Supplementary Information for “Investigating sub-city gradients of air quality: lessons learned with low-cost PM2.5 and AOD monitors and machine learning”

Michael Cheeseman[1], Bonne Ford[1], Zoey Rosen[2], Eric Wendt[3], Alex DesRosiers[1], Aaron J. Hill[1], Christian L'Orange[3], Casey Quinn[3], Marilee Long[2], Shantanu H. Jathar[3], John Volckens[3], Jeffrey R. Pierce[1]

[1]Department of Atmospheric Science, Colorado State University, Fort Collins, 80521, US
[2]Department of Journalism & Media Communication, Colorado State University, Fort Collins, 80521, US
[3]Department of Mechanical Engineering, Colorado State University, Fort Collins, 80521, US

Correspondence to: Michael Cheeseman (cheesemanmj@gmail.com)

Supplementary Tables

Table S1. Description of Environmental Protection Agency’s (EPA) monitoring sites located around Denver, CO, USA.

EPA AQS-ID	Latitude, Longitude	Sampling Frequency	Sample Collection Method	Sample Analysis Method
08-001-0010	39.83, -104.94	24 hour	R & P Model 2025 PM-2.5 Sequential Air Sampler w/VSCC	Gravimetric
08-031-0028	39.79, -104.99	1 hour	GRIMM EDM Model 180 with naphion dryer	Laser Light Scattering
08-031-0026	39.78, -105.01	1 hour	Teledyne T640 at 5.0 LPM	Broadband spectroscopy
08-031-0002	39.75, -104.99	24 hour	R & P Model 2025 PM-2.5 Sequential Air Sampler w/VSCC	Gravimetric
08-031-0027	39.73, -105.02	24 hour	R & P Model 2025 PM-2.5 Sequential Air Sampler w/VSCC	Gravimetric
08-031-0013	39.74, -104.94	1 hour	Teledyne T640 at 5.0 LPM	Broadband spectroscopy
08-005-0005	39.60, -105.02	24 hour	R & P Model 2025 PM-2.5 Sequential Air Sampler w/VSCC	Gravimetric
08-035-0004	39.53, -105.07	24 hour	R & P Model 2025 PM-2.5 Sequential Air Sampler w/VSCC	Gravimetric

Supplementary Figures

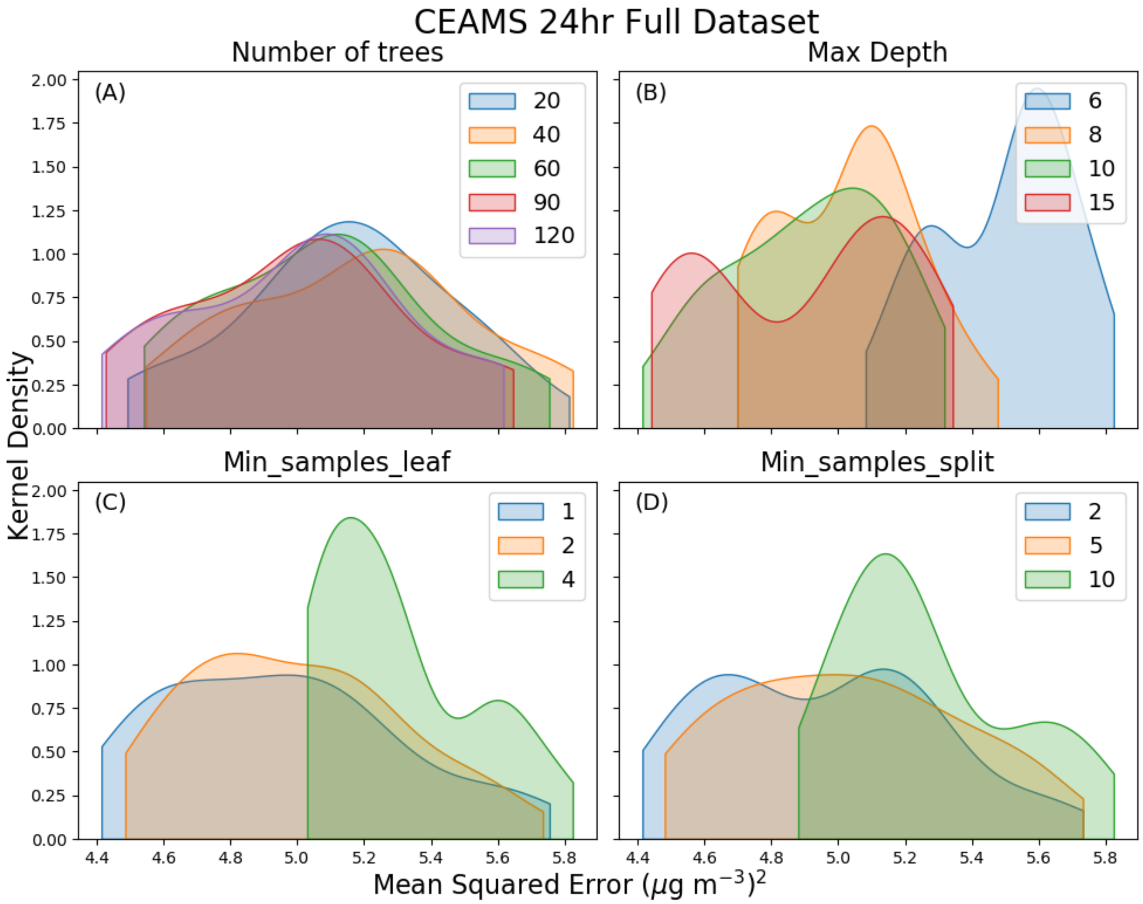


Figure S1. (a) The probability distribution, smoothed using a kernel function, of mean squared errors of model skill for predicting CEAMS 24-hour $PM_{2.5}$ grouped by the number of trees used in each Random Forest model run. The spread of the distribution is due to how the other hyperparameters are changing (see Table 2 for full list) as well as randomness introduced during the cross-validation. Similarly, the other plots show the kernel density distribution of mean squared errors of model skill grouped by (b) the maximum depth of the trees, (c) minimum samples allowed to form a leaf (end of a branch), and (d) minimum samples needed to split an internal node of a tree.

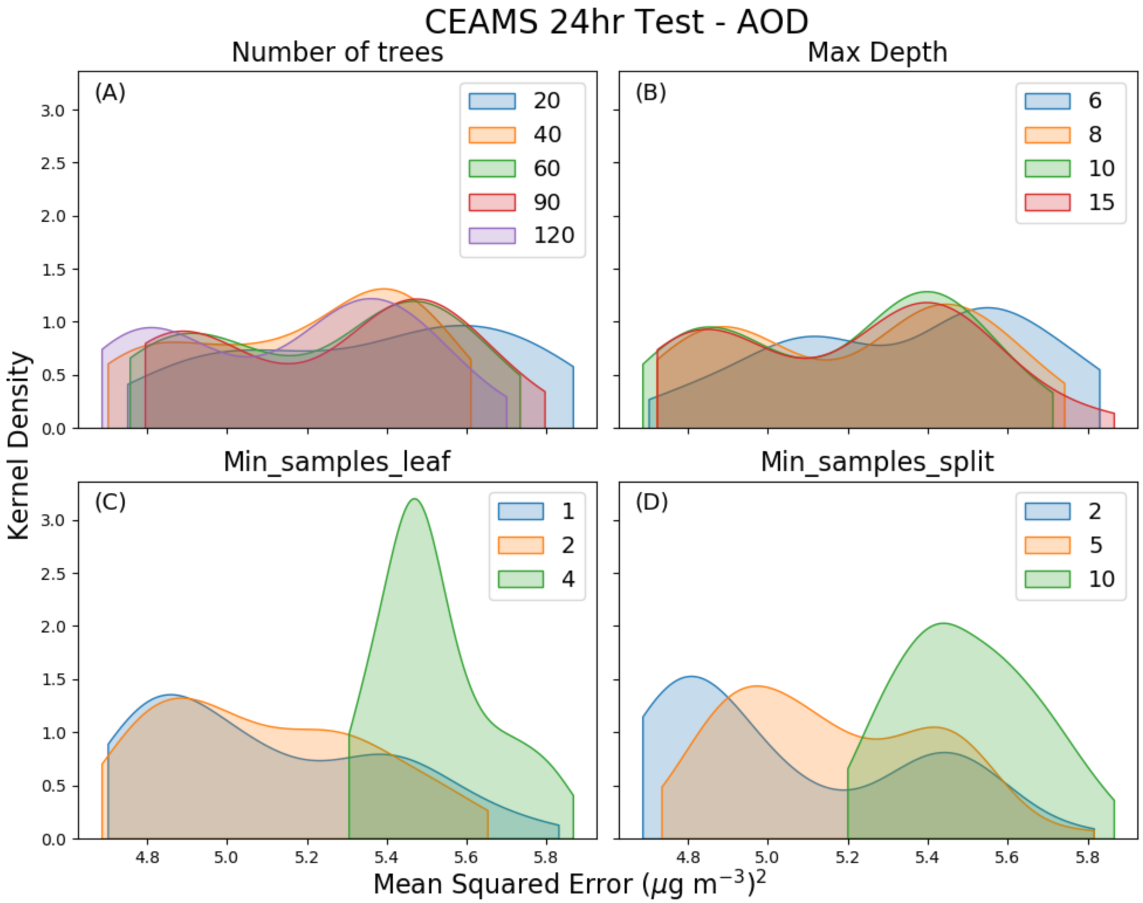


Figure S2 - (a) The probability distribution, smoothed using a kernel function, of mean squared errors of model skill for predicting the "Test - AOD" CEAMS 24-hour $PM_{2.5}$ grouped by the number of trees used in each Random Forest model run. The spread of the distribution is due to how the other hyperparameters are changing (see Table 2 for full list) as well as randomness introduced during the cross-validation. Similarly, the other plots show the kernel density distribution of mean squared errors of model skill grouped by (b) the maximum depth of the trees, (c) minimum samples allowed to form a leaf (end of a branch), and (d) minimum samples needed to split an internal node of a tree.

