Supplementary Information

Predicting peak daily maximum 8-hour ozone, and linkages to emissions and meteorology, in Southern California using machine learning methods

Ziqi Gao[1], Yifeng Wang[1], Petros Vasilakos[1], Cesunica E. Ivey[2,4], Khanh Do[2,3] Armistead G. Russell[1]

[1] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA

[2] Department of Chemical and Environmental Engineering, University of California, Riverside, Riverside, CA, USA

[3] Center for Environmental Research and Technology, Riverside, CA, USA

[4] Now at Department of Civil and Environmental Engineering, University of California, Berkeley, CA, USA

1. Detailed information of emission data

Annual average NOx and VOC emissions from 2000 to 2035 in the South Coast Air Basin (SoCAB) are projected from the emissions in 2012 (Cox et al., 2013). To backcast the NOx and VOC emissions from 1990 to 2000 based on the 2012 emission inventories, first, we computed the NOx and VOC emissions ratios (the emissions were projected from the emissions in 2008) in 1990 and 1995 to the year 2000 (Cox et al., 2013; Cox et al., 2009). Furthermore, we got the adjusted emissions in 1990 and 1995 by multiplying the ratios to the emissions in 2000 that were estimated from the inventory in 2012. Finally, we used linear interpolation to compute the emissions of the years between 1990 and 1995 and between 1995 and 2000.

Table S1. List of the data sources of the variables used to build the computational models of top 30 MDA8 days from 1990 to 2019.

Kind of Variables	**Variables**	**Units**	**Data Source**
Response Variable	Top 30 MDA8 Concentrations	ppbV	CARB/ EPA
Surface Meteorology[a]	Temperature	°C	NOAA[1]/ CARB
	Wind Speed	m/s	
	Wind Direction	Degree	
	Solar Radiation[b]	W/m^2	CARB/ EPA/ NSRD
Upper Meteorology (500 and 850 millibar) [c]	Geopotential Height	m	NOAA[2]
	Temperature	°C	
	Dew Point Temperature	°C	
	Wind Speed	m/s	
	Wind Direction[d]	Degree	
	Relative Humidity[e]	%	
Estimated Emissions[f]	NOx/ VOC	Tons/day	CARB
Large-scale Climate Index	Niño 3.4 monthly indices	°C	CPC
Temporal Variable	Day of Year	None	NA
	Day of Week		

Data Source Abbreviation: **CARB**: California Air Resources Board (https://www.arb.ca.gov/aqmis2/aqdselect.php, last access May 23, 2020); **EPA**: EPA AQS air pollutant data queries (https://aqs.epa.gov/aqsweb/airdata/download_files.html, last access

May 27, 2020); **NOAA[1]**: National Oceanic and Atmospheric Administration (https://www.arb.ca.gov/aqmis2/metselect.php, last access May 27, 2020); **NOAA[2]**: National Oceanic and Atmospheric Administration (https://ruc.noaa.gov/raobs/, last access May 23, 2020); **NSRD**: National Solar Radiation Database (https://nsrdb.nrel.gov/, last access May 27, 2020); **CPC**: Climate Prediction Center (https://www.cpc.ncep.noaa.gov/data/indices/sstoi.indices, last access May 23, 2020).

a: All the surface meteorological variables were obtained from Barstow-Daggett Airport and Los Angeles International Airport (LAX).
b: To avoid the outliers in the CARB and EPA dataset and create a continuous solar radiation (SR) from 1990 to 2019, we combined the SR data at LAX from NSRD meteorological statistical model and those at Santa Clarita site/ Los Angeles N Main Street site/ Victorville Park Avenue site from CARB and EPA AQS archives. We implemented the missing SR value using the data at Joshua Tree NP Black Rock site.
c: Upper meteorological data is at the Miramar site, close to the SoCAB, and no site has sounding data in the SoCAB. The upper meteorological data at the Miramar site that follows the standard radiosonde release time is relatively more than other sites (e.g., Edwards Air Force Base (AFB), Vandenberg AFB, Point Mugu, and San Nicolas Island).
d: We used the sine of the wind direction at upper air to represent the transport direction.
e: Relative Humidity (RH) value at 500 and 850 millibar (mb) was computed through the Clausius-Clapeyron Equation (Alduchov et al., 1996; Lawrence, 2005).

$$RH = e^{\{5321\times\{(\frac{1}{T}-\frac{1}{Td})\}\}} \text{ (Equation 1)}$$

where T is air temperature and Td is dew point temperature.
f: The full description of the estimated emissions from 1990 to 2000 is given in the SI: Detailed information of emission data.

