

1) PM_{coarse} has major implications in the US, but is understudied¹.

- Sources of coarse particulate matter (PM_{coarse}; particulate matter > 2.5 μm and < 10 μm) include deserts, dry lake beds, agriculture, and construction^{1,2}.
- PM_{coarse} impacts human health, ecosystems, climate, and visibility^{3,4,5,6}.



- In the US, data sources of PM_{coarse} are limited and lack complete spatiotemporal coverage.

2) High spatiotemporal variability necessitates hourly PM_{coarse} estimates, but coverage varies regionally⁷.

- Since PM_{coarse} is not measured directly, there needs to be co-located hourly PM₁₀ and PM_{2.5} monitors to determine PM_{coarse} estimates.
- Some states, such as California, have a denser monitoring network than other states, such as Texas where there are few co-located hourly PM₁₀ and PM_{2.5} monitors.

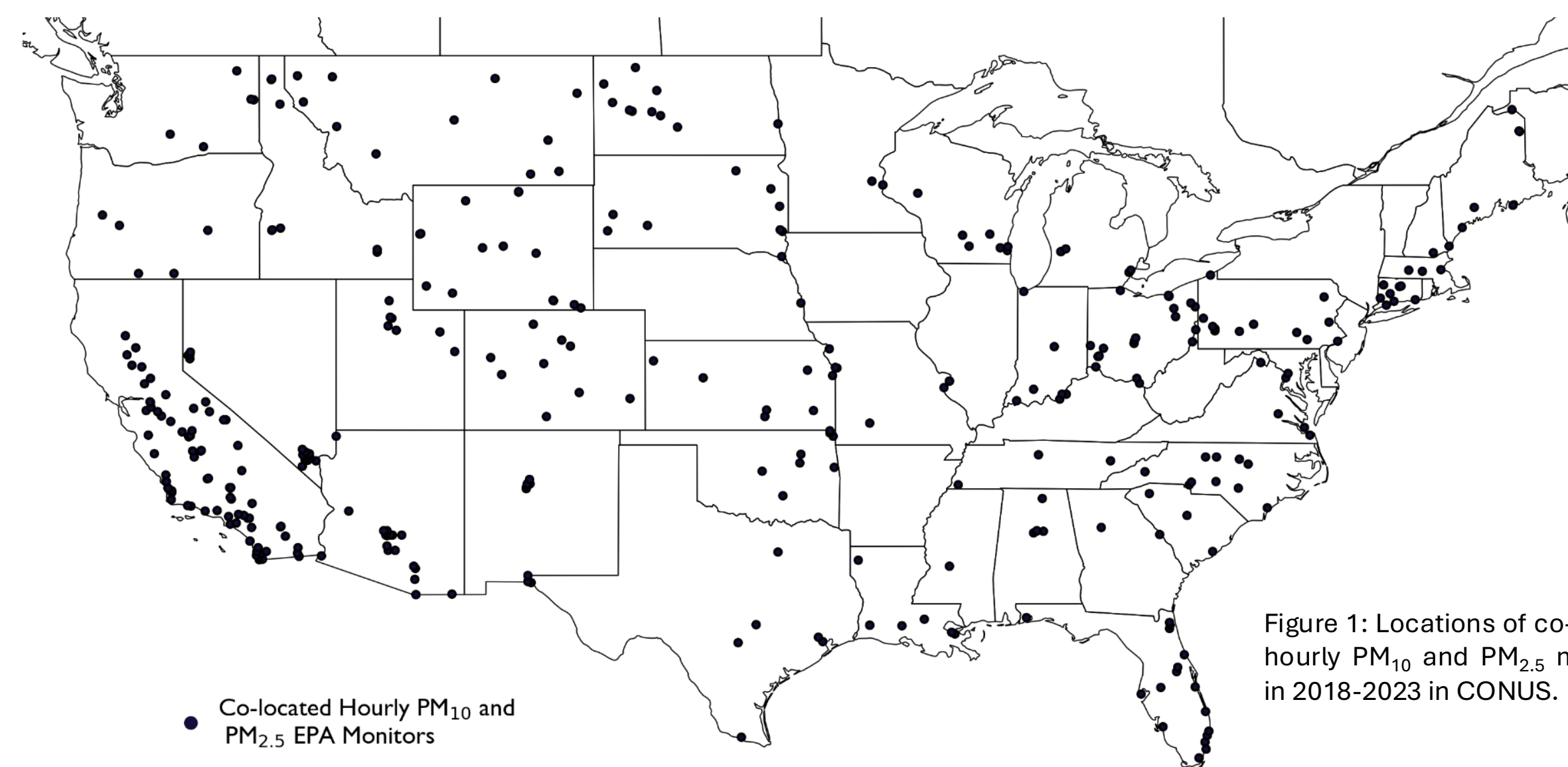
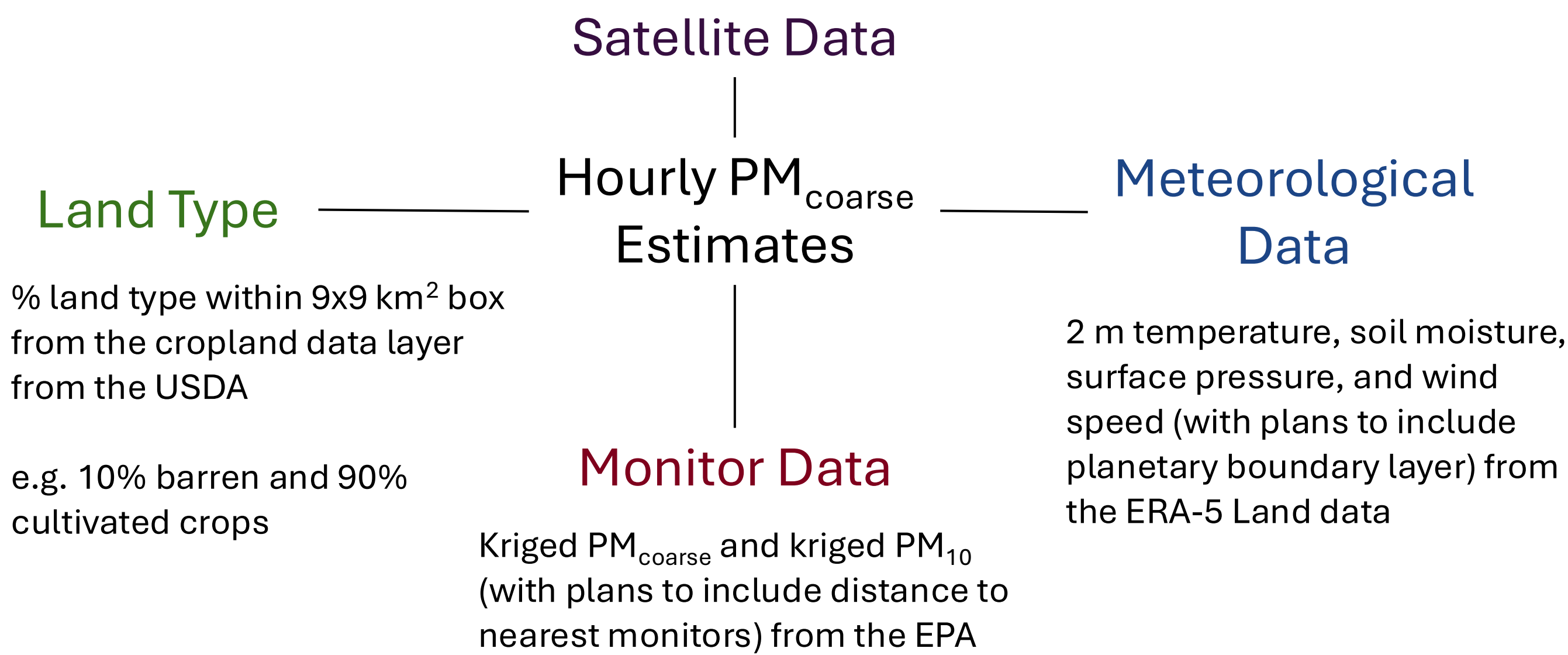