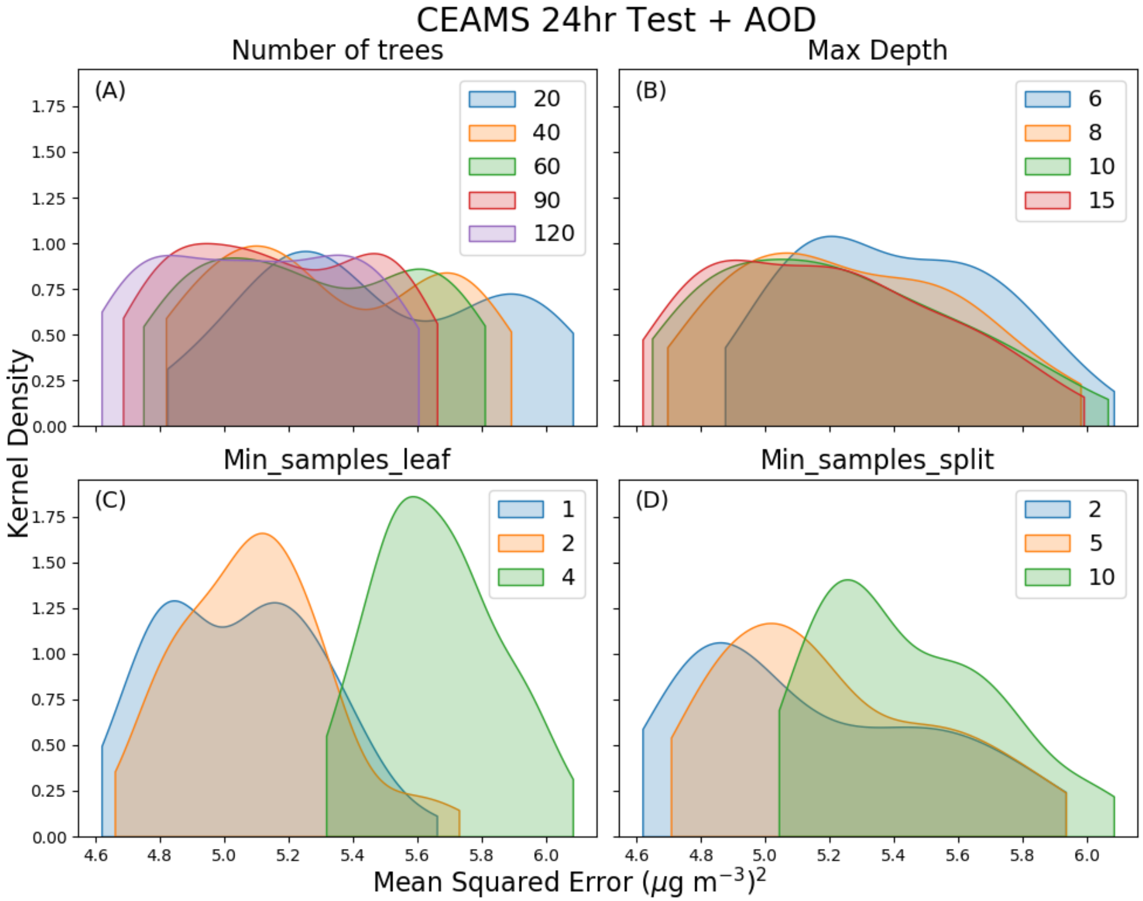


Figure S3 - (a) The probability distribution, smoothed using a kernel function, of mean squared errors of model skill for predicting the "Test + AOD" CEAMS 24-hour $PM_{2.5}$ grouped by the number of trees used in each Random Forest model run. The spread of the distribution is due to how the other hyperparameters are changing (see Table 2 for full list) as well as randomness introduced during the cross-validation. Similarly, the other plots show the kernel density distribution of mean squared errors of model skill grouped by (b) the maximum depth of the trees, (c) minimum samples allowed to form a leaf (end of a branch), and (d) minimum samples needed to split an internal node of a tree.

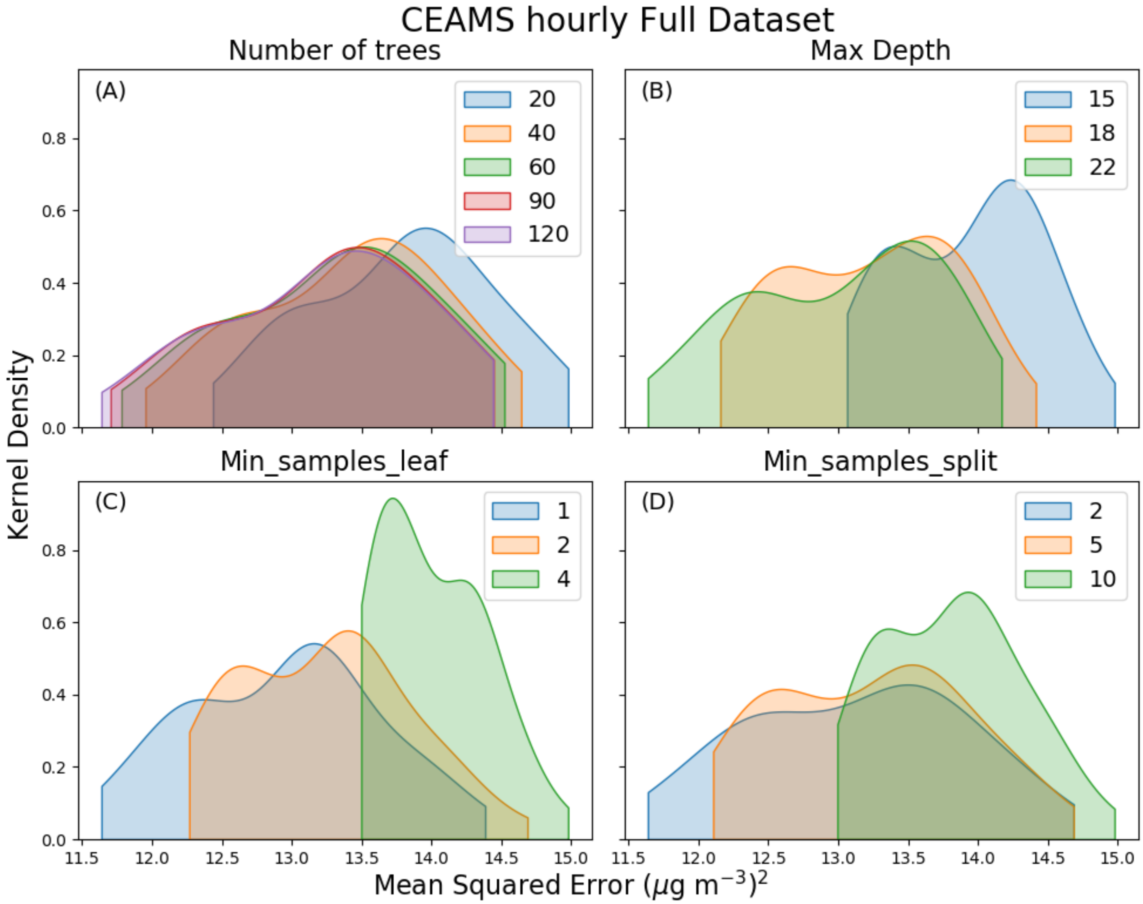


Figure S4 - (a) The probability distribution, smoothed using a kernel function, of mean squared errors of model skill for predicting the "Full dataset" CEAMS hourly $PM_{2.5}$ grouped by the number of trees used in each Random Forest model run. The spread of the distribution is due to how the other hyperparameters are changing (see Table 2 for full list) as well as randomness introduced during the cross-validation. Similarly, the other plots show the kernel density distribution of mean squared errors of model skill grouped by (b) the maximum depth of the trees, (c) minimum samples allowed to form a leaf (end of a branch), and (d) minimum samples needed to split an internal node of a tree.

