Supplemental Material

Ensemble averaging based assessment of spatiotemporal variations in ambient $PM_{2.5}$ concentrations over Delhi, India, during 2010-2016

Siddhartha Mandal, Kishore K. Madhipatla, Sarath Guttikunda, Itai Kloog, Dorairaj Prabhakaran, Joel D. Schwartz

December 17, 2019

Table S1: Description of the stages of the prediction model, including outcomes, predictors (both spatial and spatiotemporal) and modeling technique used in each stage. In the Table, r denotes the ratio of concentrations of $PM_{2.5}$ and PM_{10}.

	MODEL STAGES			
	Calibration regression	PM-AOD relation	Ensemble averaging	Spatial smoothing
Outcome	$log(r/1-r)$	$log(PM_{2.5})$	$PM_{2.5}$	Ensemble averaged prediction of $PM_{2.5}$
Method	Support Vector Regression	Generalized Additive Models, Elastic Net, Support Vector Regression, Random Forests, Neural Networks, Extreme Gradient Boosting	Generalized additive model	Generalized additive model
Spatial predictors	Airport, Bus Stops and Railway stations within 10km	Minimum distance to solid dumps and power plants, Area of builtup, openspace and vegetation, Length of roads and runways, Number of intersections, markets, malls and transport hubs, Emissions in tons per year.		
Spatiotemporal predictors	Meteorological variables and 1 day lags, Boundary layer height, PM10 concentration, Month, Days of construction at airport	Meteorological variables and 1 day lags, Boundary layer height, Ultraviolet absorption index, Population density, Carbon emissions from agricultural fires and their 5 day lagged average	Penalized splines of predictors from each learner	Tensor products of (Latitude, Longitude), Average predicted PM at nearby stations

Table S2: Number of observations and cross-validated prediction R^2, root mean squared error (RMSE) and coefficient of variation of RMSE (CV(RMSE)) for calibration regression of $PM_{2.5}$ on PM_{10} across 2010-2016. The cross-validated R^2 is based on robust linear regression across years.

YEAR	Number of observations	R^2	RMSE	CV(RMSE)
2010	24	0.547	42.234	0.383
2011	58	0.757	49.705	0.649
2012	175	0.914	34.841	0.218
2013	331	0.946	35.146	0.209
2014	242	0.846	30.933	0.286
2015	1161	0.886	30.950	0.278
2016	855	0.946	37.009	0.250

Table S3: Number of observations of $PM_{2.5}$ at each monitoring station in the National Capital Region across years, after applying data cleaning filters and implementing the calibration regression.

STATION	2010	2011	2012	2013	2014	2015	2016
Anand.Vihar	0	0	0	0	0	254	326
Civil.Lines	38	0	213	190	0	85	0
DCE	130	77	0	0	0	54	0
Dwarka	354	346	310	352	337	325	313
Faridabad	0	0	0	0	0	243	342
Gurgaon	0	0	0	0	0	29	259
IGI	90	315	358	235	102	137	0
IHBAS	262	344	305	347	351	199	253
ITO	269	349	205	0	142	3	57
Janakpuri	0	72	0	0	73	76	0
Mandir.Marg	0	0	0	0	0	254	280
Mayapuri	0	96	96	80	96	95	0
Nizamuddin	0	70	0	0	77	74	0
NY.School	0	96	96	95	90	92	0
Pitampura	0	75	0	0	75	78	0
Punjabi.Bagh	23	2	0	0	0	248	342
RK.Puram	0	0	0	2	235	270	352
Shadipur	212	306	319	363	354	351	328
Shahdara	0	75	0	0	73	74	0
Shahzada.Bagh	0	71	0	0	72	77	0
Sirifort	0	71	0	0	68	74	0
Town.Hall	0	94	96	95	90	91	0
US.Consulate	0	0	0	344	350	318	329
Total	1378	2459	1998	2103	2585	3501	3181

Table S4: Cross-validated overall, spatial and temporal R^2 for individual learners and ensemble averaged prediction across years for observations lower than 100 $\mu g/m^3$.

