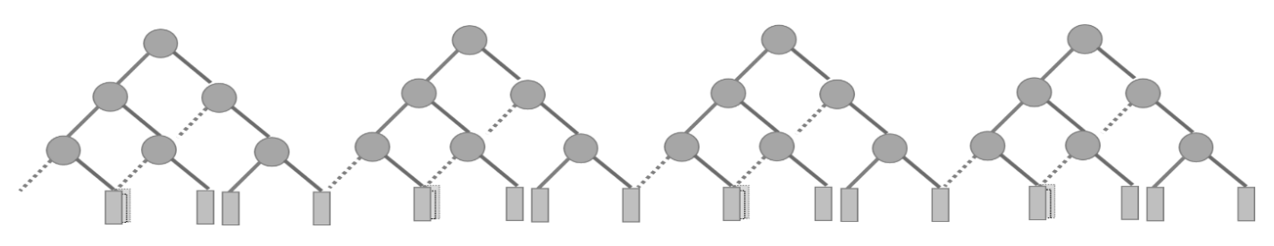
What Can We Learn From Machine Learning Models Developed for Short-term Forecasting of $PM_{2.5}$?

Meenakshi Rao[1,2,*], Aaron Fellows[2], Morgan Schafer[2], Phillip Orlando[1], and Linda Acha George[1]

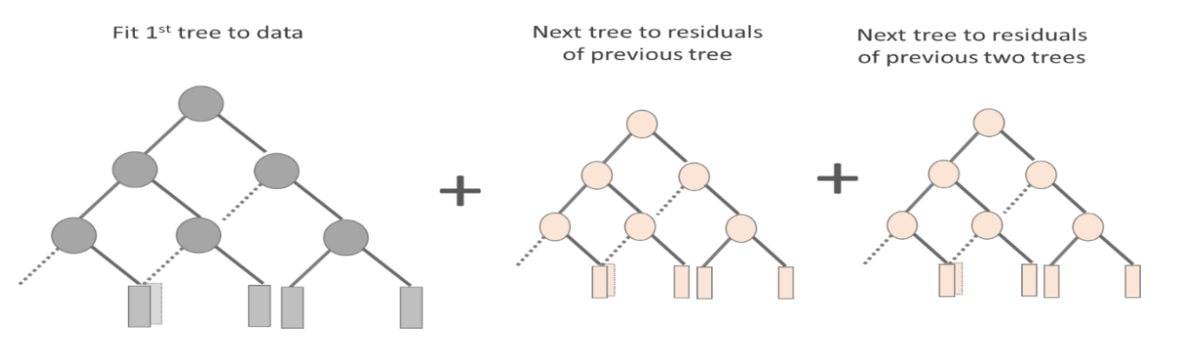
[1]Portland State University, Portland OR, United States, [2]Oregon Department of Environmental Quality, Portland, OR, United States
* Corresponding author mrao@pdx.edu

Background

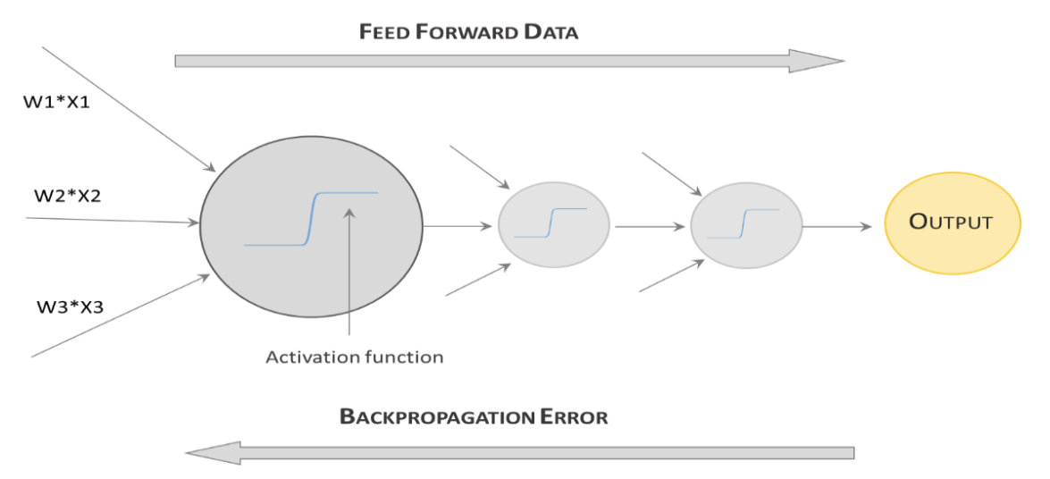
Accurate short-term forecasting of PM2.5 concentrations is important for reducing exposure to unhealthy levels of air pollution. The EPA uses a 12-hour weighted mean for its NowCast prediction of PM2.5 concentrations. We compared the performance of three Machine Learning (ML) algorithms (random forest, generalized boosted trees and multi-layer perceptron) against the EPA's NowCast.