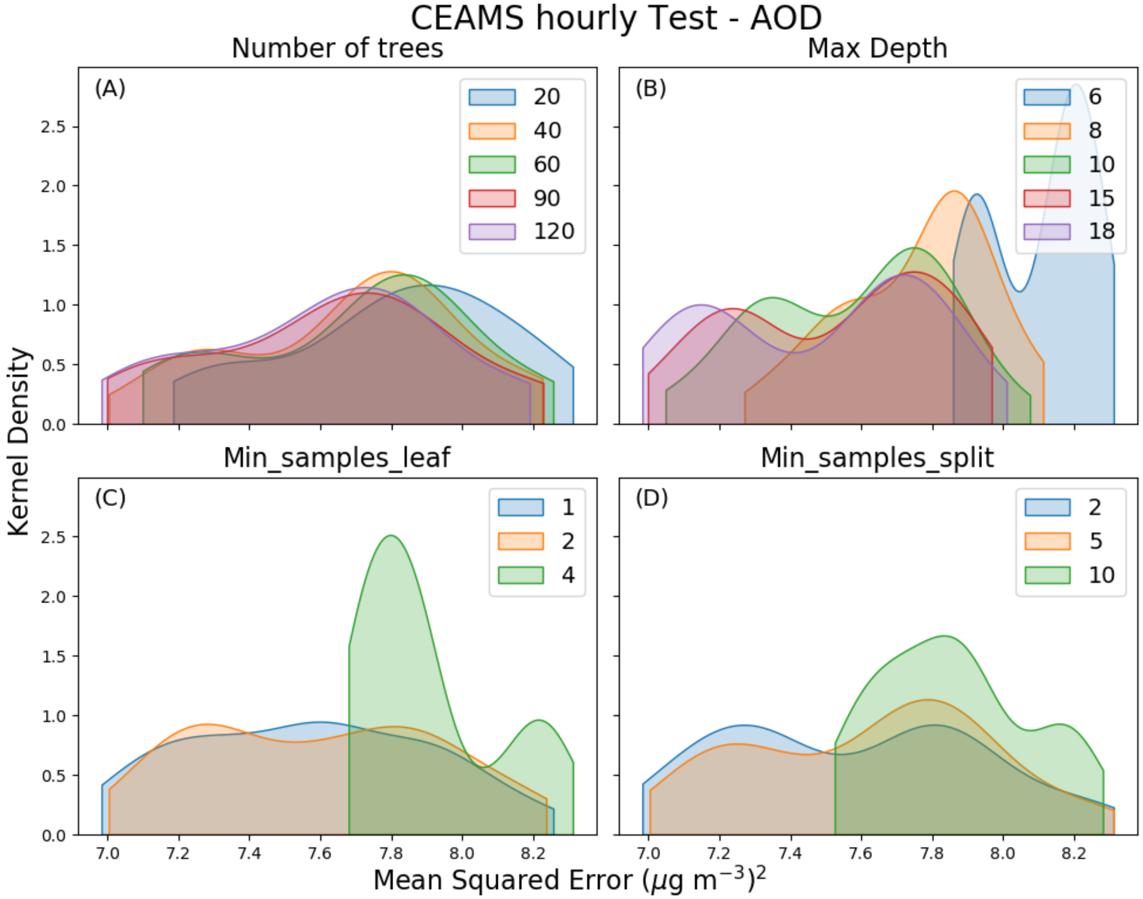


Figure S5 - (a) The probability distribution, smoothed using a kernel function, of mean squared errors of model skill for predicting the "Test - AOD" CEAMS hourly $PM_{2.5}$ grouped by the number of trees used in each Random Forest model run. The spread of the distribution is due to how the other hyperparameters are changing (see Table 2 for full list) as well as randomness introduced during the cross-validation. Similarly, the other plots show the kernel density distribution of mean squared errors of model skill grouped by (b) the maximum depth of the trees, (c) minimum samples allowed to form a leaf (end of a branch), and (d) minimum samples needed to split an internal node of a tree.

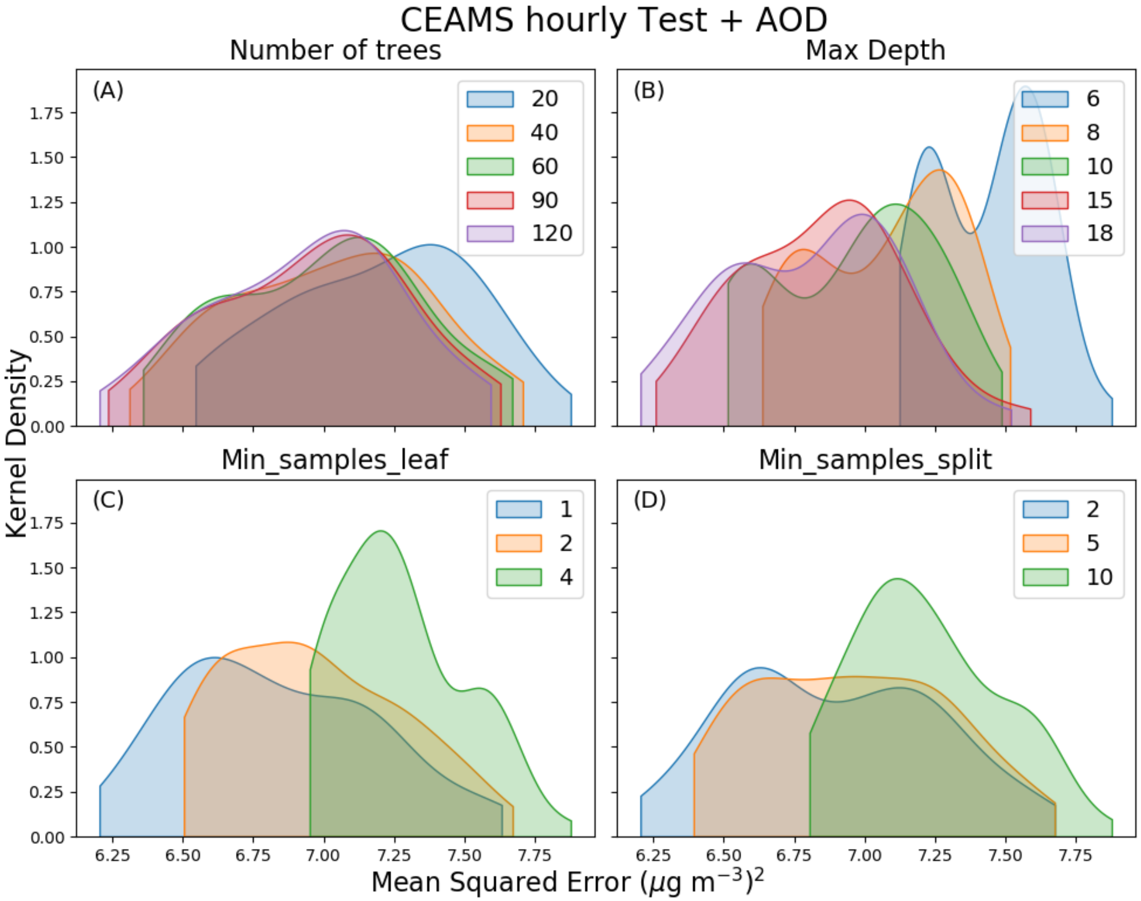


Figure S6 - (a) The probability distribution, smoothed using a kernel function, of mean squared errors of model skill for predicting the "Test + AOD" CEAMS 24-hour $PM_{2.5}$ grouped by the number of trees used in each Random Forest model run. The spread of the distribution is due to how the other hyperparameters are changing (see Table 2 for full list) as well as randomness introduced during the cross-validation. Similarly, the other plots show the kernel density distribution of mean squared errors of model skill grouped by (b) the maximum depth of the trees, (c) minimum samples allowed to form a leaf (end of a branch), and (d) minimum samples needed to split an internal node of a tree.