Figure 1: Locations of co-located hourly PM₁₀ and PM_{2.5} EPA Monitors in 2018-2023 in CONUS.

4) Combining multiple data products may help estimate PM_{coarse} concentrations.

Spatial and Temporal data: Latitude, longitude, year, hour, and day of year

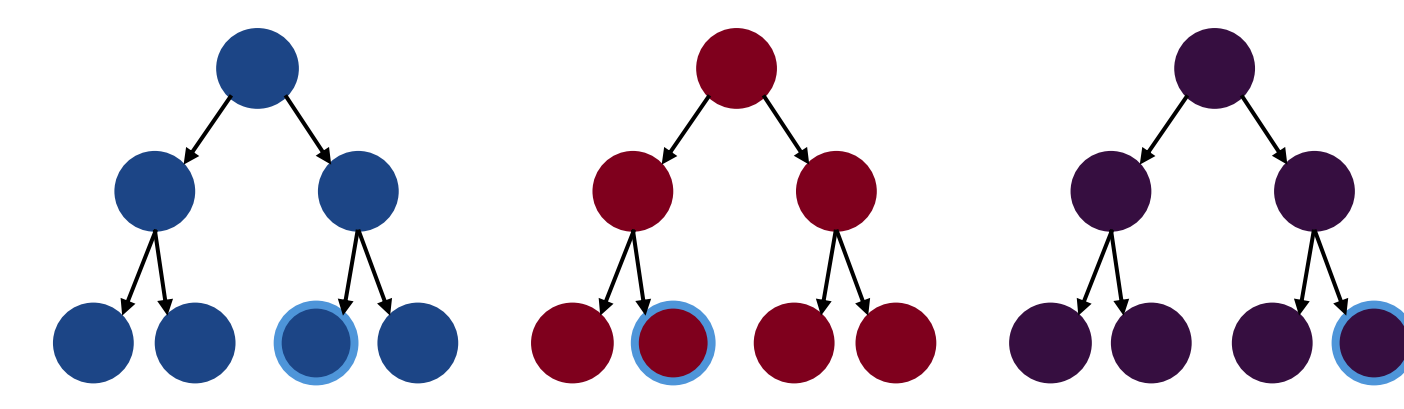
Pink dust index (long wave-based satellite algorithm used for dust detection) and Aerosol Optical Depth from GOES (with plans to include TEMPO)



5) Machine learning, specifically ensemble methods, can help integrate these complex datasets.