Table S2. Predictors used to test the final/ optimal GAM, MARS, SVR and RF model equations.

Kind of Variables	Variables	Abbreviation	Unit
Temporal Variables	Day of the week (factor, from Mon to Sun)	dayofweek	None
	Day of year (from 1 to 365/366)	dayofyear	None
Surface Meteorological Variables	Daily maximum surface temperature at the Barstow Airport/ LAX site	TmaxBarstow/ TmaxLAX	°C
	Daily minimum surface temperature at the Barstow Airport/ LAX site	TminBarstow/ TminLAX	°C
	Daily average wind speed at the Barstow Airport/ LAX site	AWNDBarstow/ AWNDLAX	m/s
	Max/ Mean solar radiation	SRmax/ SRmean	W/m^2
	Daily RH at 500/ 850 mb	Mir500RH	%

Upper Air Meteorological Variables		/Mir850RH	
	Daily dew point temperature at 500/ 850 mb	MirDewPtT500C /MirDewPtT850C	°C
	Daily temperature at 500/ 850 mb	MirTemp500C / MirTemp850C	°C
	Daily wind speed at 500/ 850 mb	MirWS500ms/ MirWS850ms	m/s
	Daily wind direction at 500/ 850 mb*	MirWD500/MirWD850	None
	Daily height at 500/ 850 mb	MirHeight500/ MirHeight850	m
Large-scale climate pattern	Monthly Niño 3.4 indices	ENSOmonthly	°C
Emissions	Annual averaged NOx emissions	eNOx	Tons/day
	Annual averaged VOC emissions	eROG	Tons/day

*: We used the sine of the wind direction at upper air to represent the transport direction.

Table S3. Summary of statistical results of the top30 MDA8 concentration using four methods at Crestline site.

Method	**Mean Bias (ppbV)**	**R^2**	**RMSE (ppbV)**
GAM model	-0.02	0.84	9.74
MARS model	-0.40	0.83	10.1
RF model[1]	-0.44	0.81	10.9
RF model[2]	-0.36	0.81	10.9
SVR model[1]	-1.2	0.81	10.8
SVR model[1]+tune	-0.74	0.81	10.9
SVR model[2]	-1.2	0.83	10.4

1 and 2: RF/ SVR model with the same variables as GAM model and RF/ SVR model with the optimal combination of the indicators.

Table S4. Summary of statistical results of the top 30 MDA8 concentrations using four methods at Crestline site using 10-fold cross validation (90% is training data and 10% is testing data).

Method	**Training Data**		**Testing Data**	
	R^2	**RMSE (ppbV)**	**R^2**	**RMSE (ppbV)**
GAM model	0.84	9.74	0.85	9.67
MARS model	0.83	10.3	0.83	10.2
RF model	0.80	11.0	0.82	10.3

SVR model[1]	0.81	10.9	0.81	10.4
SVR model[2]	0.82	10.4	0.8	10.6

1 and 2: SVR model with the same variables as GAM model and SVR model with the optimal combination of the indicators.