Random forest : Learning from experience
Random forest is an ensemble method that combines predictions from many "weak" learners (the trees) into a prediction from a "strong" learner (the forest). Each tree is develop based on a random subset of predictors. Each subsample is created using bootstrap aggregation or bagging.


Fit 1st tree to data
+
Next tree to residuals of previous tree
+
Next tree to residuals of previous two trees

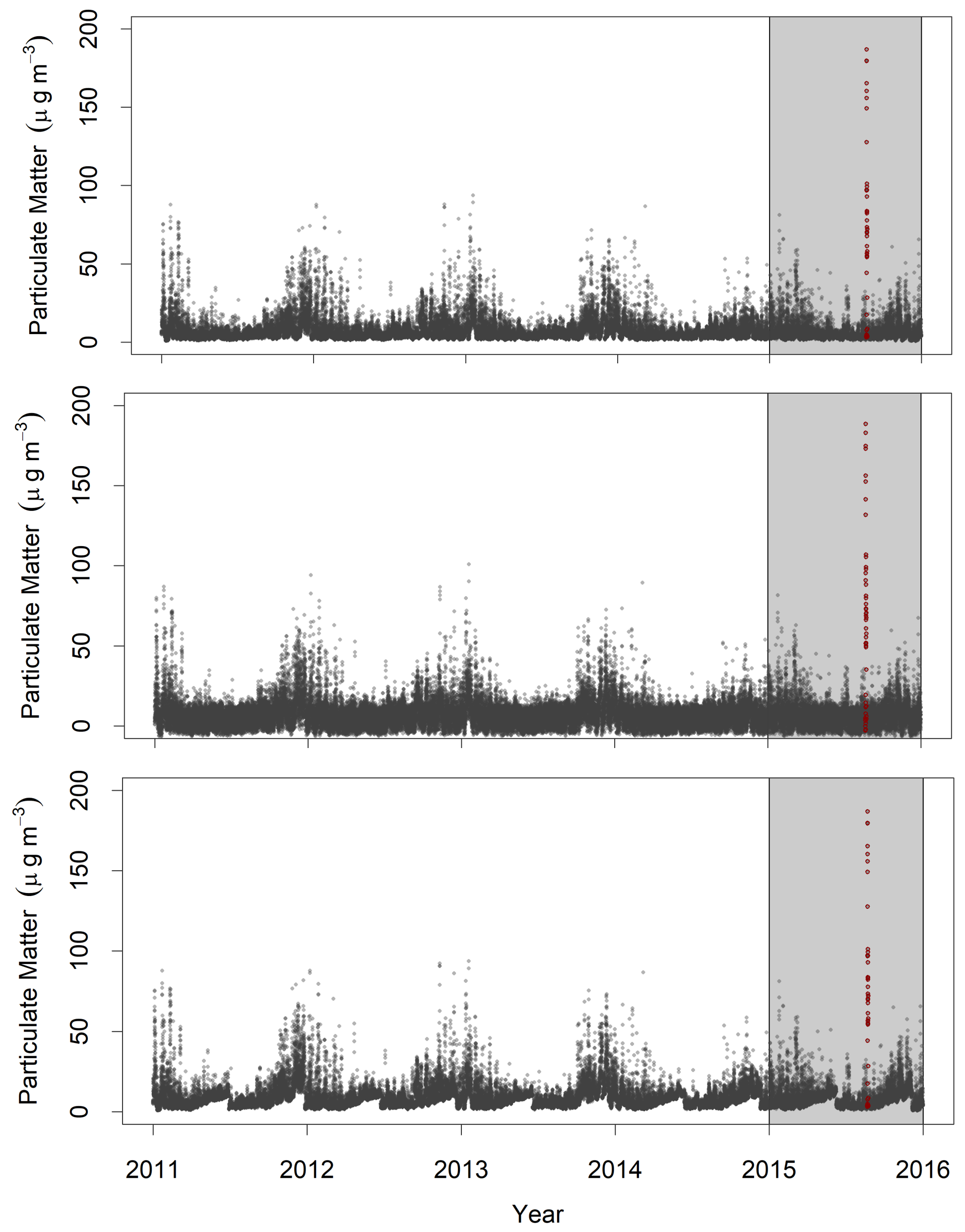
Generalized boosted models: Learning from mistakes
Generalized boosted models (gbm) is also an ensemble tree method that combines predictions from many trees. However, each tree is optimized to predict the residuals from previous trees.