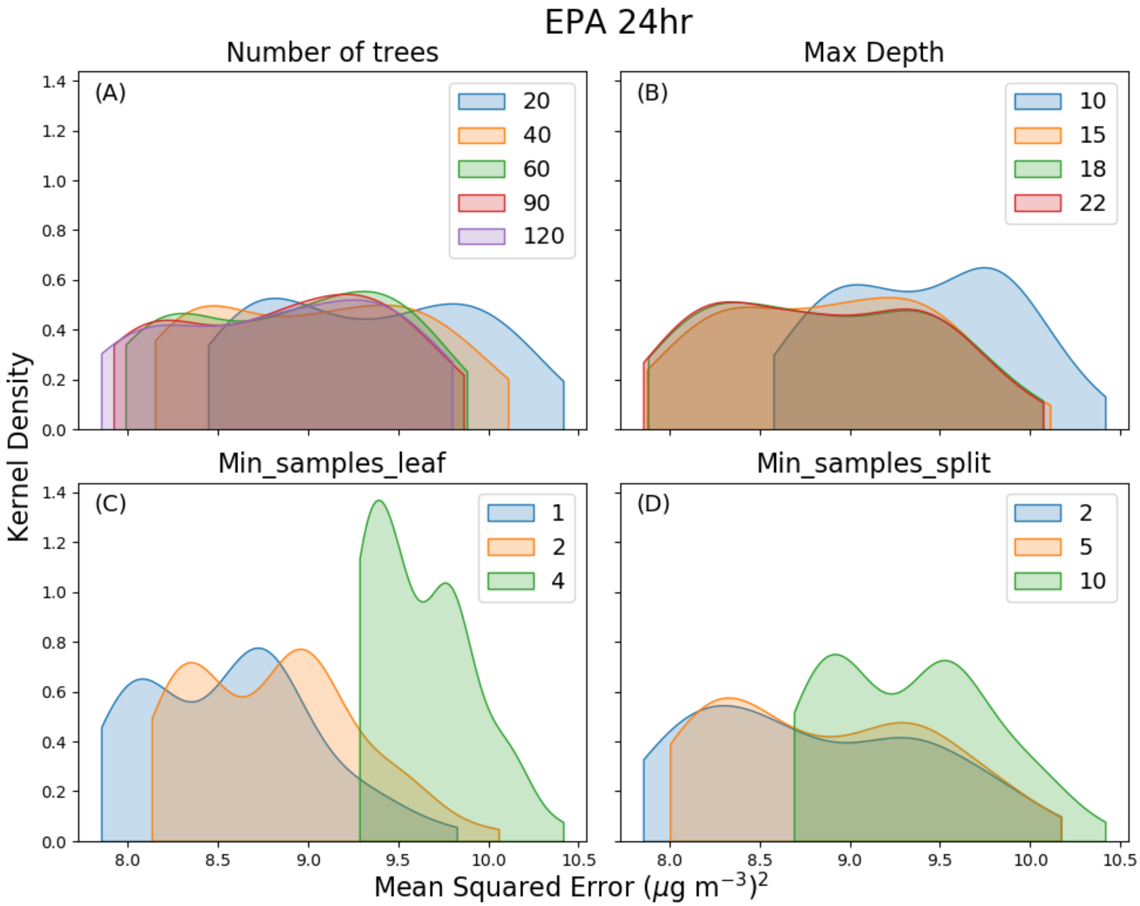


Figure S7 - (a) The probability distribution, smoothed using a kernel function, of mean squared errors of model skill for predicting the EPA 24-hour $PM_{2.5}$ grouped by the number of trees used in each Random Forest model run. The spread of the distribution is due to how the other hyperparameters are changing (see Table 2 for full list) as well as randomness introduced during the cross-validation. Similarly, the other plots show the kernel density distribution of mean squared errors of model skill grouped by (b) the maximum depth of the trees, (c) minimum samples allowed to form a leaf (end of a branch), and (d) minimum samples needed to split an internal node of a tree.

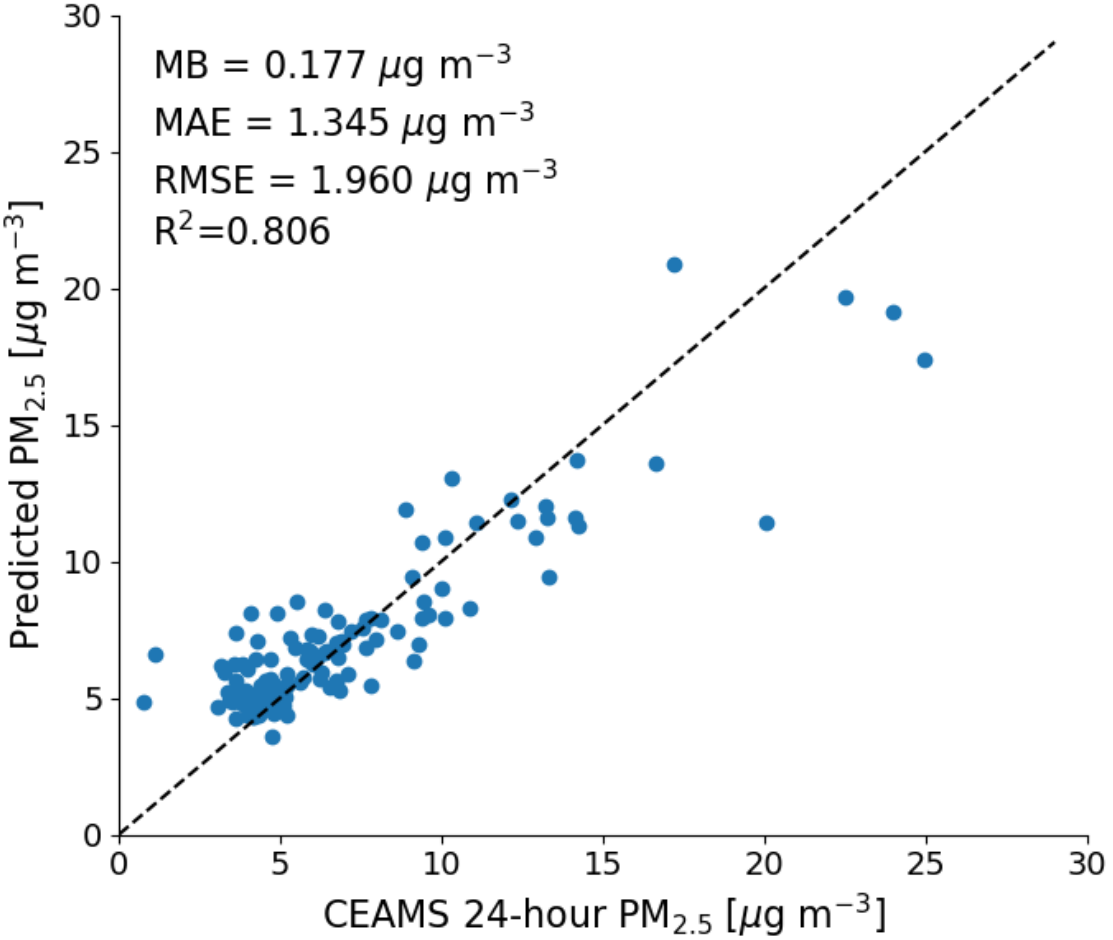


Figure S8. RF 24-hour $PM_{2.5}$ predictions from 1 testing fold during the 5-fold cross validation versus the CEAMS 24-hour $PM_{2.5}$ measurements. The mean bias (MB), mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2) are given.