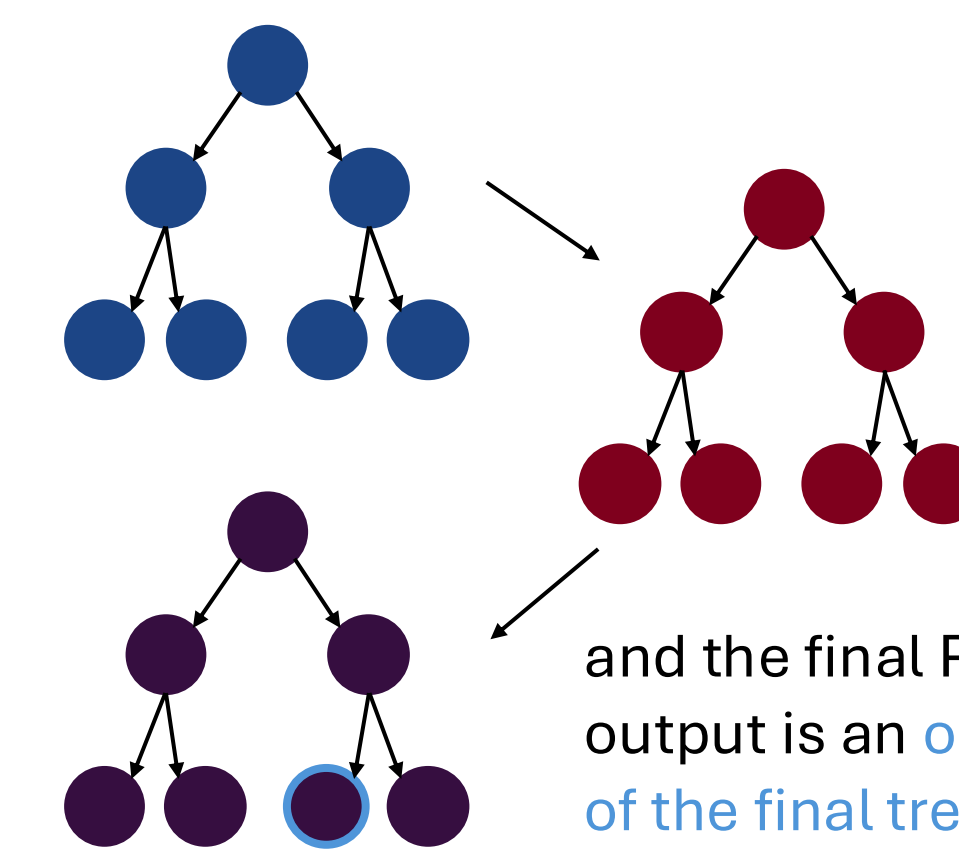
Ensemble methods use decision trees to estimate the predictand⁸.

Random Forests train decision trees in parallel,



and the final PM_{coarse} output is an average of the output of the trees.

Gradient Boosted trees train sequentially to fix the errors of the previous tree,