Multi-layer perceptron: A feedforward artificial neural network approach. Weights for connections between neurons are iteratively adjusted until error is minimized.

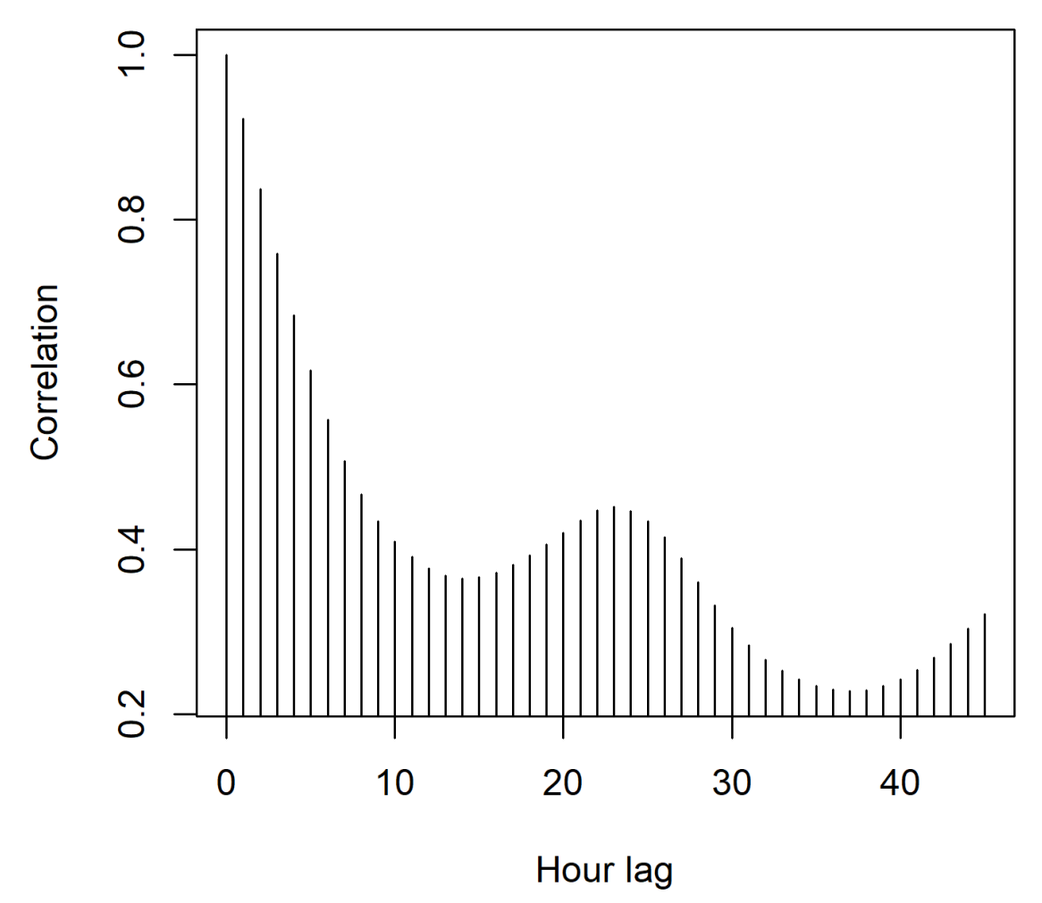
Low-cost air pollution sensors are becoming more prevalent, and provide additional opportunities for PM2.5 forecasting. However data from these sensors may have more noise and drift compared to traditional higher-end instruments. Here we examine how the EPA's NowCast and ML algorithms perform on data that have been perturbed to include noise and drift. The NowCast method was especially developed to capture rapid changes in PM2.5 associated with a wildfire or smoke event. Therefore we also examine how these algorithms perform in predicting PM2.5 during a wildfire event.