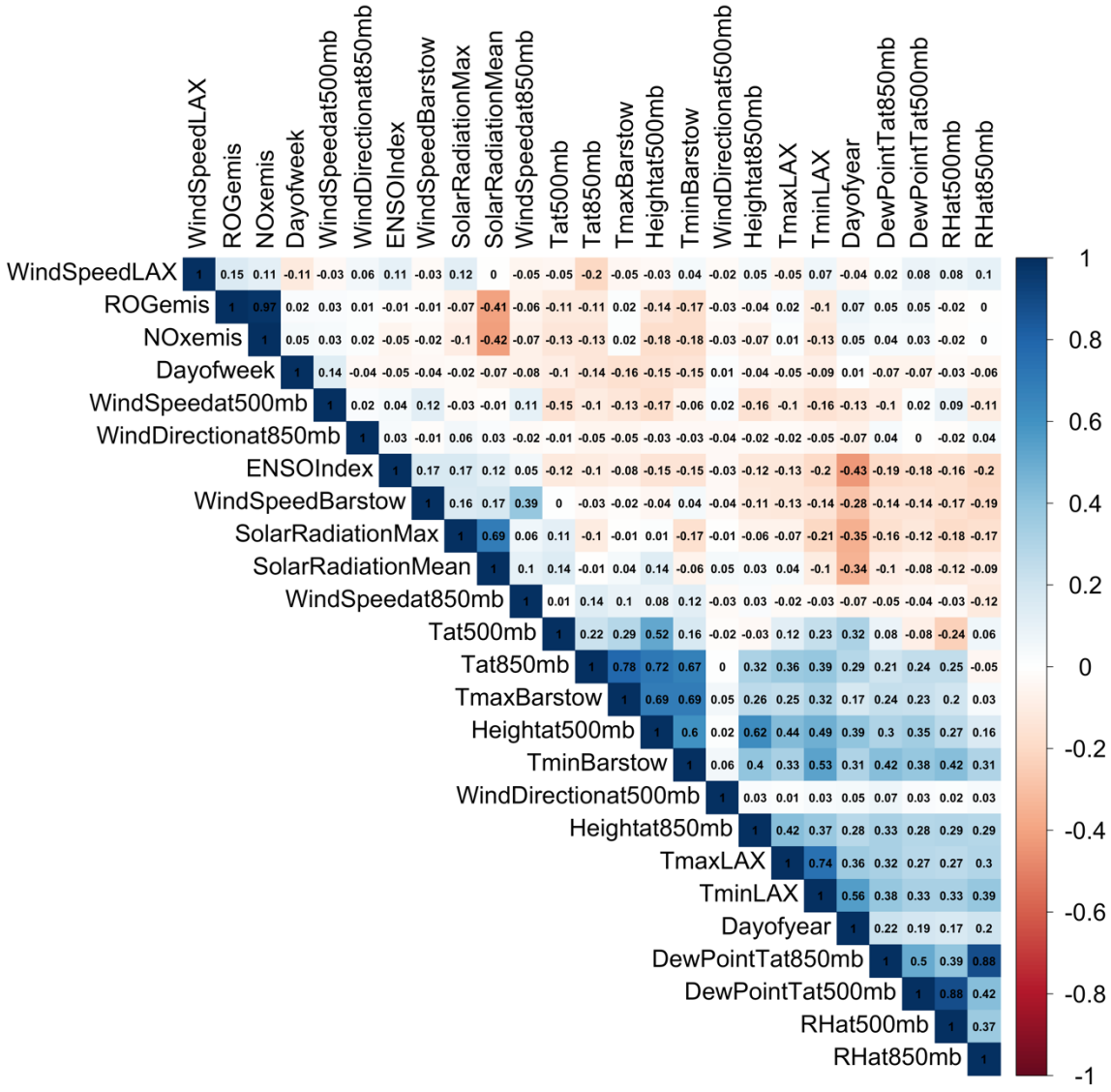

Figure S1. Correlation value among all the available independent variables.