YEAR	BAM	GLMNET	SVM	RF	NN	XGBOOST	$EAVG$	$EAVG_{Sp}$	$EAVG_T$
2010	0.246	0.270	0.485	0.549	0.440	0.657	0.706	0.966	0.668
2011	0.363	0.305	0.488	0.545	0.421	0.631	0.636	0.749	0.632
2012	0.152	0.215	0.418	0.468	0.268	0.494	0.566	0.891	0.476
2013	0.100	0.076	0.256	0.384	0.242	0.368	0.439	0.802	0.391
2014	0.084	0.063	0.280	0.323	0.187	0.319	0.425	0.747	0.394
2015	0.276	0.247	0.498	0.565	0.390	0.589	0.639	0.891	0.576
2016	0.357	0.408	0.610	0.758	0.572	0.780	0.849	0.882	0.841

Table S5: Cross-validated overall, spatial and temporal R^2 for individual learners and ensemble averaged prediction across years for observations higher than or equal to 100 $\mu g/m^3$.

YEAR	BAM	GLMNET	SVM	RF	NN	XGBOOST	$EAVG$	$EAVG_{Sp}$	$EAVG_T$
2010	0.308	0.254	0.505	0.552	0.362	0.592	0.618	0.945	0.558
2011	0.131	0.129	0.247	0.349	0.186	0.377	0.451	0.834	0.419
2012	0.175	0.176	0.361	0.456	0.294	0.411	0.493	0.929	0.455
2013	0.092	0.067	0.265	0.316	0.222	0.273	0.378	0.566	0.296
2014	0.078	0.062	0.173	0.202	0.170	0.205	0.262	0.443	0.228
2015	0.160	0.139	0.336	0.396	0.272	0.444	0.497	0.874	0.454
2016	0.264	0.249	0.468	0.559	0.328	0.613	0.650	0.871	0.617

Table S6: Cross-validated R^2 for ensemble averaged prediction across years and seasons comparing a full ensemble based on all six algorithms and a tree based ensemble using only random forests and extreme gradient boosting predictions.

Ensemble	Season	2010	2011	2012	2013	2014	2015	2016
Full ensemble	Overall	0.860	0.765	0.766	0.745	0.637	0.818	0.895
Tree based ensemble		0.856	0.765	0.759	0.736	0.613	0.812	0.893
Full ensemble	Winter	0.827	0.673	0.689	0.728	0.510	0.636	0.768
Tree based ensemble		0.827	0.662	0.671	0.719	0.475	0.631	0.776
Full ensemble	Fall	0.713	0.751	0.603	0.785	0.597	0.801	0.908
Tree based ensemble		0.701	0.758	0.573	0.756	0.527	0.795	0.901
Full ensemble	Summer	0.733	0.504	0.656	0.636	0.378	0.586	0.703
Tree based ensemble		0.713	0.495	0.647	0.631	0.372	0.574	0.693
Full ensemble	Monsoon	0.648	0.514	0.698	0.731	0.590	0.751	0.748
Tree based ensemble		0.651	0.517	0.698	0.726	0.578	0.745	0.749

Figure S1: Comparison of observed and predicted $PM_{2.5}$ concentrations at the ground monitoring stations along with measures of prediction accuracy according to year and season.