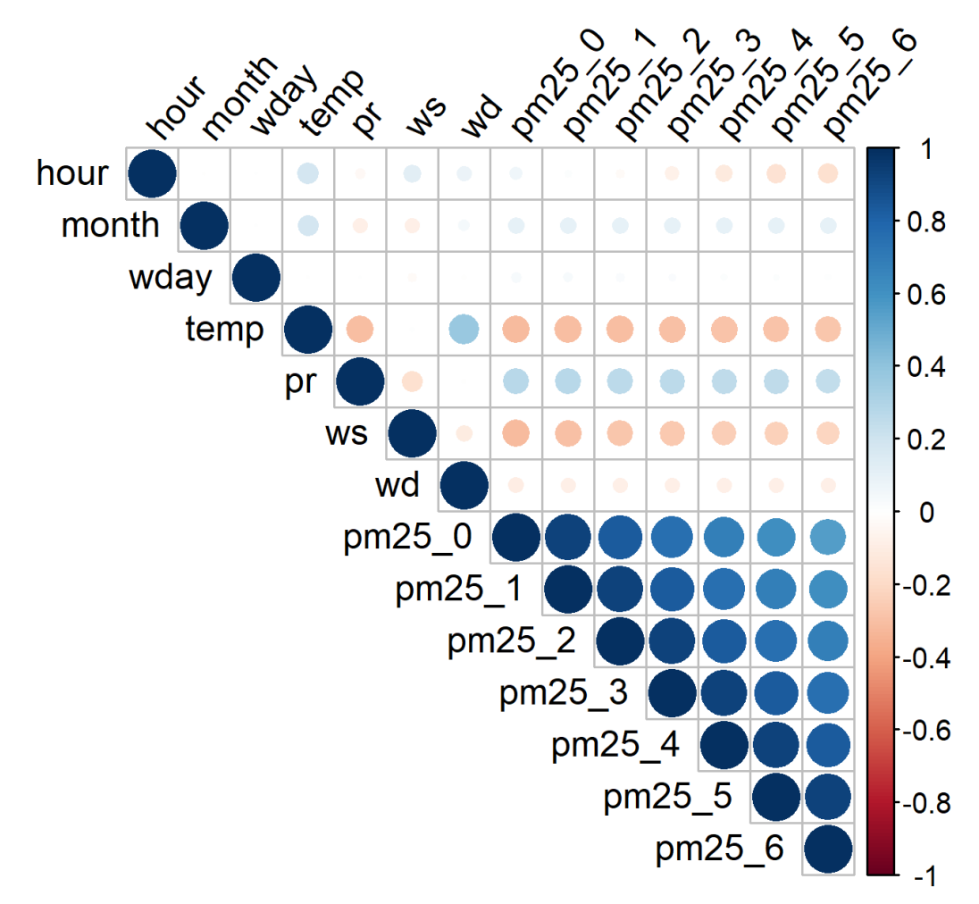
Data Input for Machine Learning (ML) Algorithms