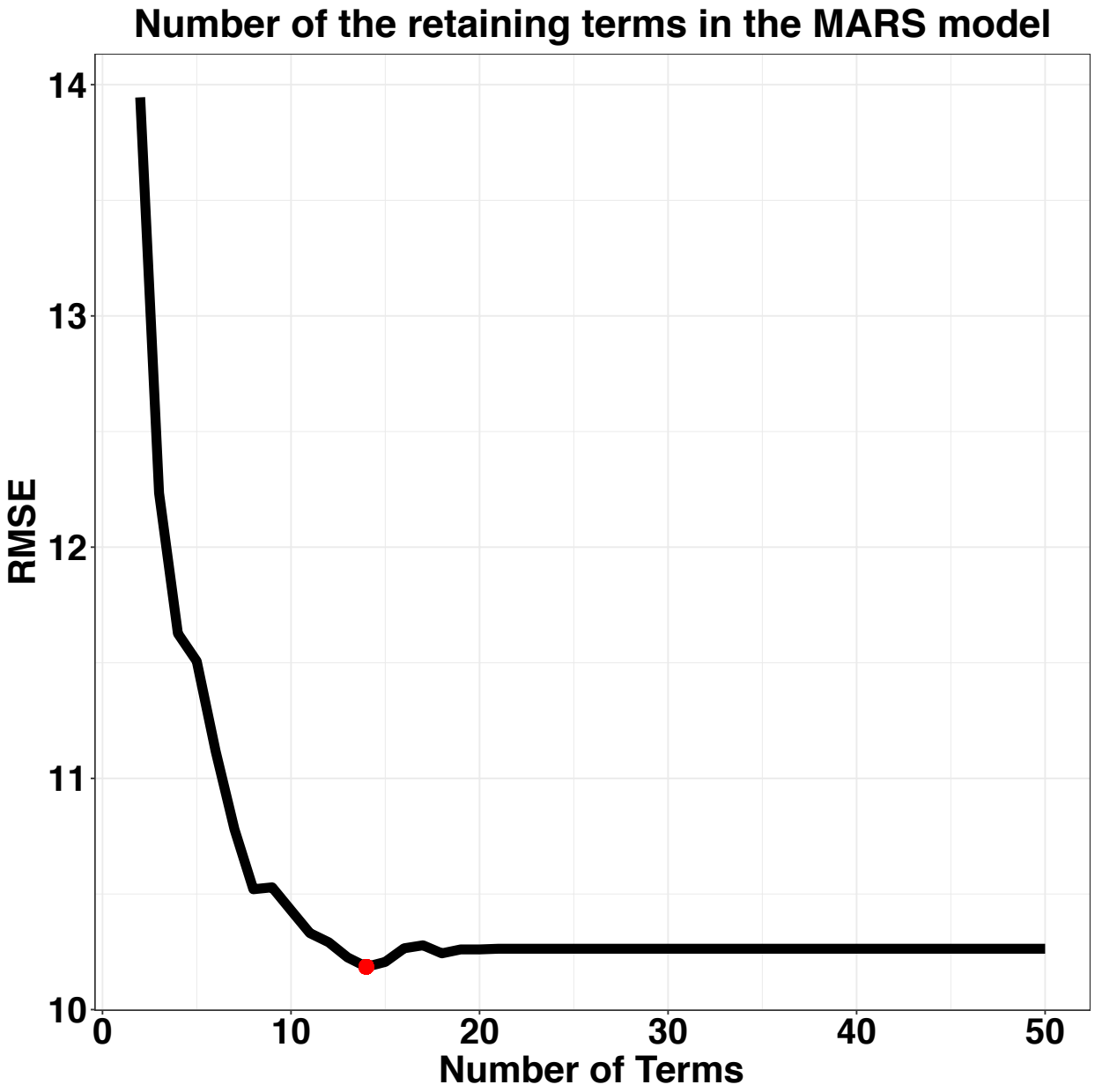

Figure S2. Number of the remaining terms in the built MARS model vs the RMSE value using 10-fold validation with the training dataset (90% of the original dataset). The red point shows the best setting that remain 14 terms in the MARS model and RMSE equals to 10.19 ppbV. The RMSE of the 16 terms MARS model is 10.27 ppbV.