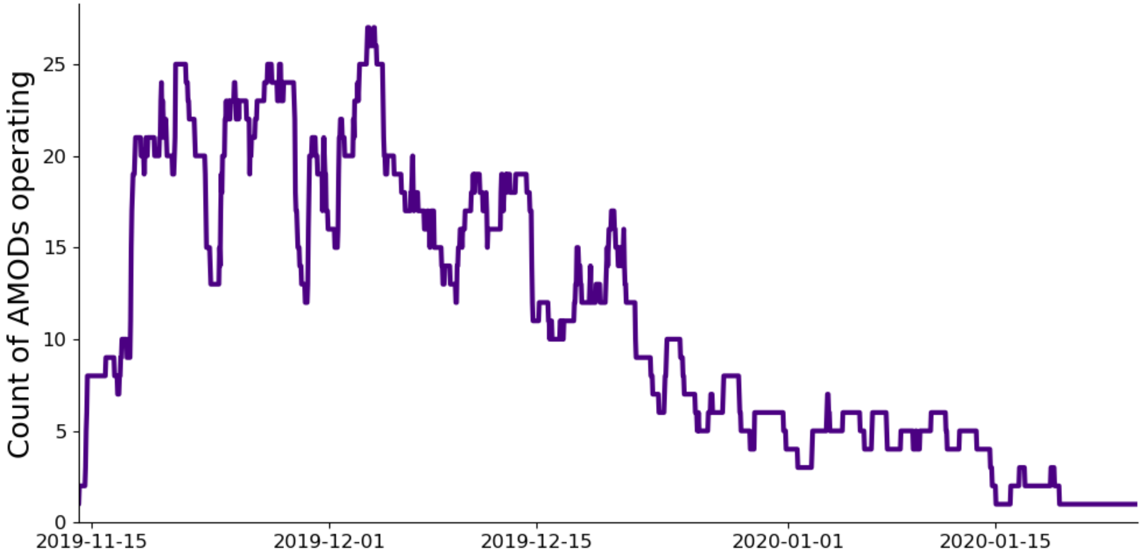
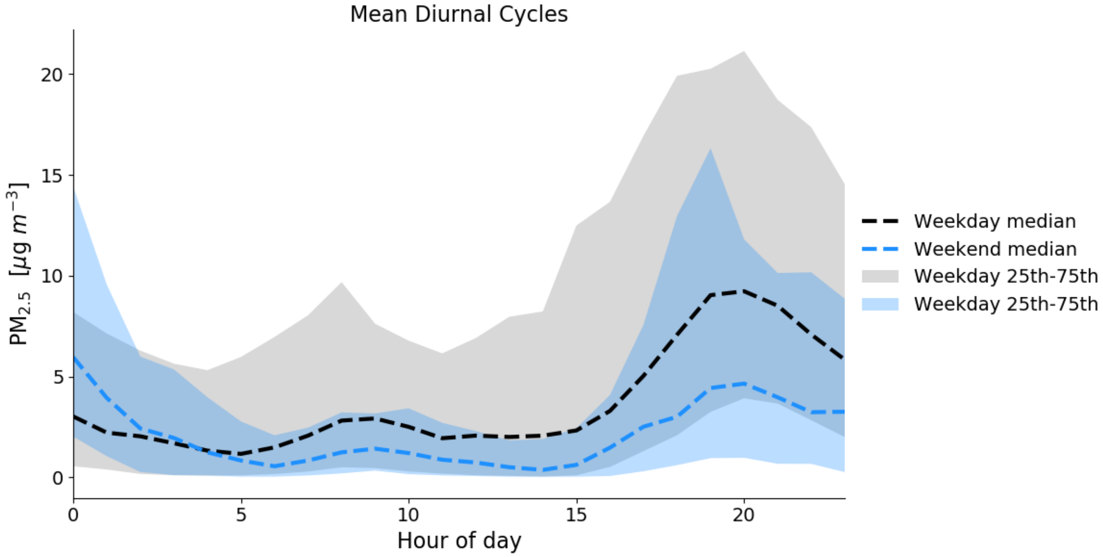
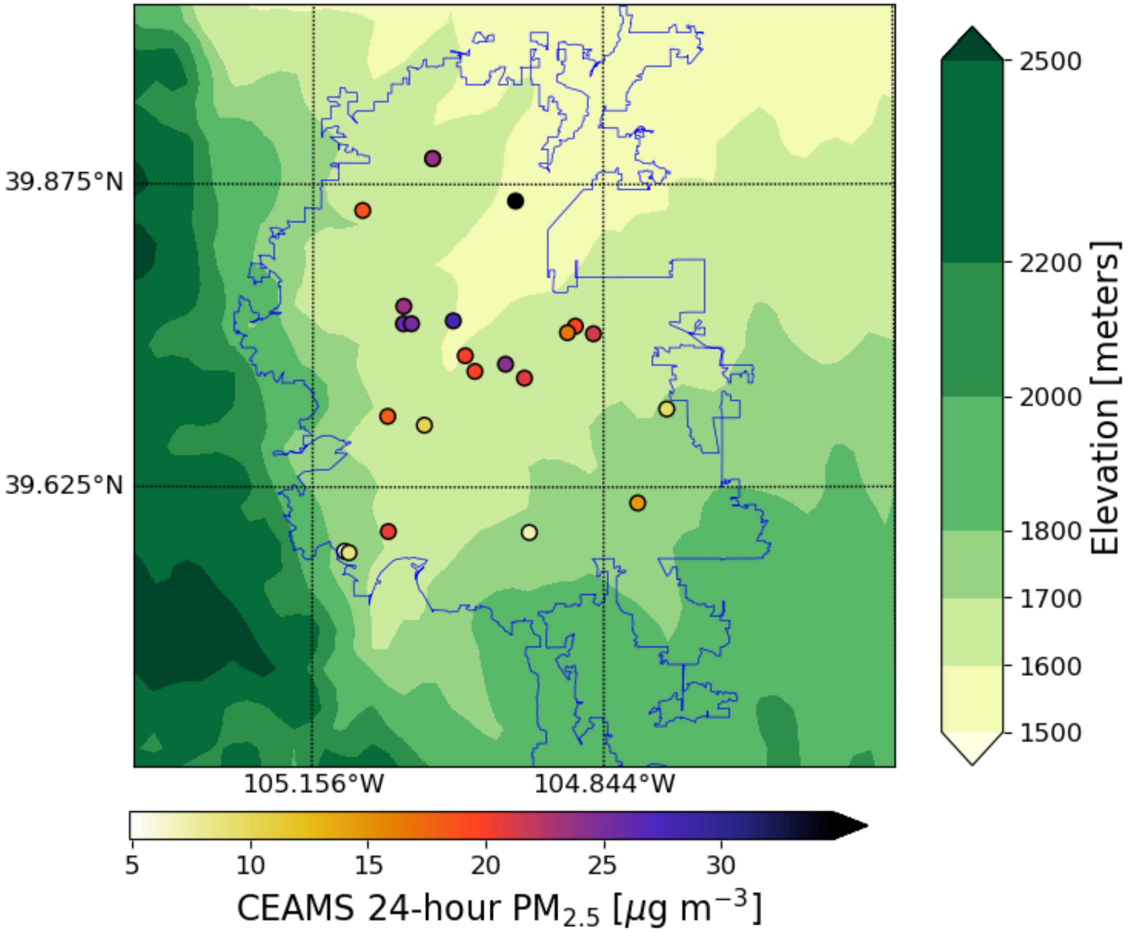


Figure S9. Time series of the number of AMODs operating per hour of the CEAMS deployment in Denver. CO.



80 **Figure S10. Median hourly averaged diurnal cycles of weekend (blue) and weekday (black) PM$_{2.5}$ from the Denver, CO wintertime CEAMS deployment. PM$_{2.5}$ measurements from major holidays were removed. The range between the 25th and 75th percentile of each hourly average is shown for the weekend (blue shading) and weekday (grey shading) averages.**



85 **Figure S11. Map of elevation (Amante and Eakins, 2009) and 24-hour $PM_{2.5}$ averages (points) from the 22 CEAMS low-cost sensors on December 6th, 2019, in Denver, CO. The greater Denver-Aurora area is outlined in blue (based on cartographic files from 2015 TIGER/Line Shapefiles).**

90

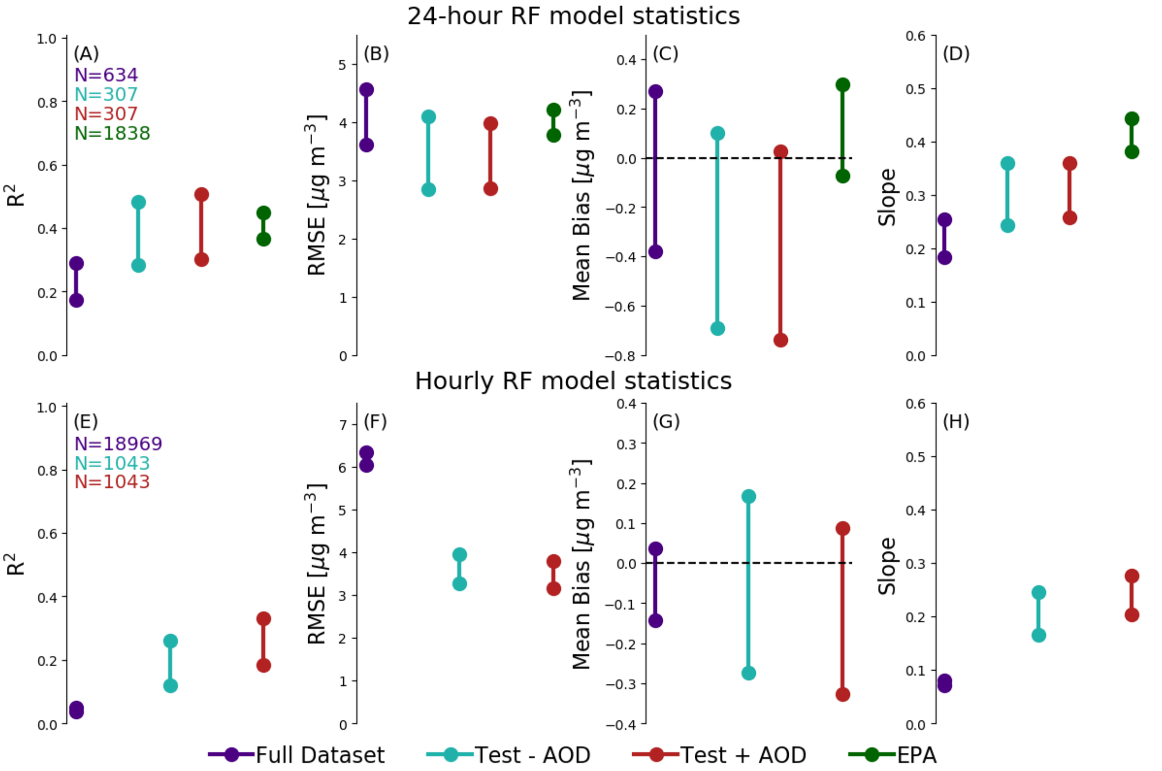


Figure S12. RF model performance metrics for $PM_{2.5}$ measurements using unshuffled or "consecutive" *k*-folds (24-hour in the top row and hourly in the bottom row). The 95% confidence interval of the error metrics for all of the CEAMS RF models (Full Dataset, Test - AOD, and Test + AOD) in predicting both 24-hour and hourly $PM_{2.5}$ and the error metrics for the 24-hour EPA model. The 95% confidence intervals show an estimate of the uncertainty range and, thus, if the intervals of two different models overlap, any difference in their error metrics are likely not statistically significant. The error metrics for each 24-hour $PM_{2.5}$ RF model includes (a) the coefficient of determination (R^2) (b) root mean squared error (RMSE), (c) mean bias, (d) and slope of the linear regression. Plots (e), (f), (g), and (h) show analogous results but for the hourly $PM_{2.5}$ predictions, which we did not predict for the EPA dataset. The size of each 24-hour and hourly dataset, before being split into *k*-folds, is shown in the top left corner of plot (a) and (e).