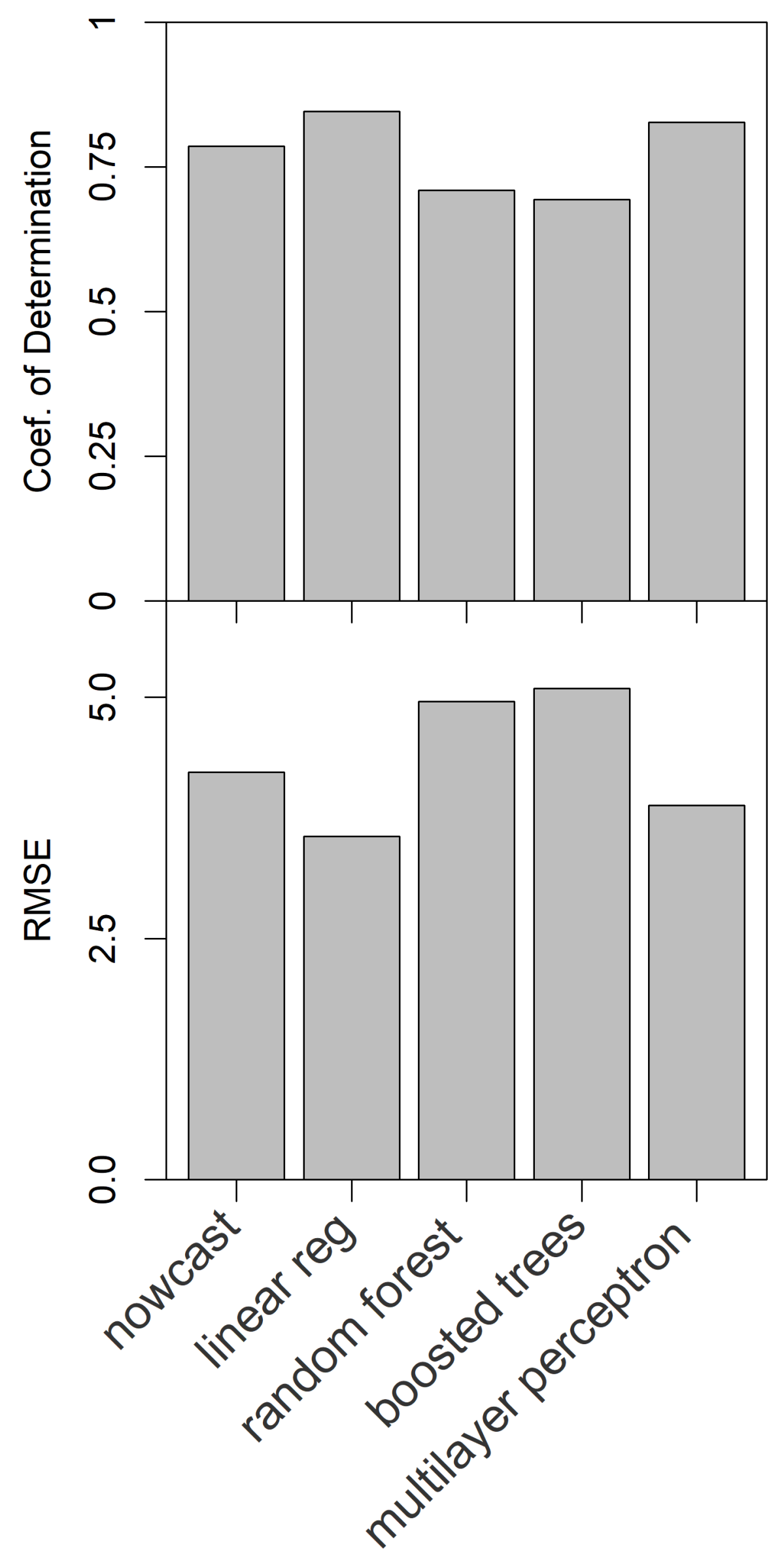


PM2.5 from Oregon DEQ's Portland air monitoring station was used for model inputs. PM2.5 from 2011-2014 was used to train the algorithms. PM2.5 from 2015 (grey background) was used for validation. Wildfire smoke created a spike in PM2.5 during 2015, and is shown in red. The top panel shows unperturbed data. The middle panel is unperturbed data with ±5 standard deviation of added noise. The bottom panel shows unperturbed data with added drift.

How do ML Algorithms Perform for Short Term Air Forecasting?



