PNAS
www.pnas.org

Supplementary Information for

From COVID-19 to Future Electrification: Assessing Traffic Impacts on Air Quality by a Machine Learning Model

Jiani Yang, Yifan Wen, Yuan Wang, Shaojun Zhang, Joseph Pinto, Elyse A. Pennington, Zhou Wang, Ye Wu, Stanley P. Sander, Johnathan H. Jiang, Jiming Hao, Yuk L. Yung, John H. Seinfeld

Email: yuan.wang@caltech.edu, zhsjun@tsinghua.edu.cn, seinfeld@caltech.edu

This PDF file includes:

Supplementary Information

Materials

The California Ambient Air Monitoring Network. The California Ambient Air Monitoring Network consists of more than 250 monitoring stations operated by federal, state, and local agencies. The network has been dedicated to measuring ambient concentrations of criteria pollutants, including ground-level ozone (O_3), particulate matter (PM_{10} and $PM_{2.5}$), nitrogen dioxide (NO_2), sulfur dioxide (SO_2) and lead (Pb). It also measures meteorological data such as dew point, temperature, pressure, precipitation, surface radiation fluxes, relative humidity, and wind (1). The distribution of the California Ambient Air Monitoring Network in the LA Basin is shown in Fig S2. In this study we took the ground-based measurement data of $PM_{2.5}$, NO_2, O_3 as observations for the training and validation of the machine-learning model. We took the measurements of temperature, wind, pressure, radiation, and relative humidity as the model input.

ERA5. ERA5 provides hourly estimates of a large number of atmospheric, land and oceanic climate variables (https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5). The data cover the Earth on a 30 km grid and resolve the atmosphere using 137 levels from the surface up to a height of 80 km. ERA5 includes information about uncertainties for all variables at reduced spatial and temporal resolutions. In this study, we took the data of boundary layer height and precipitation as model input.

Traffic Data. The real-time based freeway traffic dataset was obtained from the California Department of Transportation (PeMS, http://pems.dot.ca.gov). PeMS collects data from varies types of vehicle detector stations, including inductive loops, side-fire radar and magnetometers (2), PeMS can provide flow and speed as reported by detectors over several years. It also supports integration with common internet-based mapping service (e.g., Google Maps, Google Earth) (2). Here a python-based script was created to obtain the data of PeMS collocated with freeways around the Los Angeles (LA) basin. There are approximately 3144 sensors in this area. The distribution of PeMS sites within the LA basin is shown in Fig S10. In this study, the hourly truck and non-truck flow data from Jan 1st, 2019 to July 1st, 2020 have been used to study the change of traffic patterns in hourly, daily and monthly periods. In addition, daily total traffic volume data from 4437 arterial and residential roads within the research domain in 2019 were obtained from the Los Angeles Department of Transportation (LADOT) to provide the traffic patterns for low-ranking roads.

A random forest (RF, see technical descriptions in Methods) model served as a platform to simulate the spatial distribution of link-level traffic volumes for the entire road network of the research domain. Annual average daily traffic (AADT) of all monitored roads served as dependent variables. Functional classification (i.e., freeway, arterial road and residential road) of the road, average AADT and total length of different road types within 1 km circular buffers served as independent variables to train the RF model. Based on the spatial distribution pattern simulated by the RF model and the hourly observations from 3144 highway traffic monitoring sites, we allocated the total traffic volumes for the entire road network of the research domain from Jan 1st, 2019 to July 1st, 2020. The split between non-truck and truck fleet for each link, as well as the hourly variation, were based on hourly traffic volume data from the nearest highway traffic monitoring site. Vehicle Mileage Traveled ($\mathrm{VMT} = \sum \text{traffic volume} \times \text{road length}$) was calculated to represent traffic activities for both non-truck and truck fleet for each road segment, which would be further used as predicting variables in estimating concentrations air pollutants.

Points of Interest. The points of interest were taken from the previous methane inventory studies – Vista which is a Geographic Information System (GIS) based approach to map potential methane emissions source in the LA Basin (3). The Vista-LA datasets are available from the Oak Ridge National Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics (ORNL DAAC; https://doi.org/10.3334/ORNLDAAC/1525). The spatial distribution of all parameters in this study has been shown in Fig S2.