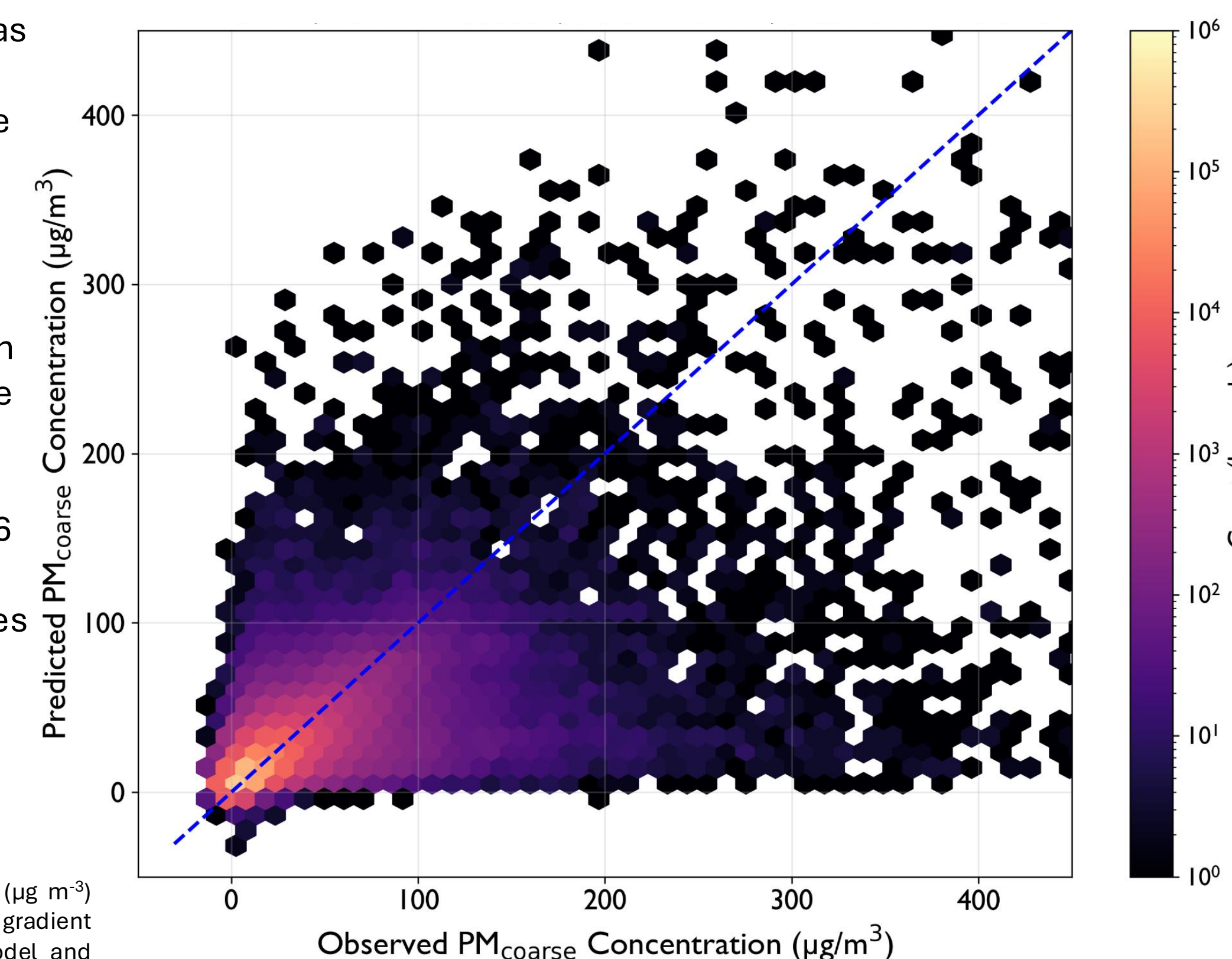
and the final PM_{coarse} output is an output of the final tree.

6) Preliminary work suggests gradient boosted decision trees can help estimate PM_{coarse} observations.

- The gradient boosted model was trained on 70% of the hourly daytime PM_{coarse} observations in 2018-2023
- Here the validation data is plotted (the next 15% of data)
- Maximum depth: 6
- Minimum examples per node: 20

R² = 0.49
RMSE = 16 μg m⁻³

Figure 3: Heatmap of PM_{coarse} (μg m⁻³) estimates predicted by gradient boosted machine learning model and observed PM_{coarse} (μg m⁻³).



3) Interpolating surface monitors does not capture the high spatiotemporal variability of PM_{coarse} concentrations.

- Hourly PM_{coarse} observations in 2018-2023
- Ordinary kriging was evaluated through leave-one-out cross validation
- Only data used in the validation data in cell 6 are shown here
- **R² = 0.18**
- **RMSE = 30 μg m⁻³**