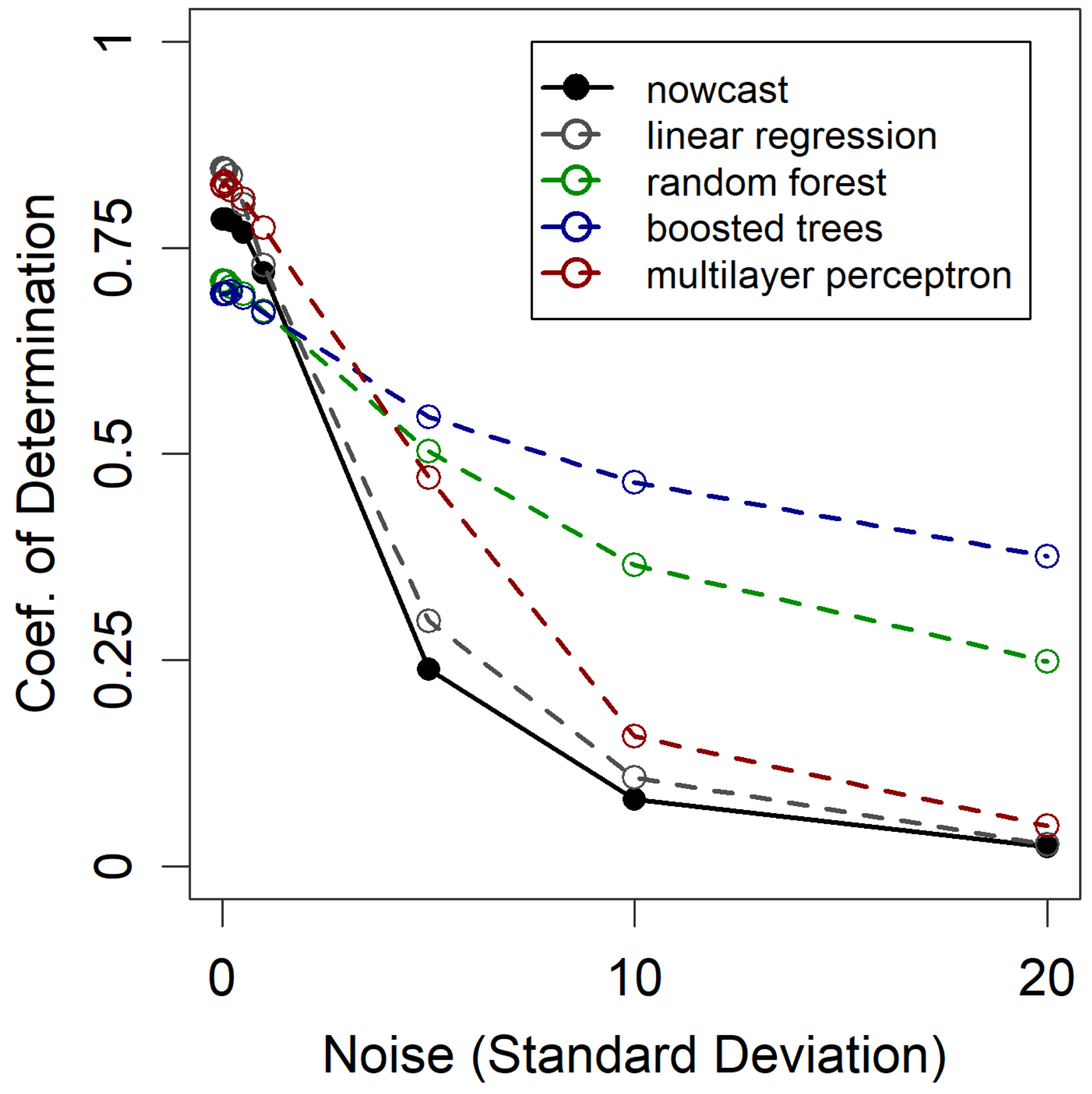




PM2.5 was first examined without noise or drift (above). The upper left panel shows autocorrelation in unperturbed PM2.5, indicating hourly and diurnal patterns. The lower left panel shows correlation for all predictors used in the machine learning models. The right panel shows the performance of the EPA's NowCast, linear regression, random forest, boosted trees, and multilayer perceptron approaches. Coefficient of determination (R^2) and Root Mean Square Error (RMSE) were used to assess performance, and were determined by comparing predicted and observed PM2.5 during the 2015 validation period.