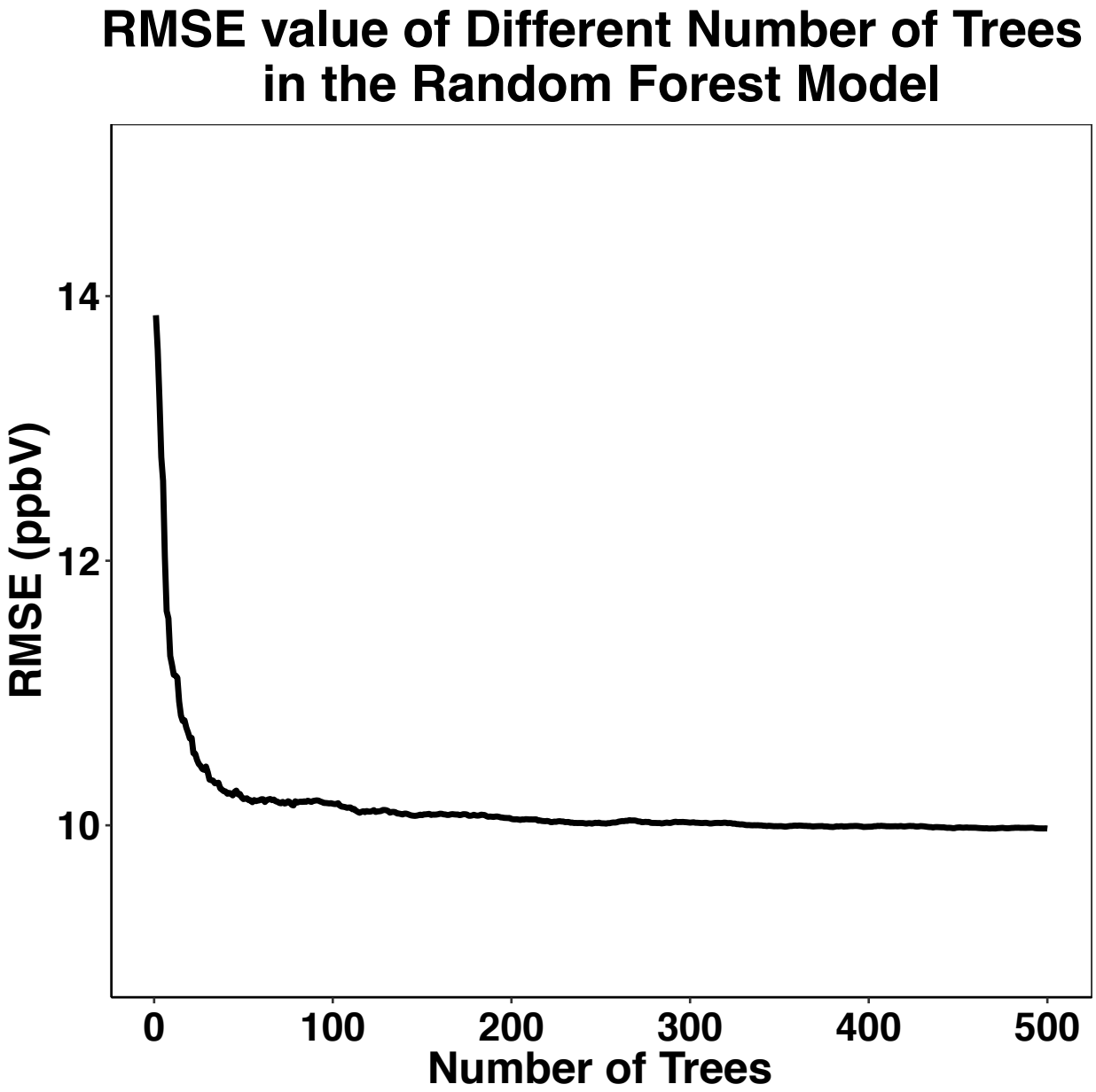

Figure S3. Number of trees in the built RF model vs the RMSE value.

