$PM_{2.5}$ Forecasting Using Machine Learning: Opportunities and Obstacles Across US Landscapes

Helen Chen[1], Angela Anqi Zhou[2], Sarah C. Kavassalis[3] (skavassalis@g.hmc.edu)

[1]Harvey Mudd College, *Claremont, CA, USA*
[2]Scripps College, *Claremont, CA, USA*
[3]Harvey Mudd College, *Hixon Center for Climate and the Environment and Department of Chemistry, Claremont, CA USA*

Motivation

Particulate matter with an aerodynamic diameter of less than 2.5 microns (PM2.5) is noted for its ability to deposit in small airways and alveoli.[1] The long and short-term health effects and climate impacts of aerosols motivate the need for simulation at both high spatiotemporal resolution and long timescales, two needs typically not computationally tractable within the same model. PM2.5 concentration and composition vary regionally, sensitive to diverse precursor emissions, regional primary sources, and meteorological controls on transport and chemistry. While explicit chemical simulation might be the goal, data-built models have shown recent successes in PM2.5 prediction.[2] We have trained a random forest using surface observations of PM2.5 concentration and traditional gridded chemical transport model meteorology and emissions inventories to produce a computationally efficient but large-scale model-compatible aerosol prediction scheme.

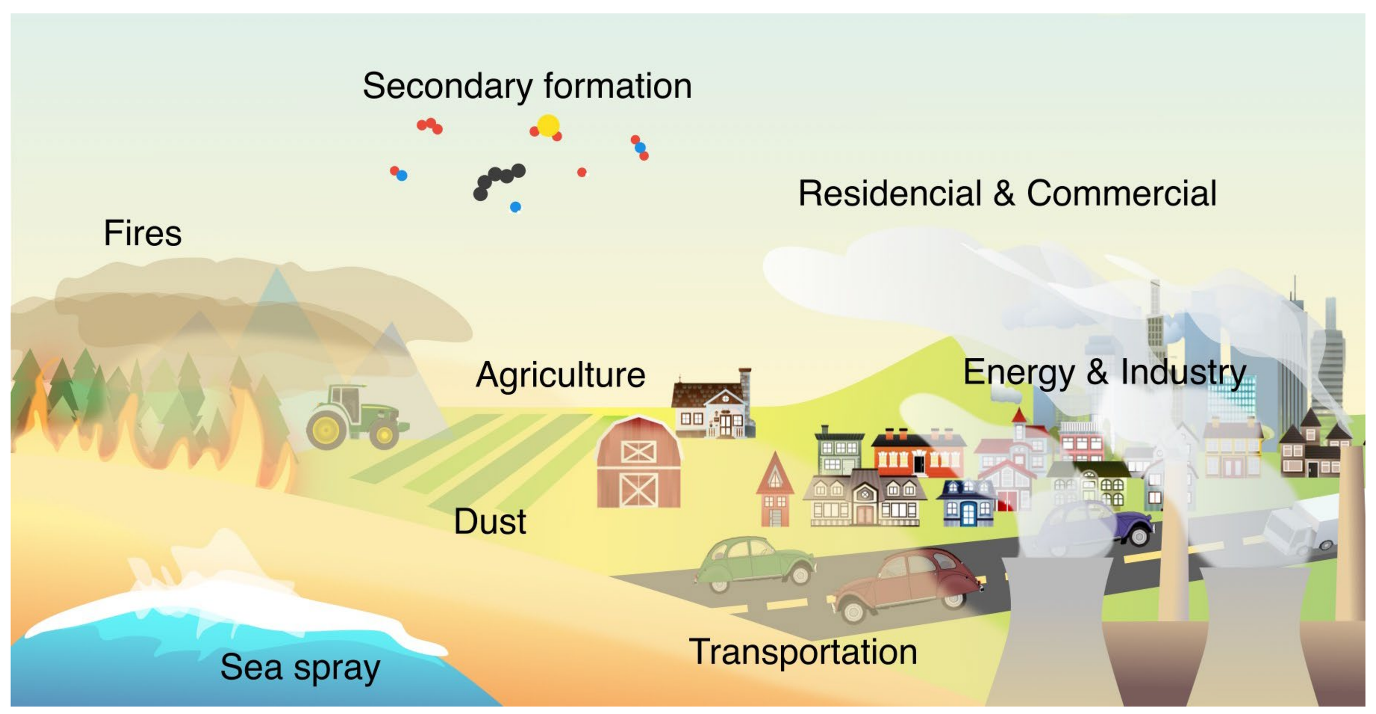
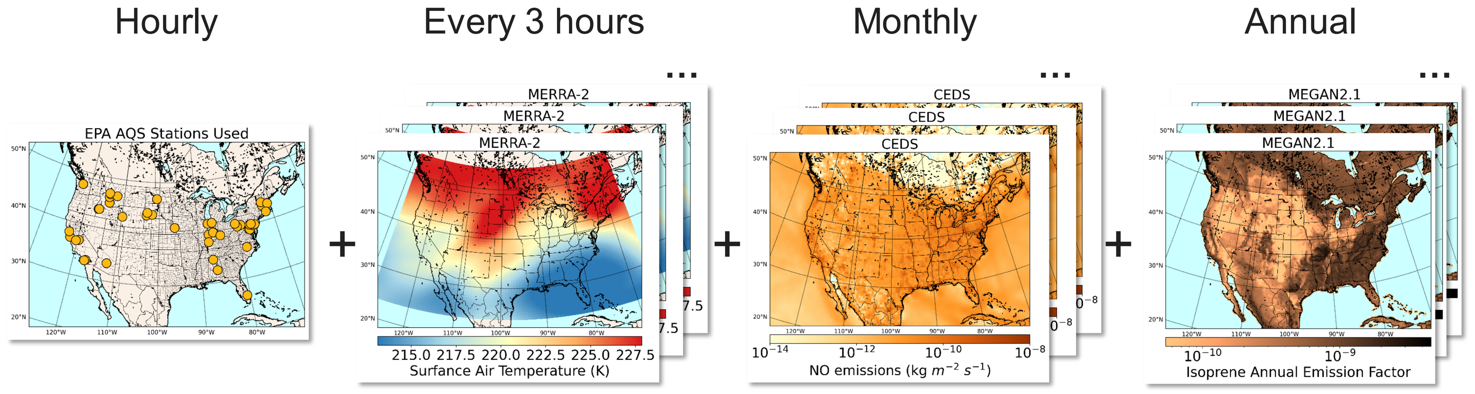


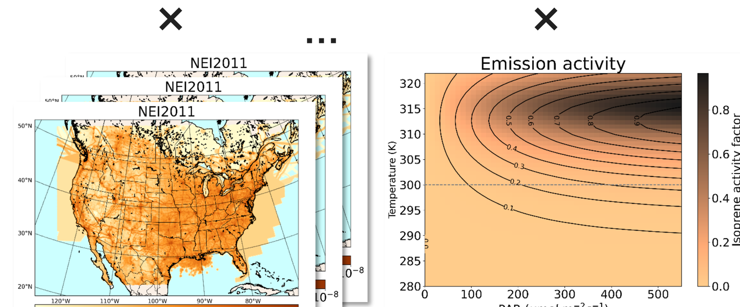
Figure 1. Schematic showing variability and complexity of primary and secondary sources of PM2.5 in the US. Representing all of these sources and chemical processes in any modelling framework is a challenge, often requiring sacrifices of spatial resolution or length of simulation possible.

Training Data

PM2.5 observations were acquired from the EPA's Air Quality System (AQS) monitoring stations. Only stations with 10 years of PM2.5 data (2010-2019) were included allowing less than 10% missing values. MERRA-2 meteorology and HEMCO emissions (CEDS monthly profiles from 2010-2019 were scaled with NEI2011's hourly and day-of-week profiles). MEGAN2.1 was used for biogenic VOC emissions.