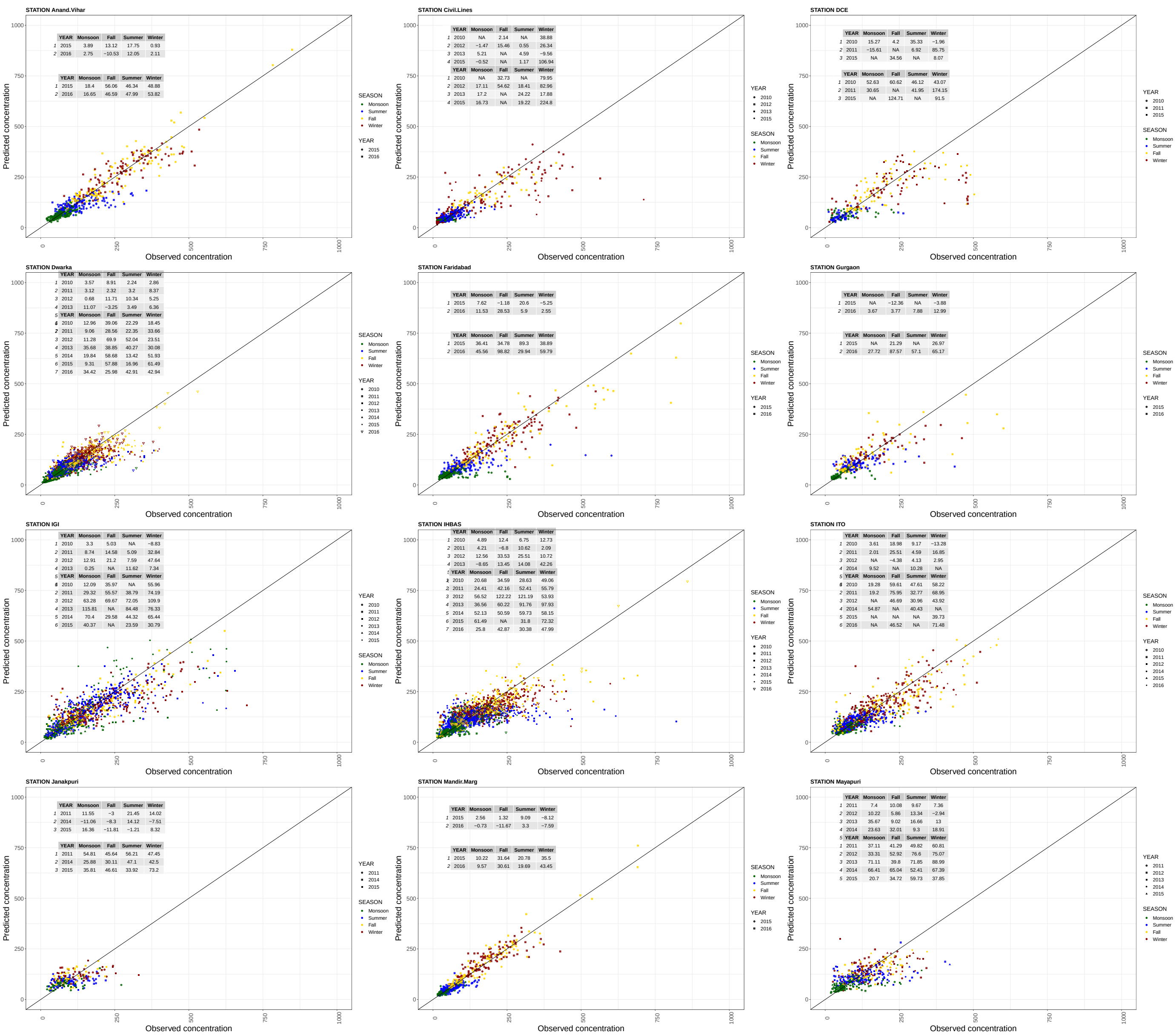


Figure S2: Comparison of observed and predicted $PM_{2.5}$ concentrations at the ground monitoring stations along with measures of prediction accuracy according to year and season.