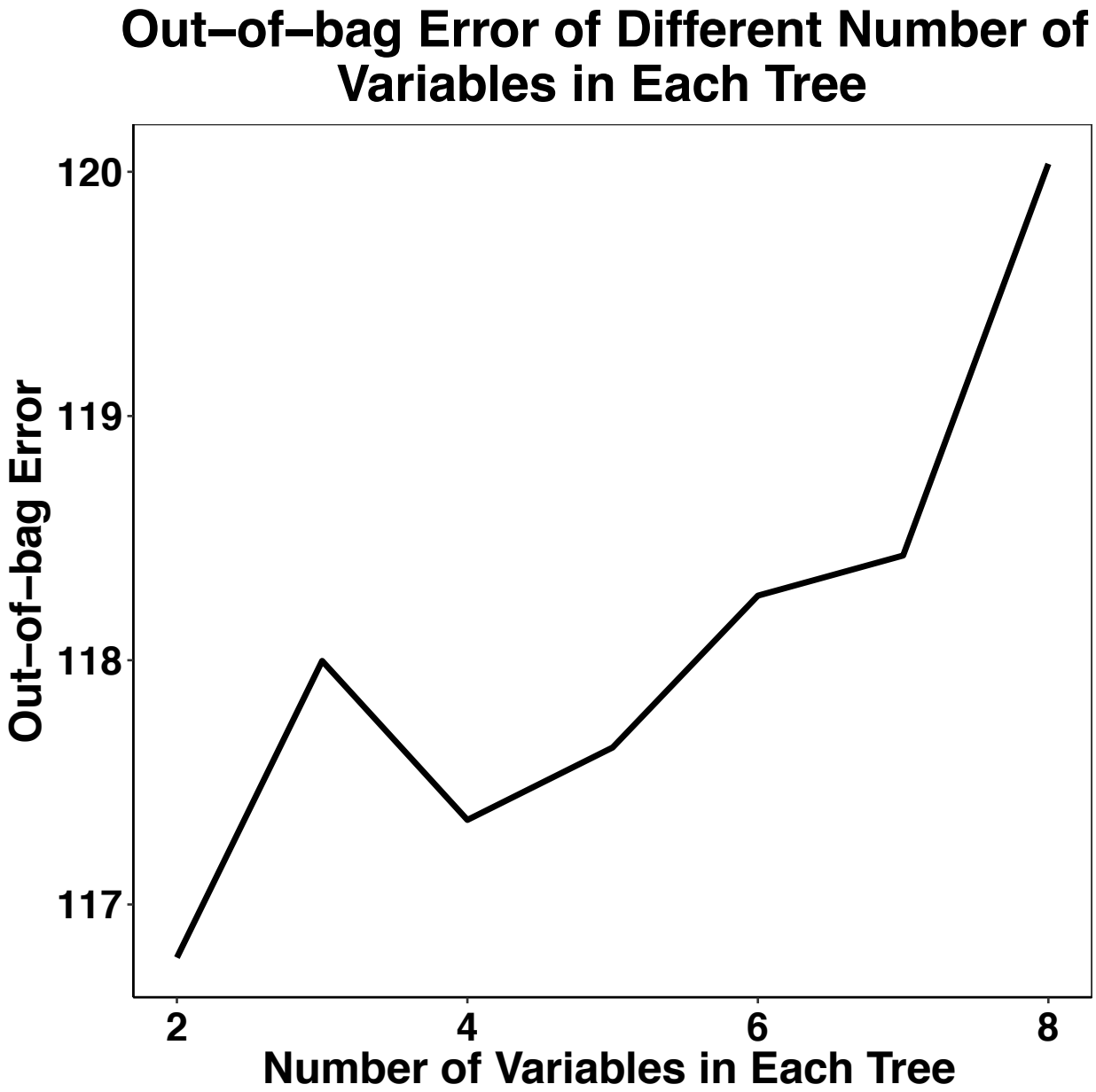

Figure S4. Number of variables in each tree of the built RF model vs the out-of-bag (OOB) value.