PM2.5 Observations	Meteorology	Anthropogenic Emissions	Biogenic VOCs
EPA AQS	MERRA-2	CEDS	MEGAN 2.1
36 sites	0.5°×0.675°	0.5°×0.5°	0.025°×0.03125°
Hourly	Every 3 hours	Monthly	Annual





NEI2011	met. sensitivity
0.1°×0.1°	Guenther et al., 2012
Hourly	

Figure 2. The data sources for our model.

Model

- All features were resampled to be hourly resolution.
- Observation gaps were filled with mean-values.
- The random forest was trained on 7 years of data (2010-2016) and validated on 3 years of data (2017-2019).
- Performance varied across the stations in our dataset (Fig. 4)

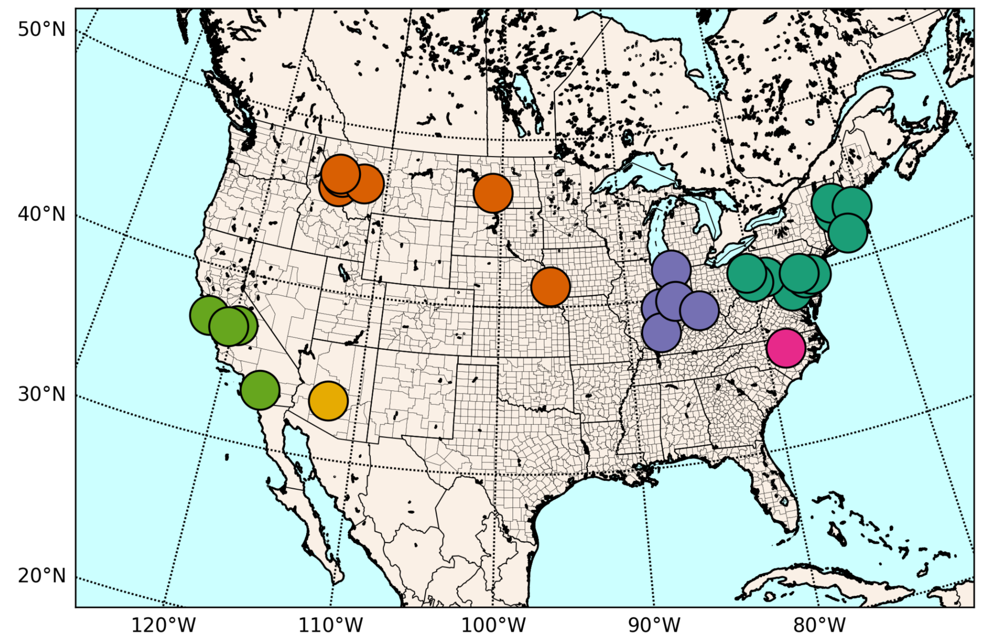


Figure 3. EPA AQS sites with >10 years of PM2.5 measurements; color-coded by climate region.