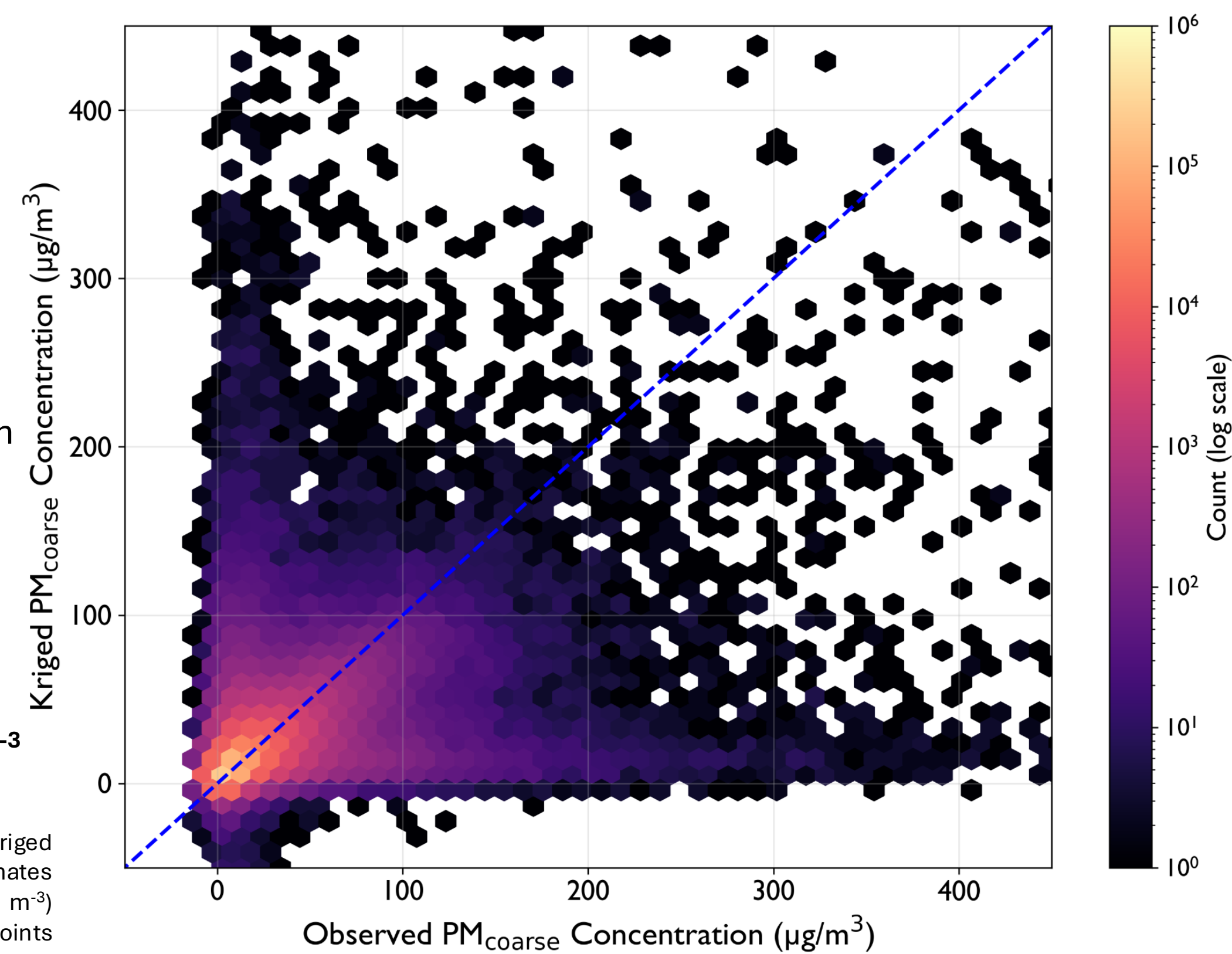


Figure 2: Heatmap of kriged hourly PM_{coarse} (μg m⁻³) estimates and hourly PM_{coarse} (μg m⁻³) observations for the data points used in Figure 3.