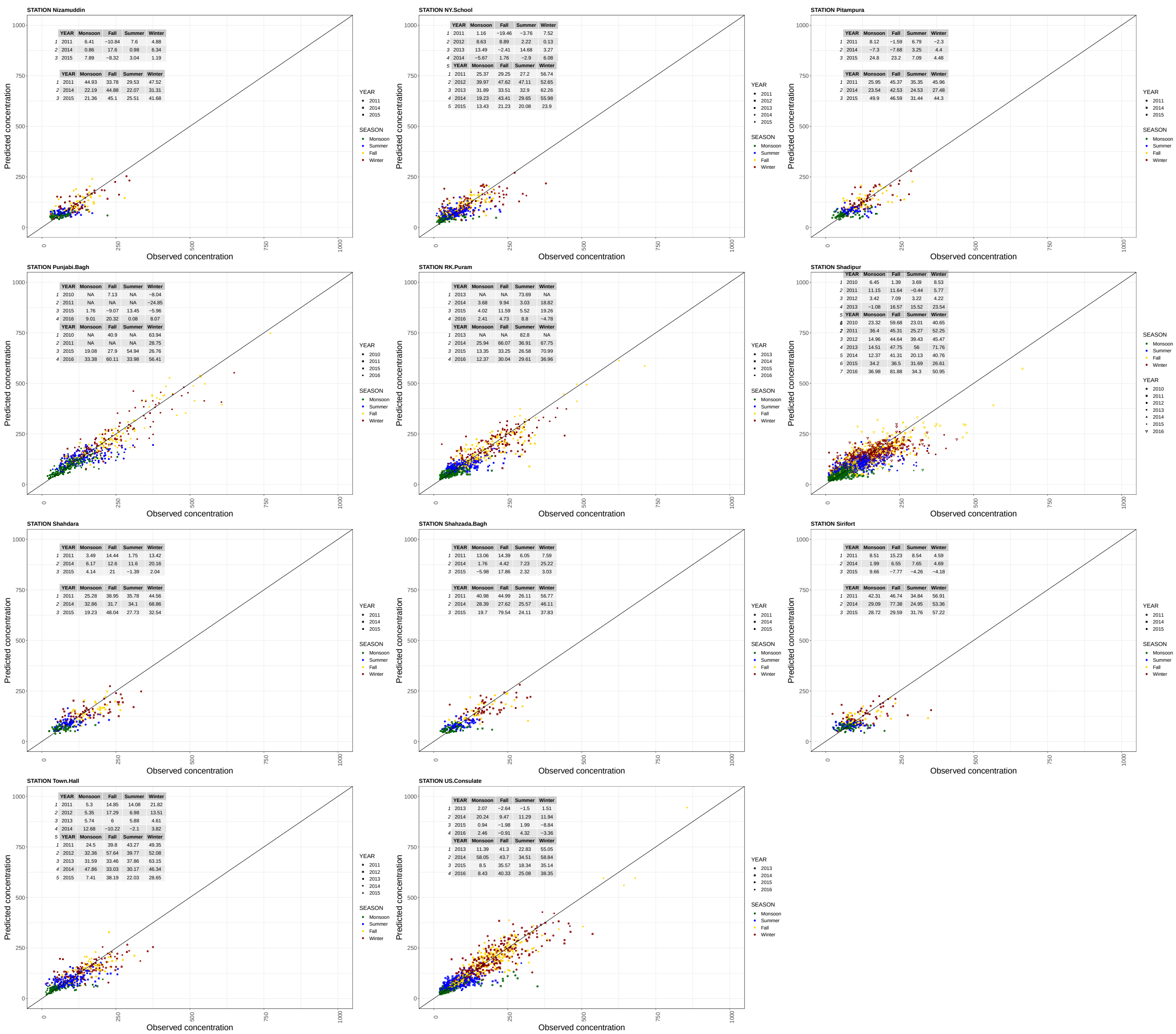
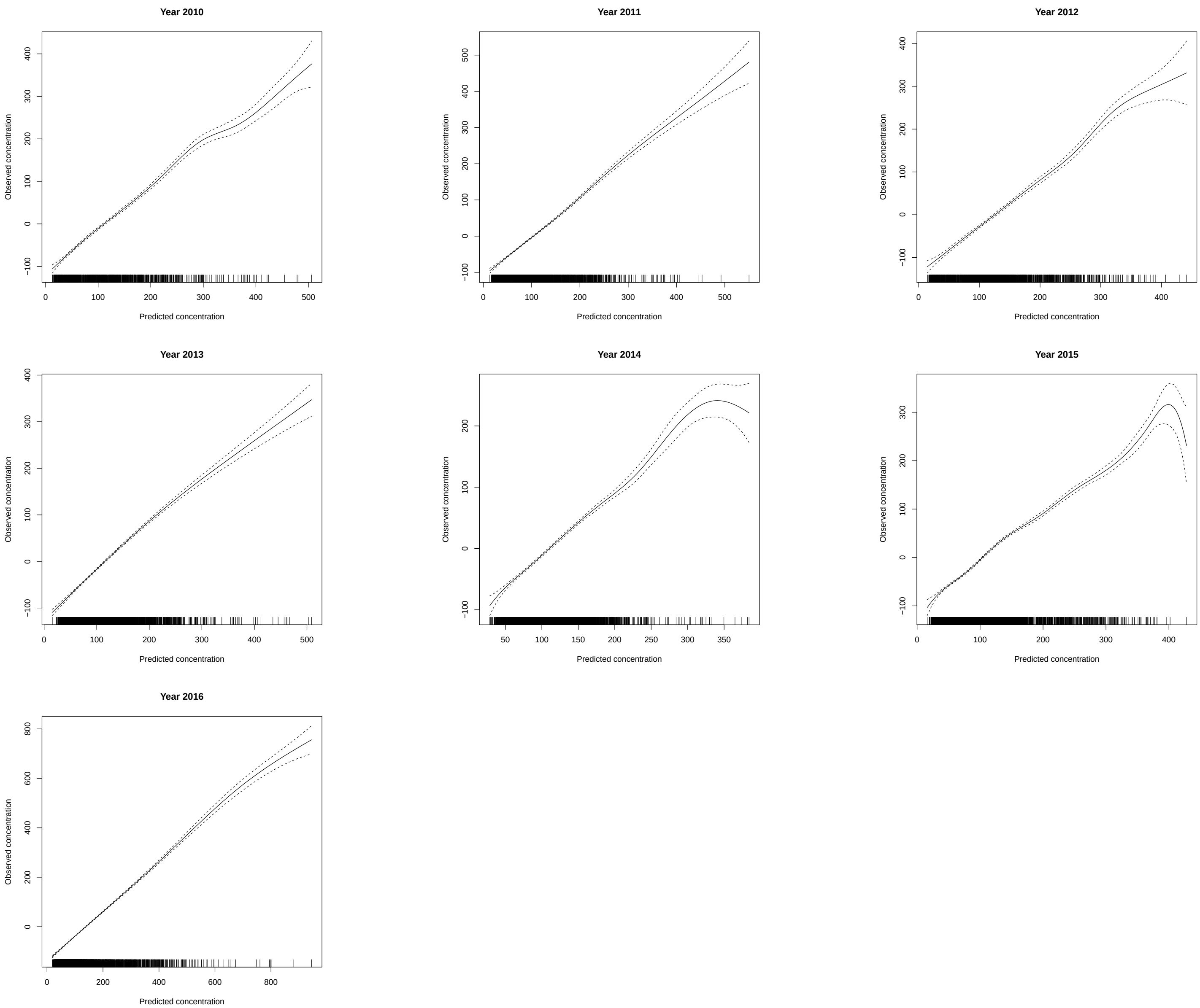