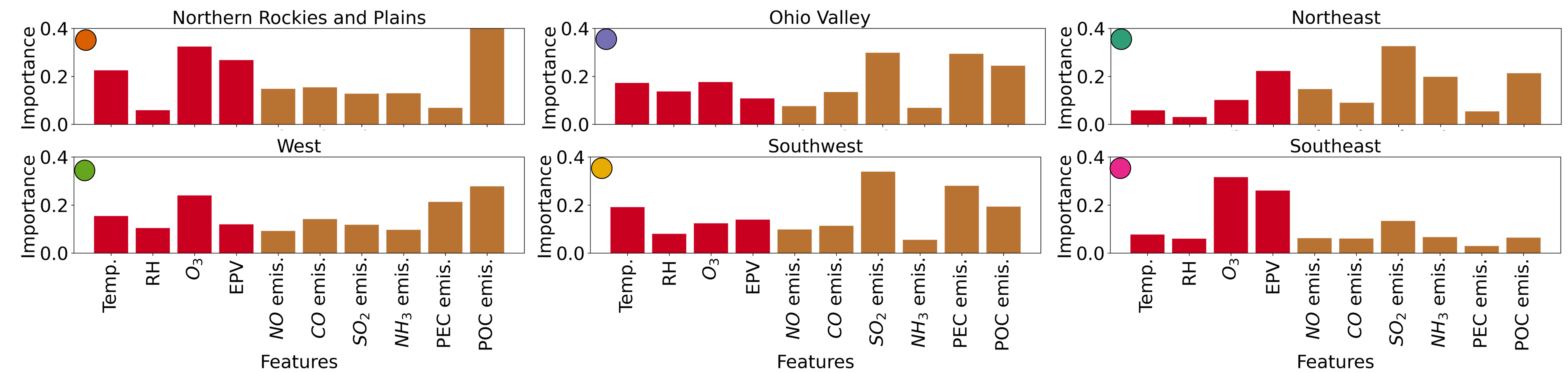


Figure 4. Feature importance plots by climate zone showing only the subset of features that consistently were ranked 'important' to the prediction. Features from MERRA-2 are in red and emissions in brown. Emissions based features vary regionally in ways somewhat consistent with observed sources and composition.[3] The (presumably spurious) importance of Ertel's potential vorticity (EPV) is curious and in some regions explained by poor predictive ability (features aren't meaningfully important when the model performs poorly).