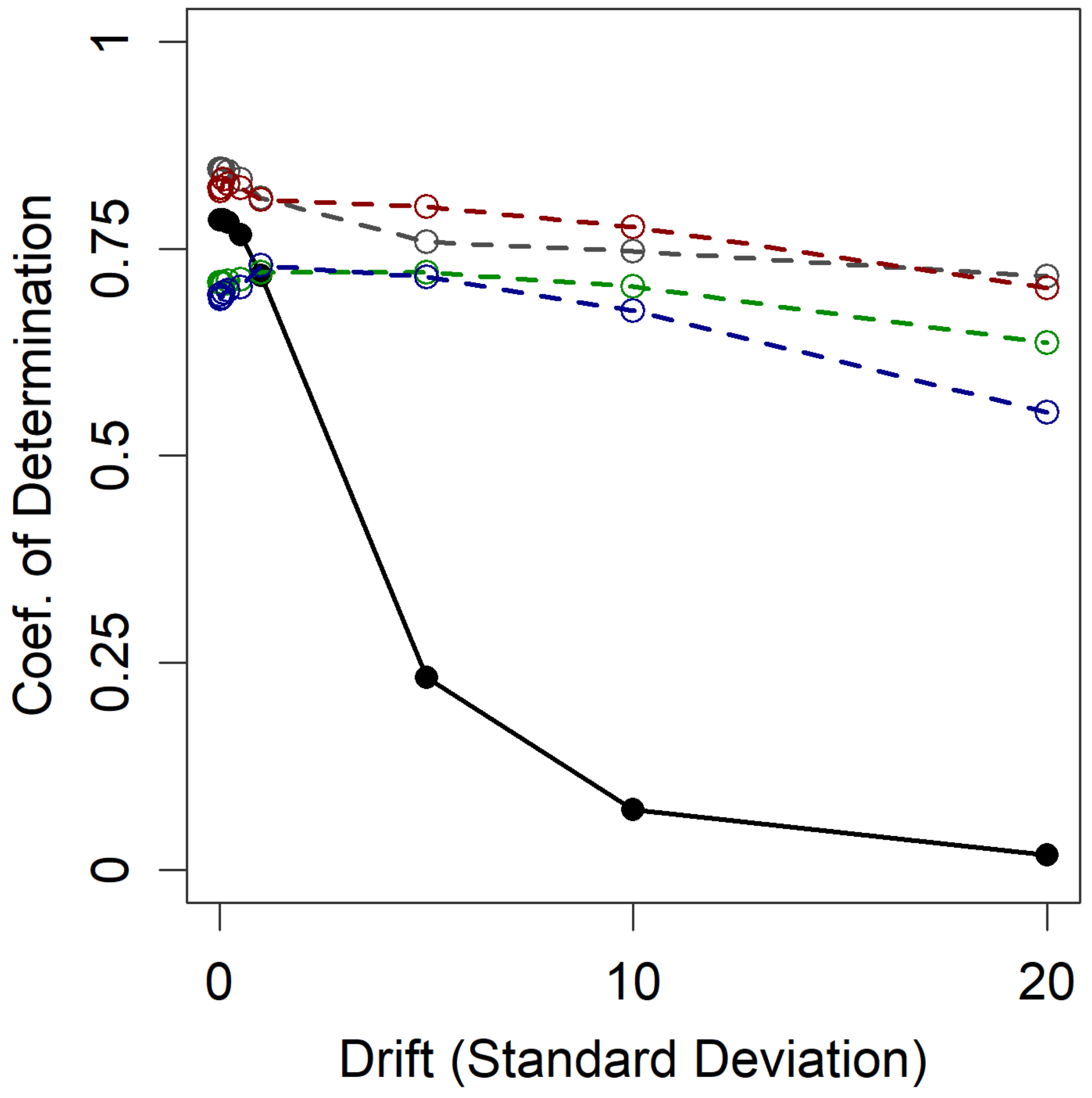
How do ML Algorithms Perform with Noise, Drift, and Outliers?