Figure S3: Plots showing observed against cross-validated predictions for $PM_{2.5}$ according to generalized additive models using penalized splines without specified degrees of freedom, across different years.



Tuning parameters for machine learning algorithms used in the paper

1. ***Calibration regression***: A Support Vector Machines with Radial Basis Function Kernel was employed to calibrate the ratio of $PM_{2.5}$ and PM_{10}. The *svmradial* package was used through the *caret* package in R. The two tuning parameters in the algorithm were σ (with values 0.05, 0.075, 0.10, 0,125, 0.15, 0.175, 0.20, 0.225 and 0.25)and C (with values 1,2,5,7,10).

2. ***Calibration of MAIAC AOD with CAMS AOD***: A random forest model was implemented in the R caret package through the *ranger* package. The number of predictors ($mtry$) was the only tuning parameter with values 10, 15, 20, 25 and 30, along with minimum node size as 5.

3. ***Machine learning algorithms for ensemble averaging***:

   (a) Elastic net: Implemented through *enet* and *caret* packages in R. The tuning parameters were the weights ($\alpha = 0.1$, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8) and penalty ($\lambda = 0.001$, 0.01, 0.02, 0.05, 0.1) with the grid search being over combinations of α and λ.

   (b) Support vector regression: Same as (1) earlier.

   (c) Neural networks: Implemented through *nnet* and *caret* packages in R. Tuning parameters and associated values were decay (0.01,0.05, 0.1) and size (4,5,6,7), with the grid search being over combinations of all tuning parameters.

   (d) Random Forests: Same as (2) earlier.

   (e) Extreme gradient boosting: Implemented through *xgbTree* and *caret* packages in R. Tuning parameters and associated values were nrounds (100,500,1000), max depth (5,10,15), eta (0.1,0.2,0.5), gamma = 0, colsample bytree (2/3,1/2,1/3), min child weight= 1 and subsample (.5,.75), with the grid search being over combinations of all tuning parameters.