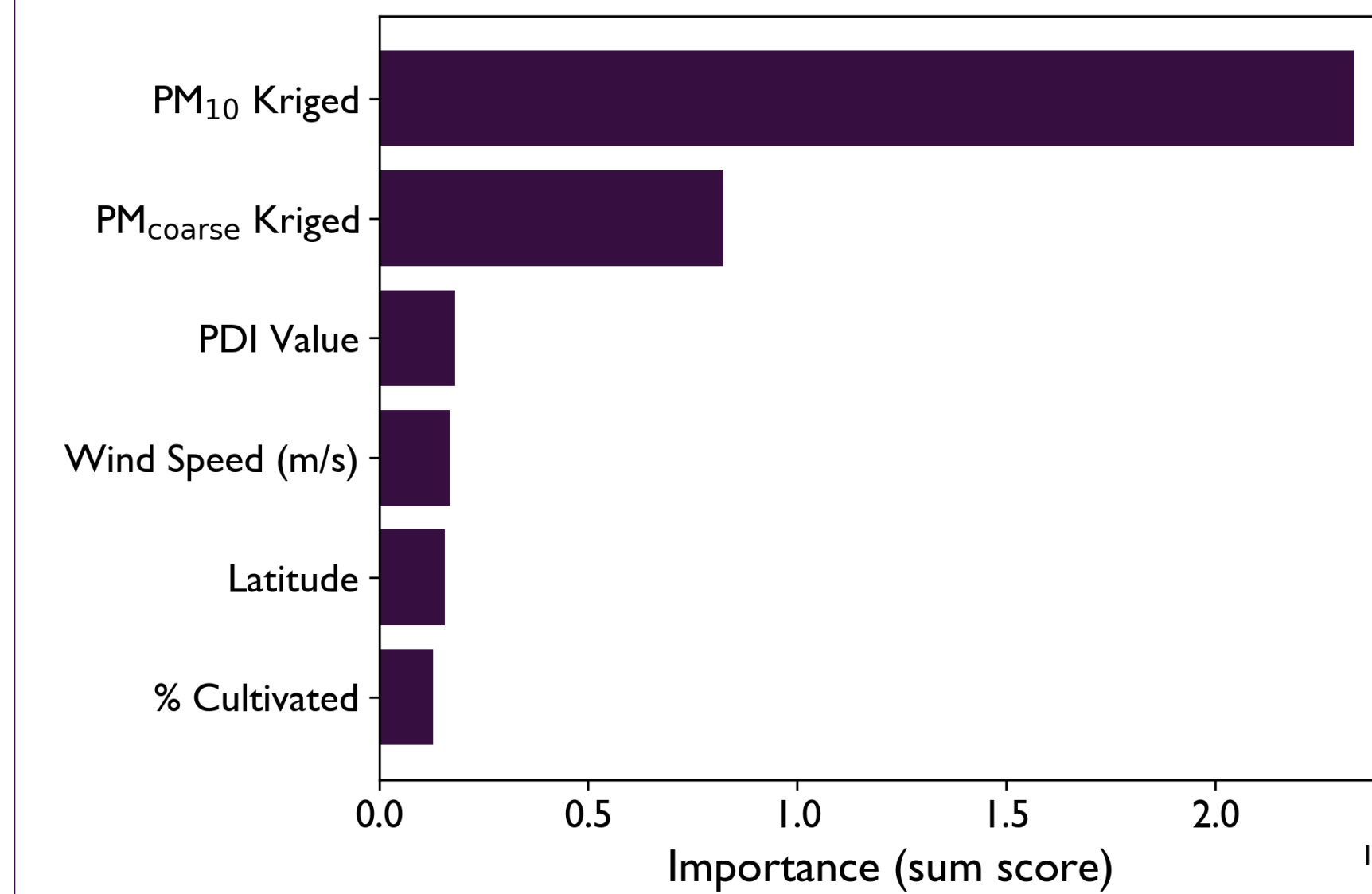
7) Combining multiple data products may help estimate PM_{coarse} concentrations.

- Regions with higher monitor density seem to have better performance in the random forest model
- Regions of higher elevation have worse performance



Figure 4: Scatter plot of the monitor-level Pearson r values across CONUS

8) The machine learning model relies most on monitor data to estimate PM_{coarse}.



- The sum score measures the sum of the split scores using a specific feature (larger scores indicate more important features)⁹
- PM₁₀ is likely more important than PM_{coarse} because of the denser PM₁₀ monitoring network
- The satellite product identifying the presence of dust, PDI value, is next most important in estimating PM_{coarse}

Figure 5: Bar chart of the importances of the top 6 features in predicting PM_{coarse} (μg m⁻³).