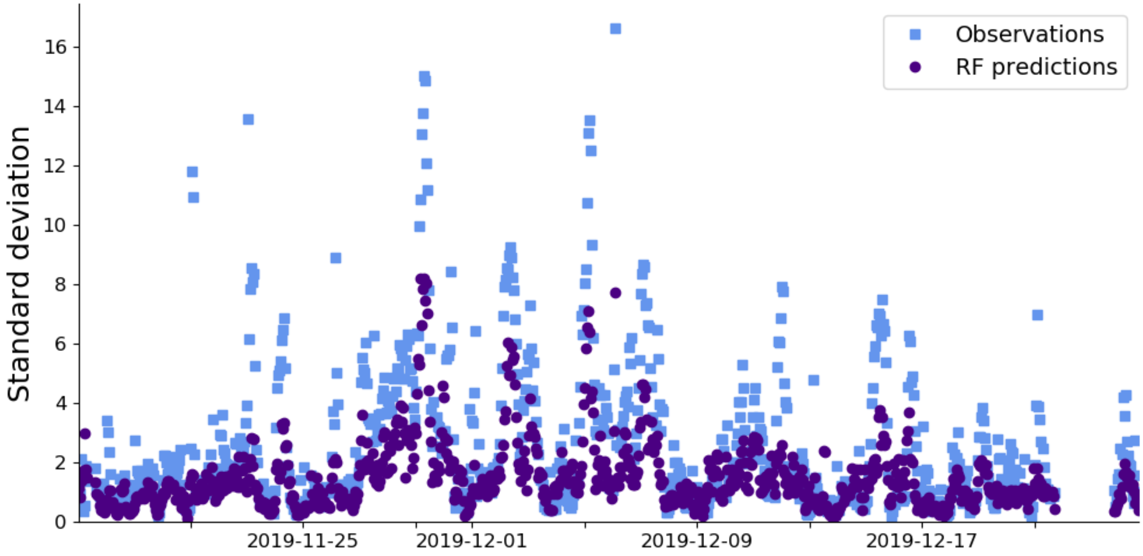


Figure S13. Time series that shows the standard deviation of the hourly $PM_{2.5}$ for each hour of observations (light blue squares) and CEAMS RF predictions (dark blue circles) for hours that had at least 10 monitors operating at the same time. This plot shows results from RF models that used shuffled *k*-folds. 105

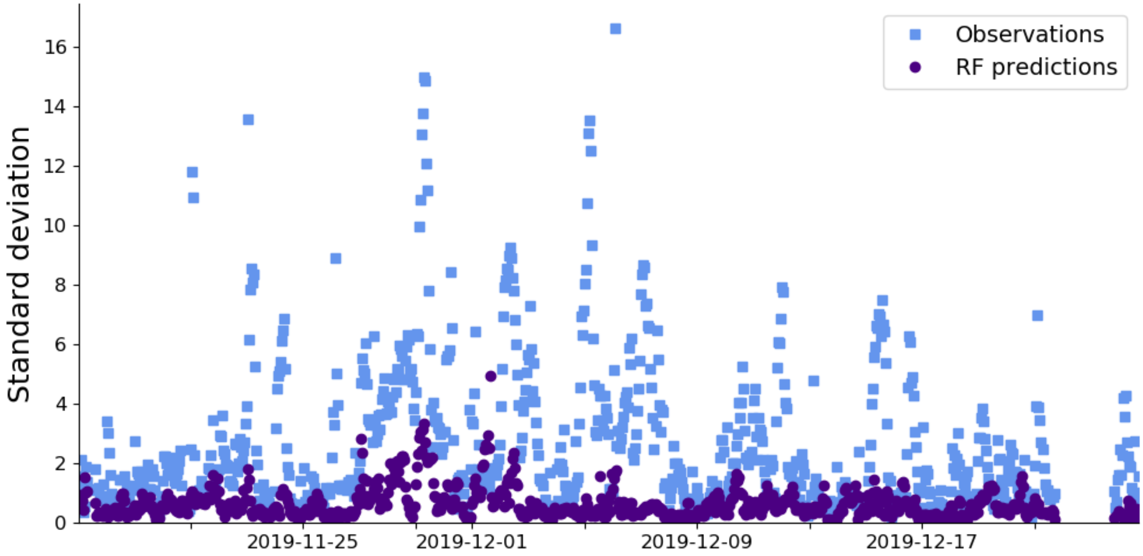


Figure S14. The same as Figure S13 but when unshuffled *k*-folds were used during the training and validation of the RF models predicting CEAMS hourly $PM_{2.5}$.

110

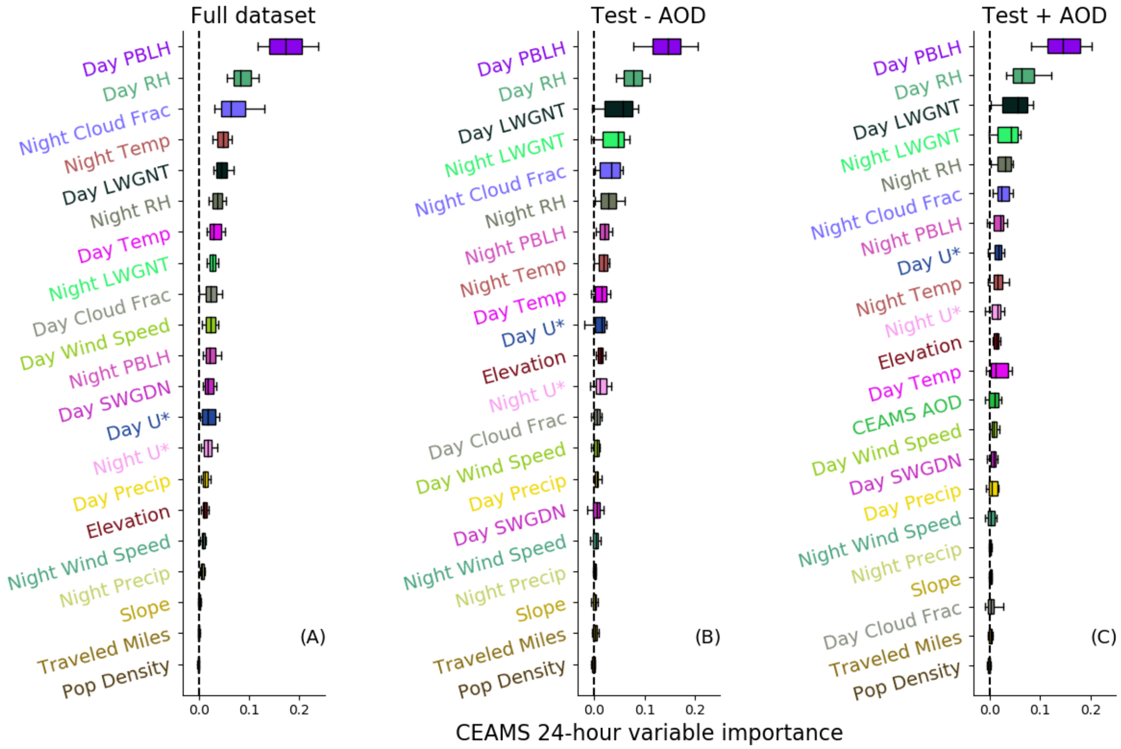


Figure S15. Box-and-whisker plots of 500 permutation importance values for all of the predictors in the RF models that predict CEAMS 24-hour $PM_{2.5}$. The 500 permutation importance values are taken from 100 repeats of permutation importance from each of the 5 testing folds. The whiskers of each box are the 10th and 90th percentile of the permutation importance distribution. The edges of each box represent the 25th and 75th percentile and, finally, the centerline of each box represents the median (i.e., 50th percentile) of the permutation importance distribution. (a) The 24-hour $PM_{2.5}$ predictions of the CEAMS "Full dataset", which contained all of the available 24-hour $PM_{2.5}$ averages regardless of whether daily AOD was available from each location and day. (b) The 24-hour $PM_{2.5}$ predictions of the CEAMS "Test - AOD'' dataset, which only contained 24-hour $PM_{2.5}$ averaged and the associated predictors at locations and days where daily AOD was also available, but we did not use AOD as a predictor for this model. (c) The 24-hour $PM_{2.5}$ predictions of the CEAMS "Test + AOD'' dataset, which only contained $PM_{2.5}$ data where AOD was available and we used AOD as an additional predictor to the meteorological and geographical predictors.