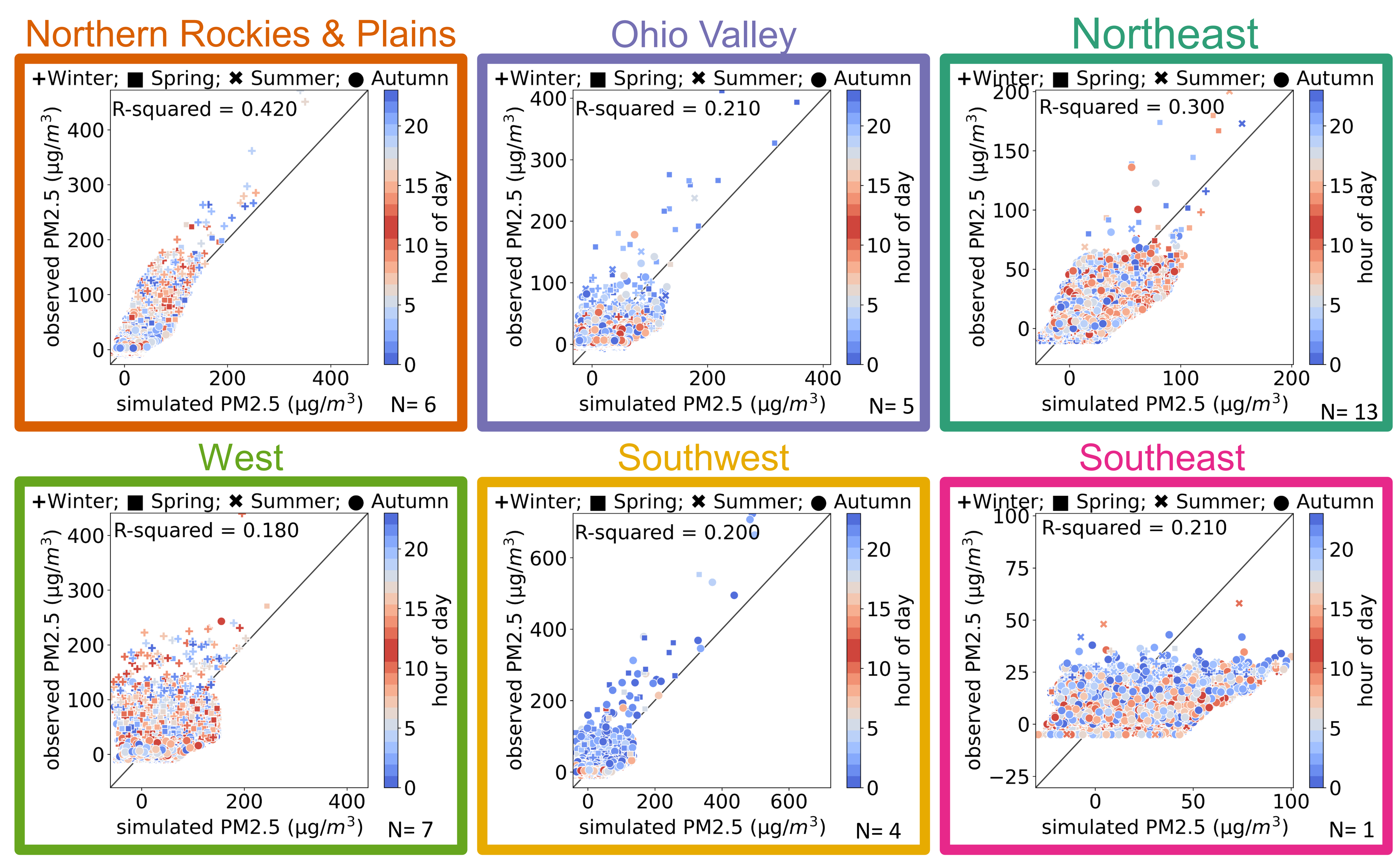


Figure 5. Scatter plots showing hourly observed PM2.5 versus simulated PM2.5 concentration for each station separated by climate zone. Clear regional differences in model performance are apparent. Data points are colored by hour of day and markers indicate seasons to help biases.

Challenges – Infrequent Events

Fire events occur sporadically and are poorly simulated by our model. No feature included in our training data explicitly accounts for fire on PM2.5 and including such features from coarse gridded data is potentially unfeasible.

