for estimating $PM_{2.5}$ adjacent to the coast.
- **More research must be done tuning the different models.** The machine learning models show promising preliminary results, but more research must be done to uncover their full potential.
- **More research must be done on choosing the most pertinent parameters.** This study has shown that a broad array of parameters can lead to promising results, but a more careful analysis of which parameters work best together is needed.
- **A continuous, 1 km² resolution dataset should be made available to public health researchers.** This study suggests that a carefully crafted non-linear model can exhibit the necessary accuracy to be valuable to public health research.

Acknowledgments

Thanks to the **Center for Applied Atmospheric Research and Education** (CAARE), NASA Office of Education's Minority University Research and Education Project (MUREP) Institutional Research Opportunity (**MIRO**) Program, NASA/MSFC/Earth Science Office, NASA/MSFC/Academic Affairs Office, and CAARE Director **Dr. Sen Chiao** from San Jose State University. A special thank you to mentors **Dr. Mohammad Al-Hamdan** and **Christian White** for their timely support, guidance, and encouragement throughout this whole process. We appreciate the help of CAARE interns **Emily Lill** and **Lucas Cohen**, whose input has been invaluable.

References

[1] Delfino, R. J., Brummel, S., Wu, J., Stern, H., Ostro, B., Lipsett, M., Winer, A., Street, D. H., Zhang, L., Tjoa, T., & Gillen, D. L. (2009). The relationship of respiratory and cardiovascular hospital admissions to the southern California wildfires of 2003. *Occupational and Environmental Medicine; London*, *66*(3), 189.
[2] Westerling, A. L., & Bryant, B. P. (2008). Climate change and wildfire in California. *Climatic Change*, *87*(1), 231–249.
[3] Hu, Xuefei, et al. (2014). Estimating Ground-Level PM2.5 Concentrations in the Southeastern United States Using MAIAC AOD Retrievals and a Two-Stage Model. *Remote Sensing of Environment*, vol. 140, pp. 220–232., doi:10.1016/j.rse.2013.08.032.
[4] Zamani Joharestani, M.; Cao, C.; Ni, X.; Bashir, B.; Talebiesfandarani, S. (2019). $PM_{2.5}$ Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. *Atmosphere*, *10*, 373.
[5] Stowell, Jennifer D, et al. (2020). Estimating PM2.5 in Southern California Using Satellite Data: Factors That Affect Model Performance. *Environmental Research Letters*, doi:10.1088/1748-9326/ab9334.
[6] Al-Hamdan, Mohammad Z., et al.(2009). Methods for Characterizing Fine Particulate Matter Using Ground Observations and Remotely Sensed Data: Potential Use for Environmental Public Health Surveillance. *Journal of the Air & Waste Management Association*, vol. 59, no. 7, pp. 865–881., doi:10.3155/1047-3289.59.7.865.
[7] Gupta, P., Doraiswamy, P., Levy, R., Pikelnaya, O., Maibach, J., Feenstra, B., et al. (2018). Impact of California fires on local and regional air quality: The role of a low-cost sensor network and satellite observations. GeoHealth, 2, 172–181. https://doi.org/10.1029/2018GH000136