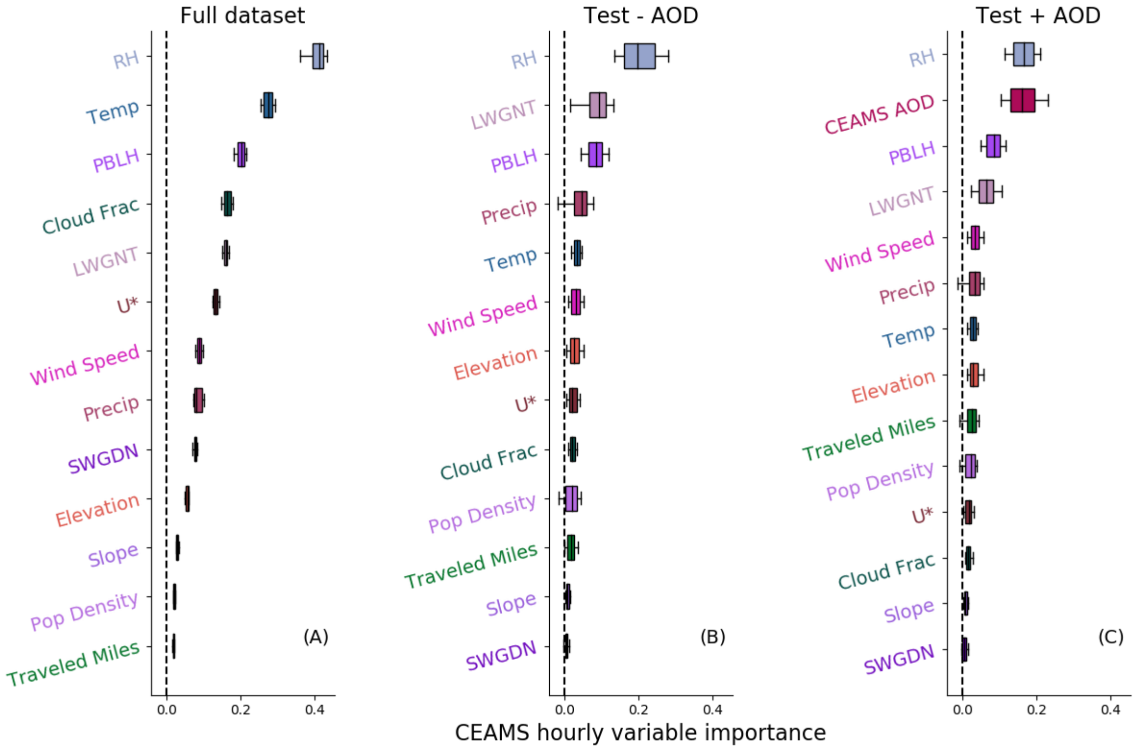


Figure S16. Box-and-whisker plots of 500 permutation importance values for all of the predictors in the RF models that predict CEAMS hourly $PM_{2.5}$. The 500 permutation importance values are taken from 100 repeats of permutation importance from each of the 5 testing folds. The whiskers of each box are the 10th and 90th percentile of the permutation importance distribution. The edges of each box represent the 25th and 75th percentile and, finally, the centerline of each box represents the median (i.e., 50th percentile) of the permutation importance distribution. (a) The hourly $PM_{2.5}$ predictions of the CEAMS "Full dataset", which contained all of the available 24-hour $PM_{2.5}$ averages regardless of whether daily AOD was available from each location and hour. (b) The hourly $PM_{2.5}$ predictions of the CEAMS "Test - AOD" dataset, which only contained 24-hour $PM_{2.5}$ averaged and the associated predictors at locations and days where daily AOD was also available, but we did not use AOD as a predictor for this model. (c) The hourly $PM_{2.5}$ predictions of the CEAMS "Test + AOD" dataset, which only contained hourly $PM_{2.5}$ data where AOD was available and we used AOD as an additional predictor to the meteorological and geographical predictors.

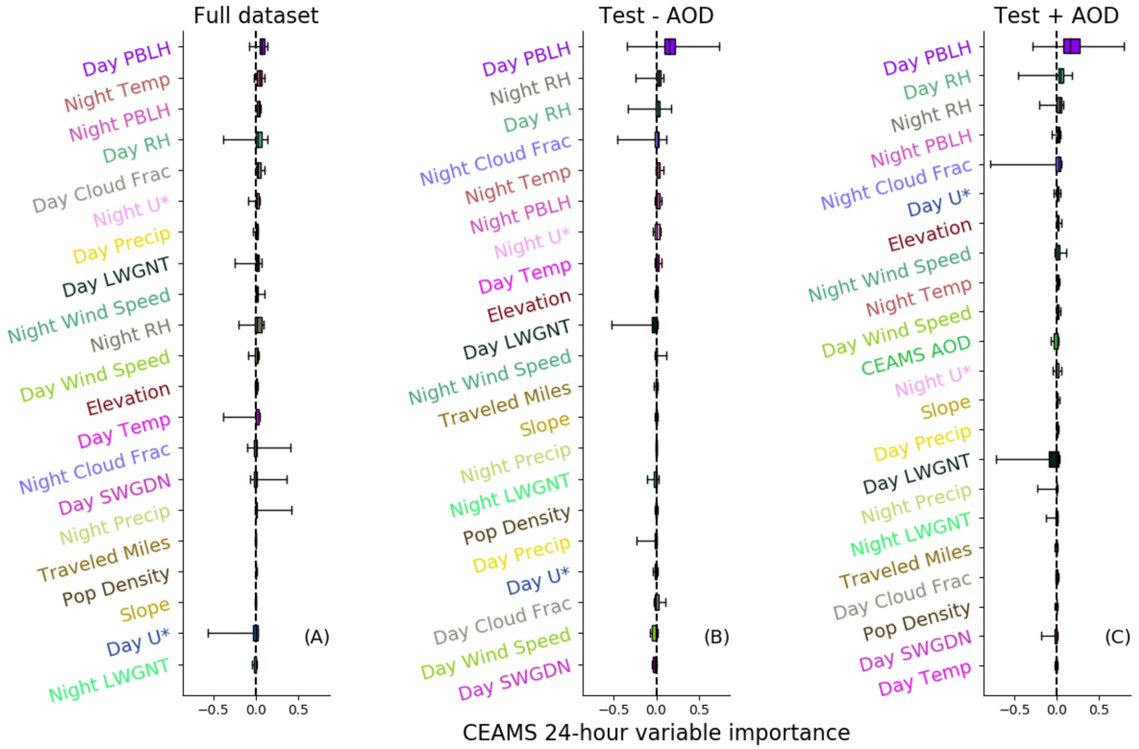


Figure S17. The same as Figure S14 but when unshuffled *k*-folds were used during the training and validation of the RF models predicting CEAMS 24-hour $PM_{2.5}$.

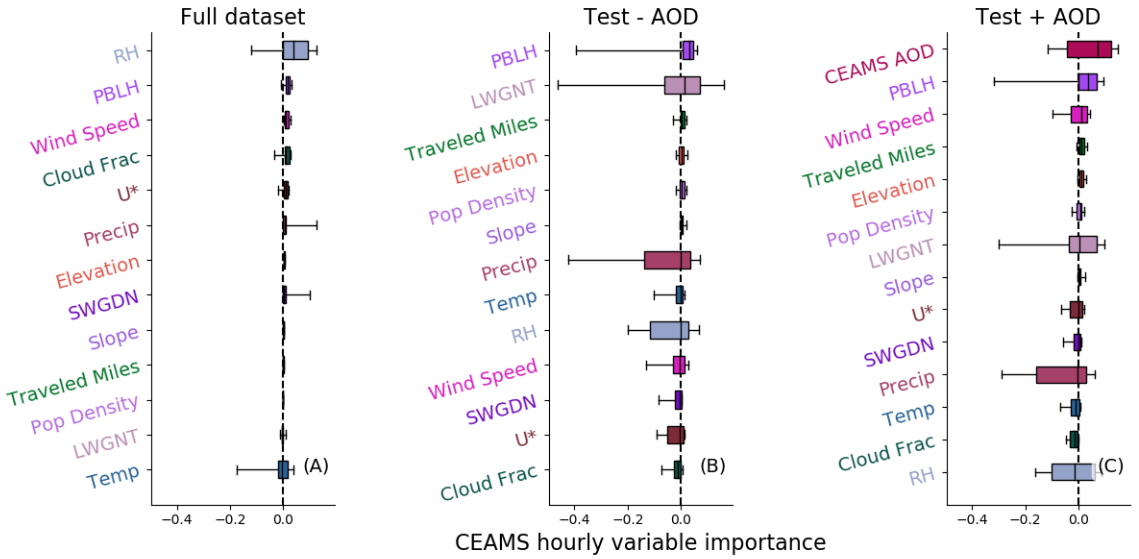


Figure S18. The same as Figure S16 but when unshuffled *k*-folds were used during the training and validation of the RF models predicting CEAMS hourly $PM_{2.5}$.