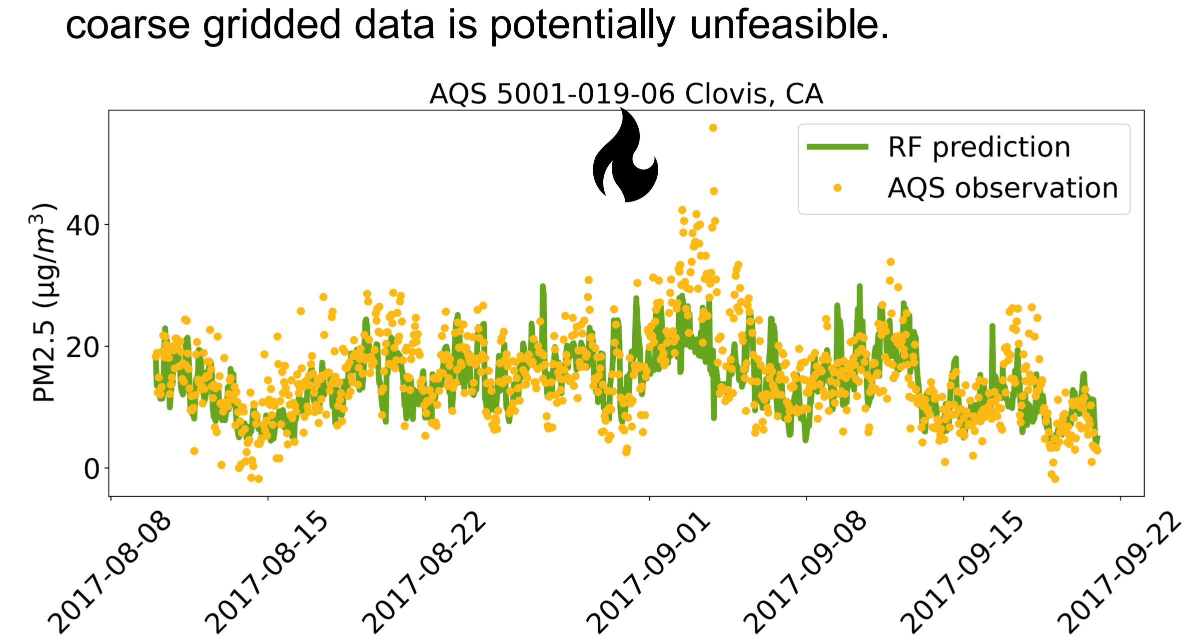


Figure 5. An example time series during the validation period with fire influence. While the basic structure of the observed PM2.5 is captured for typical hours, for the 'wildfire' flagged episode, the model performs poorly.

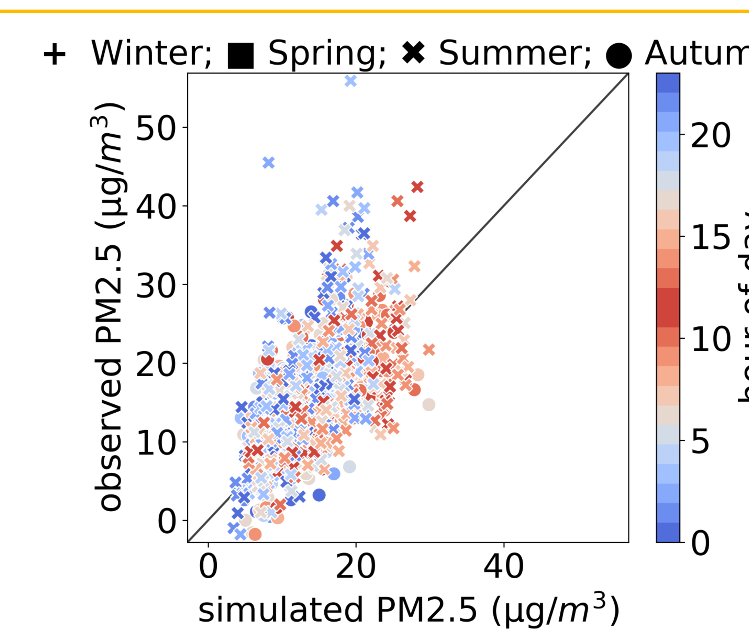
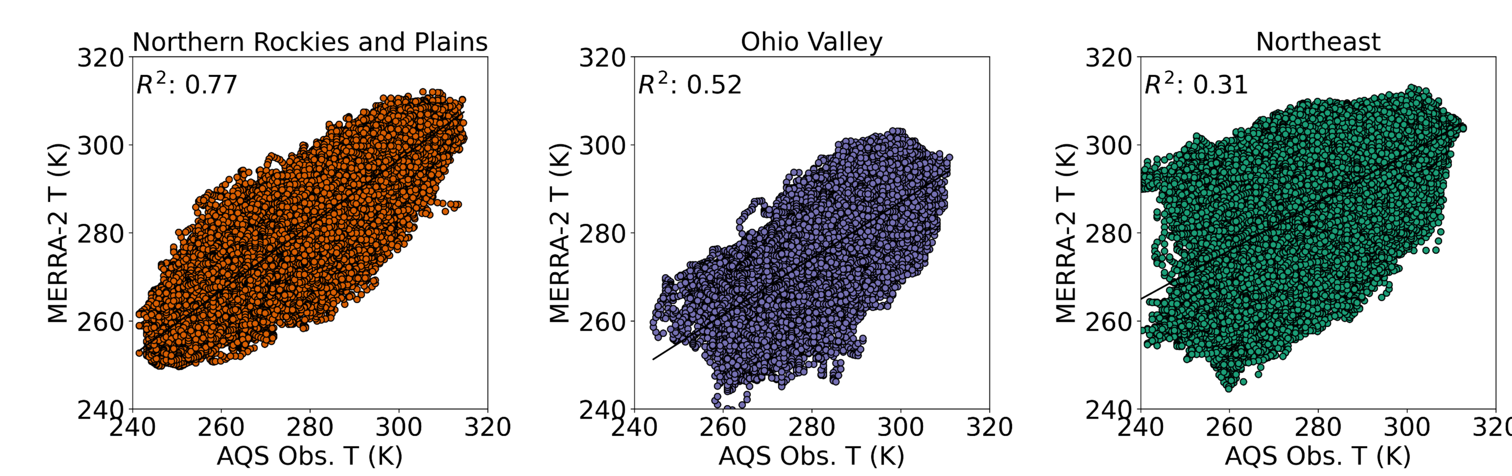