Conclusions:

- Machine learning can help integrate many data sources together to produce better estimates of PM_{coarse}.
- In preliminary work, the R² value increased from 0.18 to 0.49 between kriged only estimates, and machine learning predictions.
- So far, regional performance varies.
- Adding more features, such as distance to monitor, additional meteorological data, and TEMPO data, should help.
- The machine learning model relies most heavily on monitor data to predict PM_{coarse}.

Acknowledgements and References:

We would like to thank the NASA Health and Air Quality Applied Sciences Team grants that supported this work: #80NSSC21K0429 and #80NSSC25K0207

- Hand, J. L., Gill, T. E., & Schichtel, B. A. (2019). Urban and rural coarse aerosol mass across the United States: Spatial and seasonal variability and long-term trends. *Atmospheric Environment*, 218, 117025. <https://doi.org/10.1016/j.atmosenv.2019.117025>
- Hand, J. L., Gill, T. E., & Schichtel, B. A. (2017). Spatial and seasonal variability in fine mineral dust and coarse aerosol mass at remote sites across the United States. *Journal of Geophysical Research: Atmospheres*, 122(5), 3080–3097. <https://doi.org/10.1002/2016JD026290>
- Achakulwisut, P., Anenberg, S. C., Neumann, J. E., Penn, S. L., Weiss, N., Crimmins, A., Fann, N., Martinich, J., Roman, H., & Mickley, L. J. (2019). Effects of Increasing Aridity on Ambient Dust and Public Health in the U.S. Southwest Under Climate Change. *GeoHealth*, 3(5), 127–144. <https://doi.org/10.1029/2019GH000187>
- Ahmadzai, H., Malhotra, A., & Tutundjian, S. (2023). Assessing the impact of sand and dust storm on agriculture: Empirical evidence from Mongolia. *PLOS ONE*, 18(2), e0269271. <https://doi.org/10.1371/journal.pone.0269271>
- Yang, D., Zhang, H., Wang, Z., Zhao, S., & Li, J. (2022). Changes in anthropogenic particulate matters and resulting global climate effects since the Industrial Revolution. *International Journal of Climatology*, 42(1), 315–330. <https://doi.org/10.1002/joc.7245>
- Tong, D., Feng, L., Gill, T. E., Schepanski, K., & Wang, J. (2023). How Many People Were Killed by Windblown Dust Events in the United States? *Bulletin of the American Meteorological Society*, 104(5), E1067–E1084. <https://doi.org/10.1175/BAMS-D-22-0186.1>
- Ardon-Dryer, K., & Aziz, T. (2025). Times Matter, the Impact of Convective Dust Events on Air Quality in the Greater Phoenix Area, Arizona. *GeoHealth*, 9(3), e2024GH001209. <https://doi.org/10.1029/2024GH001209>
- GeeksforGeeks. (2026). *Ensemble learning - multiple ML models for better prediction*. [https://www.geeksforgeeks.org/machine-learning/a-comprehensive-guide-to-ensemble-learning/cli-c-user-manual-cli-c-user-manual-ydf-documentation.\(n.d.\)](https://www.geeksforgeeks.org/machine-learning/a-comprehensive-guide-to-ensemble-learning/cli-c-user-manual-cli-c-user-manual-ydf-documentation.(n.d.))
- GeeksforGeeks. (2026). *Ensemble learning - multiple ML models for better prediction*. [https://www.geeksforgeeks.org/machine-learning/a-comprehensive-guide-to-ensemble-learning/cli-c-user-manual-cli-c-user-manual-ydf-documentation.\(n.d.\)](https://www.geeksforgeeks.org/machine-learning/a-comprehensive-guide-to-ensemble-learning/cli-c-user-manual-cli-c-user-manual-ydf-documentation.(n.d.))