CMIP6 Climate Projection. The most recent Coupled Model Inter-comparison Project Phase 6 (CMIP6) is a project of the World Climate Research Programme (WCRP)'s Working Group of Coupled Modelling (WGCM) (https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6). CMIP6 model simulations have been regularly assessed as part of the IPCC Climate Assessments Reports and various national assessments. In this study, we took the ensemble mean of multiple runs from six CMIP6 models that predicts wind direction, wind speed, precipitation, solar radiation, surface air pressure, temperature, relative humidity as the meteorological predictions for 2035 and 2050. The monthly meteorological ratio patterns of 2035/2019 and 2050/2019 are shown in Fig S9.

Extended Methods

Development of RF Prediction Model. The key input to the machine learning model and related experiments is summarized in Fig S1. RF is a machine learning method based on decision trees (4). It constructs each tree by bootstrapping (random resampling with replacement) and splits each point in the tree according to the best of a subset of randomly chosen predictors at each point. Versus traditional parametric modelling techniques, RF presents some advantages. Firstly, it can handle non-linear relationships with a low likelihood of overfitting. RF algorithm also copes with highly correlated variables since once a variable has been selected to build a tree from the subset of variables, a highly correlated variable will be less likely selected to grow the tree (5). It also gives higher predictive accuracy compared to parametric techniques, and provides information about the underlying mechanism reporting as variable importance of the predictor variables (4). Considering the advances above, we chose RF as a platform to model air pollutant concentrations, and then evaluate the contribution of traffic decline to air quality changes during COVID-19 lockdown by applying traffic-as-usual scenarios to the model.

Three RF models for hourly NO_2, O_3, and $PM_{2.5}$ concentrations were developed based on site-specific air quality, meteorological and traffic data at hourly accuracy from 1 Jan 2019 to 30 June 2020, as well as up-to-date demographic and land use data (all predictors are listed in Table S2). Meteorological data for each air quality site (AQS) were extracted from the nearest meteorological monitoring site (i.e., the California Ambient Air Monitoring Network and ERA5). Traffic and demographic variables were calculated around each AQ site in 10 circular regions with radii of 50 m to 5 km as either an average (e.g., population density) or sum (e.g., vehicle mileage traveled, VMT) within each buffer. In particular, non-truck and truck fleets were considered separately in the prediction parameters of the RF models. Distance to importance point of interest (POIs) serves as land use indictors that reflect the potential impact from nearby industrial manufacturers or transportation facilities. In order to allow the RF model to utilize the temporal correlations in the data, we considered temporal variables such as day of a week and public holidays in model training.

RF models for NO_2, O_3, and $PM_{2.5}$ concentrations were trained in R (a programming language) using the *ranger* package. Two user-defined parameters: the optimum number

of trees (ntree) and variables randomly tried at each split in the random forests (mtry) were determined by maximizing the out-of-bag (OOB) correlation coefficient squared (R^2) calculated as 1-MSE/var(Y), where Y is the observed values and MSE is the mean of the OOB errors for all the prediction points. Then the model performance was validated using a 5-fold cross validation scheme. R^2 and root mean squared error (RMSE) between the predictions and observations were computed as indicators of the model performance. A key advantage of the RF model over alternative machine learning algorithms is the ability to measure variable importance. We utilized Gini importance, a commonly-used indicator that could be calculated in R (i.e., *ranger*) (6), to rank the important factors affecting the air quality. Gini importance is computed by the sum of Gini impurity decrease for all nodes in the forest when a split on one variable has been conducted (7, 8). A split with a large decrease of Gini impurity is considered important, and variables used for splitting at important splits are also considered important. Thus, higher Gini importance indicates higher variable importance.

To account for the secondary $PM_{2.5}$ formations due to atmospheric oxidation processes, we experimented to include O_3 as a predictor in the $PM_{2.5}$ model. The results indicate that including concurrent O_3 concentrations would much improve R^2 from 0.53 to 0.65, which has been used in the $PM_{2.5}$ RF model (see Fig S11(a) and (b)). However, including O_3 concentrations does not significantly increase R^2 for NO_2 prediction (see Fig S11(c) and (d)). Thus, we opt to not consider O_3 concentrations in predicting NO_2 concentrations.