Figure 6. Model-measurement agreement for this window. We see the random forest under-predicts infrequent, high PM2.5 concentrations in summer, aligning with California's fire season.

Challenges – Gridded Meteorology Data

The coarseness of the MERRA-2 dataset meant that assumed meteorology was only an approximation of local conditions. When compared with 2m temperature data directly measured at the stations, we can see some regions are better represented than others.



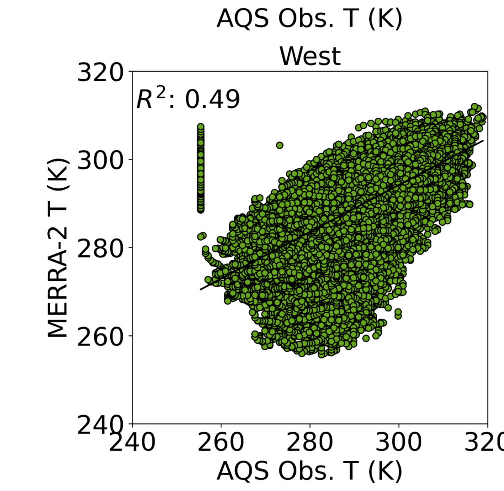
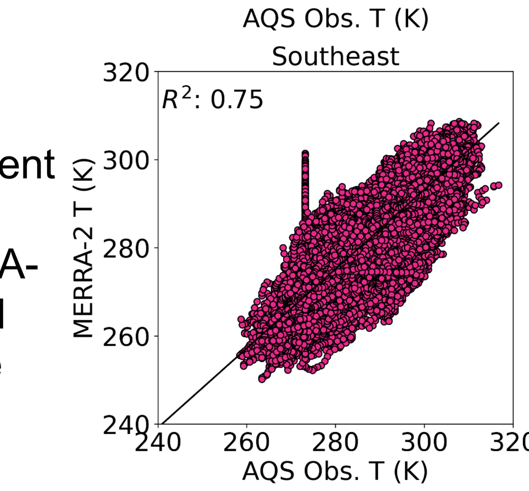


Figure 7. Scatterplots showing station agreement between observed temperature and MERRA-2 meteorology. Regional variation is present here too!



Takeaways

The emissions related feature importances and model performance for non-fire events suggests that the random forest is capturing some of the complex, nonlinear processes regulating PM2.5 and may even be able to provide hints to composition (with much greater computationally efficiently than traditional chemical transport models). The lack of 'fire' related features means this approach cannot be used to simulate events of high public health importance, however.

Acknowledgements

Funding for this research was generously provided by Harvey Mudd College through the Program in Interdisciplinary Computation and the Leeds Student Travel Fund. We would like to thank the Harvey Mudd Chemistry Department and Hixon Center for Climate and the Environment for providing space and computing resources for the project and Bruce and Sharon DePriester for their donation to the FICUS lab.

References: 1) Guarnieri & Balmes, *Lancet*, 2014. 2) Lu et al., *Environ. Res.*, 2021. 3) van Donkelaar et al., ES&T, 2019