Noise



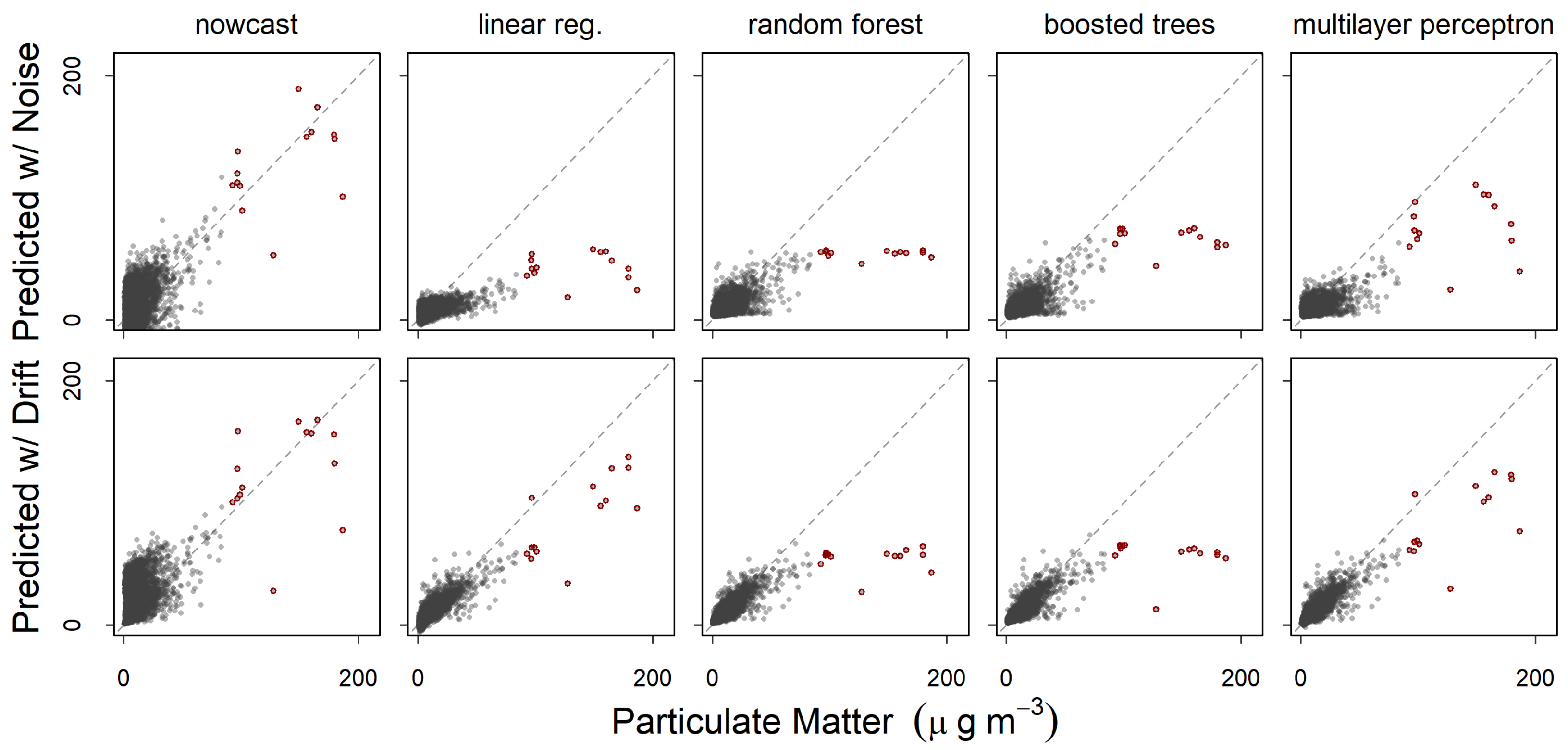
The coefficient of determination (R^2) between predicted and observed 2015 validation data for model runs with added noise. Model performance remained high at small-to-moderate noise. R^2 dropped with large increases in noise for all models. Random forest and boosted trees performed best at high noise.

Drift



The coefficient of determination (R^2) between predicted and observed 2015 validation data for model runs with added drift. Performance of all algorithms (with the exception of the NowCast) remained relatively stable with increased drift.

Outliers



Predicted vs. Observed PM2.5 (μg m^{-3}) with ±5 S.D. of added noise and drift for all model runs. Smoke from a fire created a spike in PM2.5 during Aug. 2015. Capturing these PM2.5 spikes is particularly important for short-term air quality forecasting. The EPA NowCast and multilayer perceptron models performed well, whereas other models under predicted PM2.5. These figures also show several models often under-predicted PM2.5 despite high R^2.

Conclusions

ML algorithms performed well in the presence of noise and drift, although they did not perform as well on predicting values outside the range of the training datasets (outliers). Overall, ML algorithms show promise in generating good predictions based on relatively noisy datasets characteristic of low-cost sensor networks.

R code, input data, and this poster have been uploaded to: https://github.com/MRaoODEQ/ml_nowcast