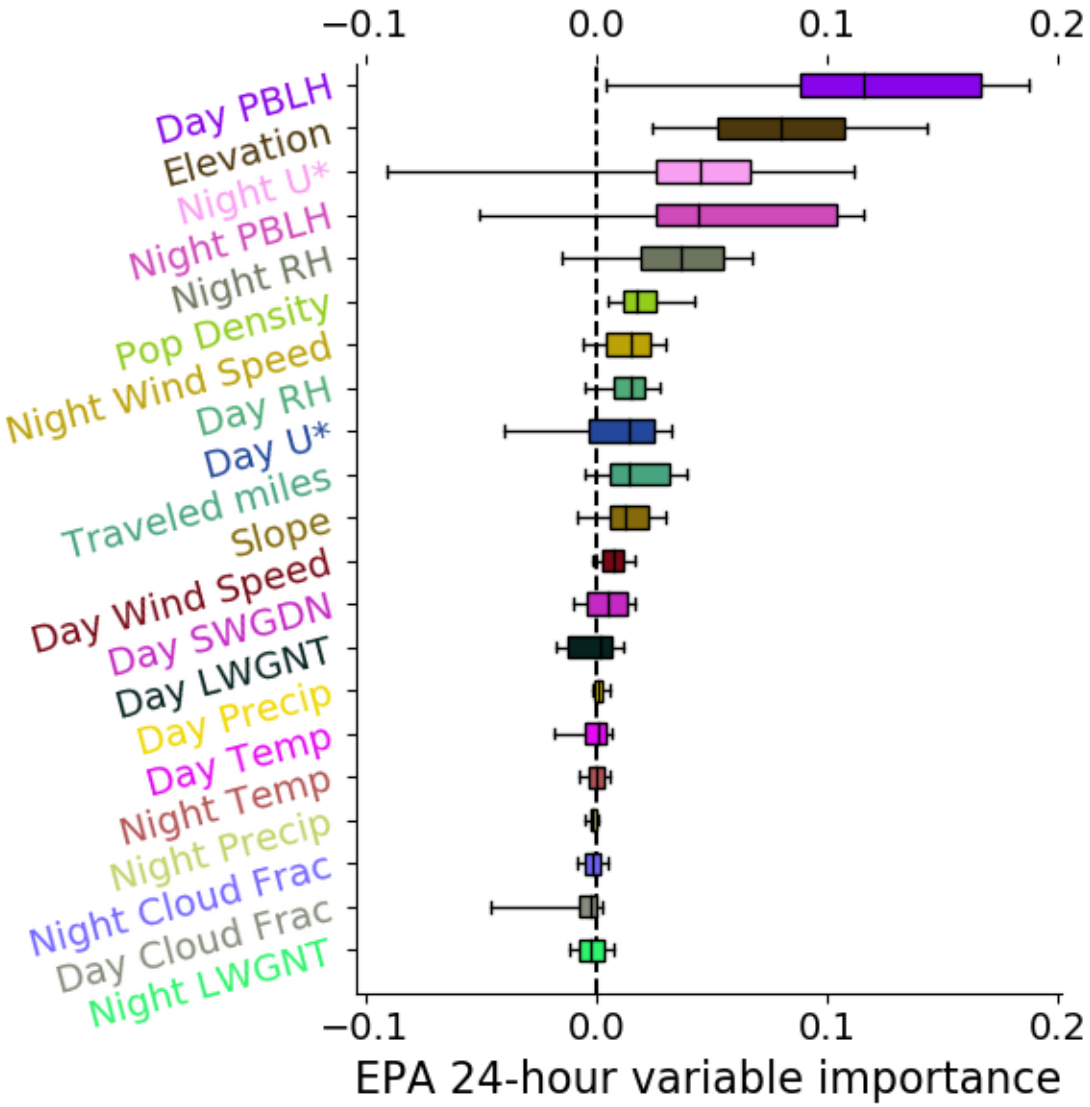


Figure S19. Box-and-whisker plots of 500 permutation importance values for all of the predictors in the RF models that predict EPA 24-hour $PM_{2.5}$ using consecutive *k*-folds. The 500 permutation importance values are taken from 100 repeats of permutation importance from each of the 5 testing folds. The whiskers of each box are the 10th and 90th percentile of the permutation importance distribution. The edges of each box represent the 25th and 75th percentile and, finally, the centerline of each box represents the median (i.e., 50th percentile) of the permutation importance distribution.

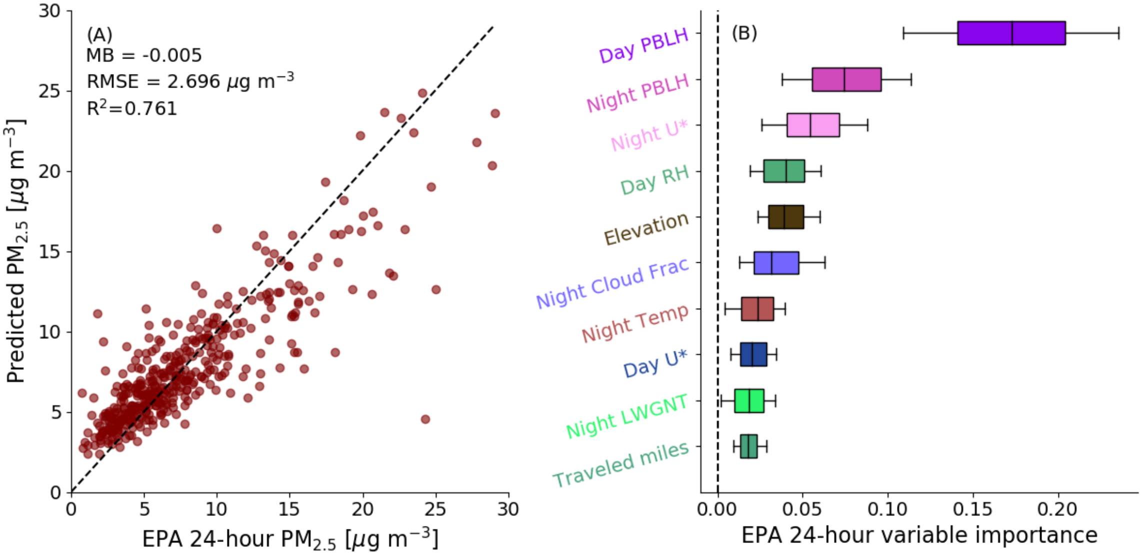


Figure S20. (a) All points from the testing folds of the 5-fold CV for the EPA 24-hour RF model for only one winter (Dec. 15 - Jan. 15, 2019) and shuffled *k*-folds. (b) Box-and-whisker plots of the distribution of 100 permutation importance metrics for the top 10 ranked predictors of the 24-hour EPA $PM_{2.5}$ for one winter (Dec. 15 - Jan. 15, 2019) and shuffled *k*-folds.

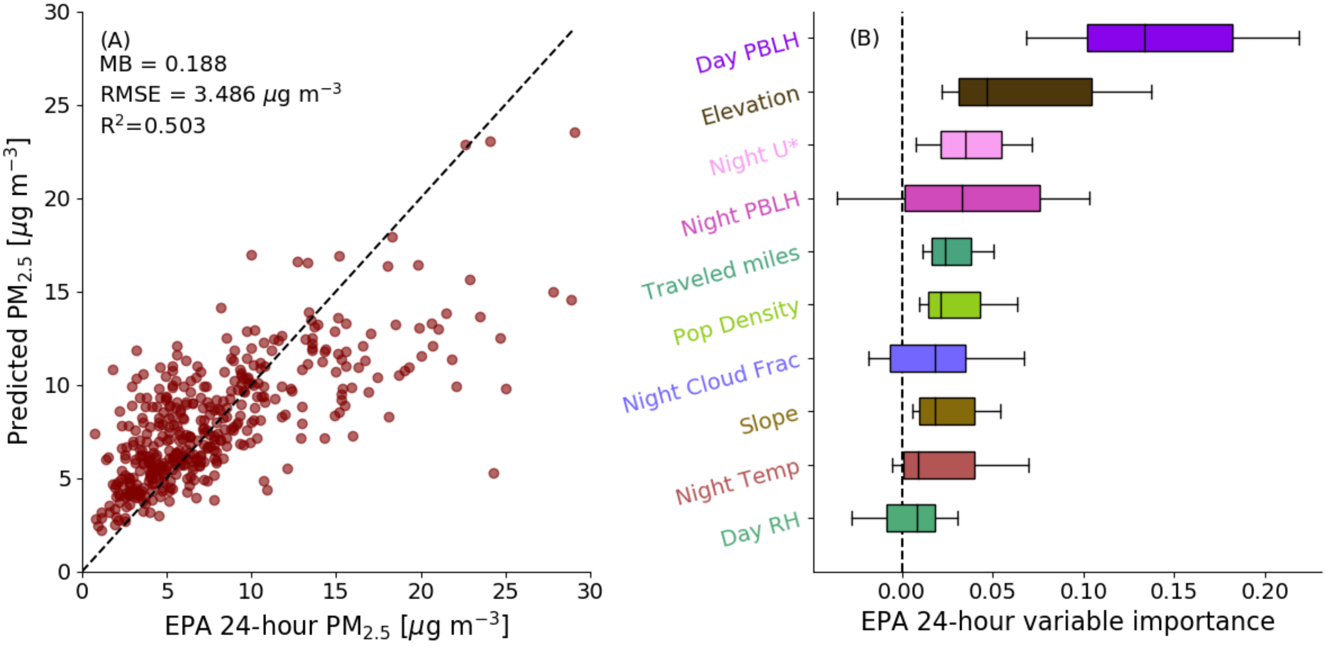


Figure S21. (a) All points from the testing folds of the 5-fold CV for the EPA 24-hour RF model for one winter (Dec. 15 - Jan. 15, 2019) and consecutive *k*-folds. (b) Box-and-whisker plots of the distribution of 100 permutation importance metrics for the top 10 ranked predictors of the 24-hour EPA $PM_{2.5}$ for one winter (Dec. 15 - Jan. 15, 2019) and consecutive *k*-folds.