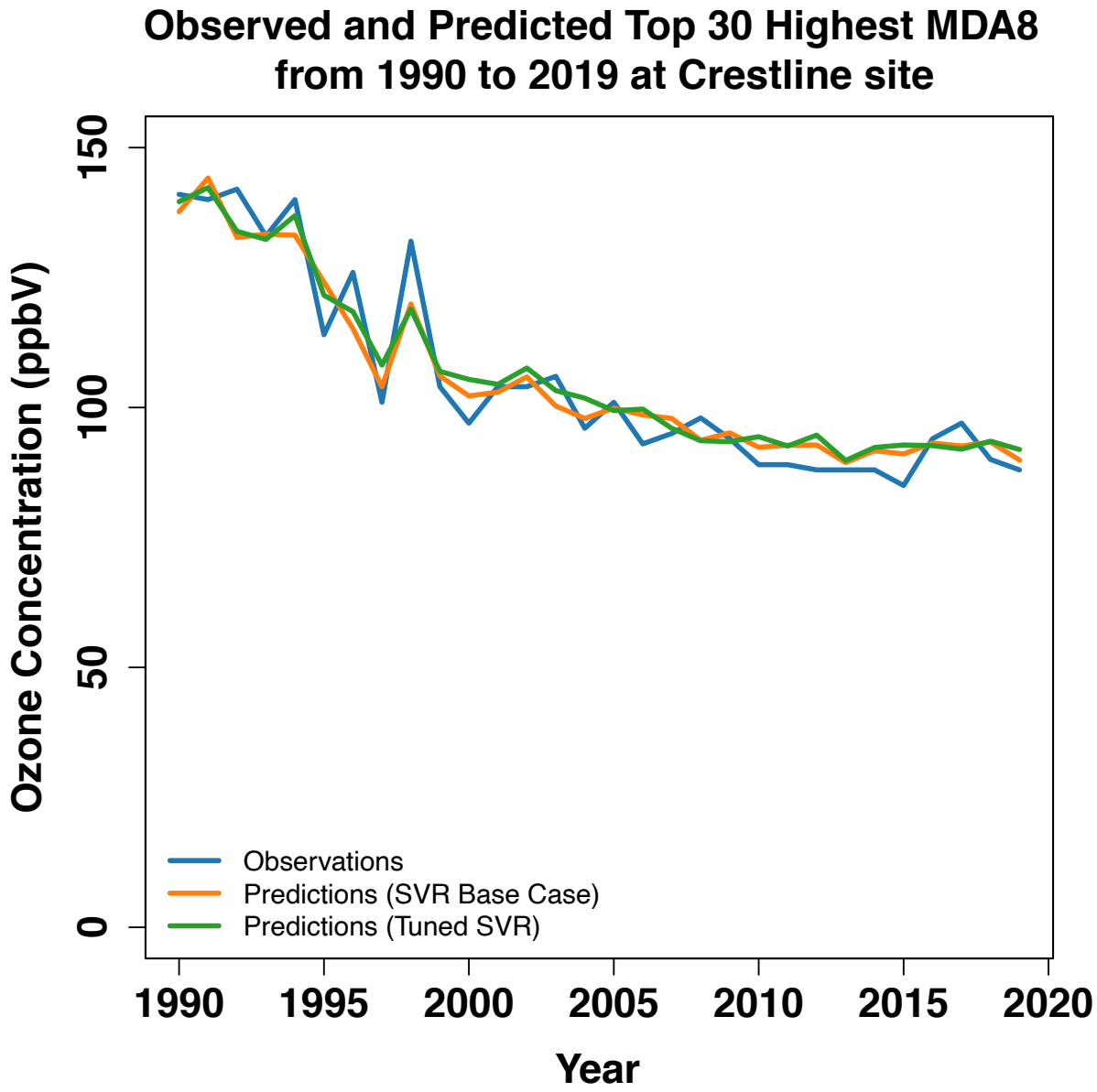

Figure S5. Observed (blue) and predicted top 30 MDA8 concentrations using original (orange) and tuned (green) SVR models from 1990 to 2019 at Crestline site.

References

Alduchov, O. A., & Eskridge, R. E. (1996). Improved Magnus Form Approximation of Saturation Vapor Pressure. *Journal of Applied Meteorology, 35*(4), 601-609.

Cox, P., Delao, A., & Komorniczak, A. (2013). The California almanac of emissions and air quality. *California Air Resources Board, Sacramento, CA*.

Cox, P., Delao, A., Komorniczak, A., & Weller, R. (2009). The California almanac of emissions and air quality. *California Air Resources Board, Sacramento, CA*.

Lawrence, M. G. (2005). The relationship between relative humidity and the dewpoint temperature in moist

air - A simple conversion and applications. *Bulletin of the American Meteorological Society, 86*(2), 225-233.