Quantifying Impacts of Traffic Decline on Air Quality During COVID-19 Lockdown. We established two prediction scenarios (i.e., Normal Traffic and Normal Truck) assuming the traffic activities of the entire fleet and the truck fleet follow the average weekly and diurnal patterns from 1 Jan to 4 Mar, 2020 (see Table S3). Then we predicted the hourly pollutant concentrations during COVID-19 using the established RF models based on the assumed traffic activities and the actual meteorological conditions and calculated the difference ratio between predicted and actual concentrations to represent the impact of total traffic and truck fleet respectively (Eq. 1 and 2).

$$\text{Traffic impact ratio} = \frac{\text{Observation-Prediction of Normal traffic scenario}}{\text{Prediction of Normal traffic scenario}} \quad \text{Eq. 1}$$

$$\text{Truck impact ratio} = \frac{\text{Observation-Prediction of Normal truck scenario}}{\text{Prediction of Normal truck scenario}} \quad \text{Eq. 2}$$

Projection of Future Air Pollution Changes by Traffic and Climate

The RF models were also utilized to predict the ambient concentrations of $PM_{2.5}$, O_3 and NO_2 in the future based on the predicted meteorological profiles and traffic emissions. We analyze CIMIP6 simulations to predict the future meteorological changes in 2035 and 2050 comparing with the pattern in 2019. We use output from six different climate models for the predication of key parameters such as wind speed, wind direction, precipitation, solar radiation at surface, surface air pressure, surface temperature and relative humidity, as listed in Table S4. To minimize the influence of large interannual variations of climate states in the analysis, we use 6-year means of 2015-2020, 2030-2035, 2045-2050 to represent 2019, 2035, 2050, respectively. We first take grids from our research domain (latitude: 33.68N - 34.22N, longitude: 118.60W - 117.39W), and then calculate the

ensemble mean of each parameter for each model. We then calculate the ratio between 2035 and 2019 and the ratio between 2050 and 2019 for wind speed, precipitation, solar radiation at surface, surface air pressure, surface temperature, and relative humidity. We also calculate the difference between 2035 and 2019 and the difference between 2050 and 2019 for wind direction. In order to calculate the ratio mean and variance from all models for a specific month, we take the ratio between 2035 and 2019 as an example: The difference percentage ratio between 2019 and 2035 for a random variable in a specific model can be calculated as:

$$y_{n_i} = \left(\frac{x_{n_{i_{2035}}}}{x_{n_{i_{2019}}}} - 1\right) \times 100\% \quad \text{Eq. 3}$$

Here, y_{n_i} represents the difference percentage ratio in month i of 2035 compared with the month i of 2019 for model n for a random variable; $x_{n_{i_{2035}}}$ is a random variable value for the model n in the month i of 2035, $x_{n_{i_{2019}}}$ is the variable value for the model n in month i of 2035.

We followed the method in Kendall's Advanced Theory of Statistics (9) and Survival Models and Data Analysis (10). Given random variables in month i of 2019 and 2035: $X_{i_{2035}}$ and $X_{i_{2019}}$ where X_{2019} either has no mass at 0 (discrete) or has support [0, ∞). Let $G = g(X_{i_{2035}}, X_{i_{2019}})$. The approximations for the mean of the ratio of random variable X in month i of 2035 and 2019 $E(X_{i_{2035}}/X_{i_{2019}})$after 2 Taylor expansion can be improved approximately as:

$$E(X_{i_{2035}}/X_{i_{2019}}) \approx \frac{\mu_{x_{i_{2035}}}}{\mu_{x_{i_{2019}}}} - \frac{Cov(X_{i_{2035}}, X_{i_{2019}})}{\left(\mu_{x_{i_{2019}}}\right)^2} + \frac{Var\left(\mu_{x_{i_{2019}}}\right)\mu_{x_{i_{2035}}}}{\left(\mu_{x_{i_{2019}}}\right)^3} \quad \text{Eq. 4}$$

$\mu_{x_{i_{2035}}}$ and $\mu_{x_{i_{2019}}}$ are the mean the random variable X in month i of 2035 and 2019; $\mathrm{Cov}(X_{i_{2035}}, X_{i_{2019}})$ is the covariance of random variable X in month i of 2035 and 2019; $\mathrm{Var}(\mu_{x_{i_{2019}}})$ is the variance of random variable X in month i of 2019.

The approximations of the variance of the ratio of random variable X in month i of 2035 and 2019 $\mathrm{Var}(X_{i_{2035}}/X_{i_{2019}})$ after 2 Taylor expansion can be improved approximately as:

$$Var(X_{i_{2035}}/X_{i_{2019}}) \approx \frac{(\mu_{x_{i_{2035}}})^2}{(\mu_{x_{i_{2019}}})^2}\left[\frac{Var\left(\mu_{x_{i_{2035}}}\right)}{\left(\mu_{x_{i_{2035}}}\right)^2} - 2\frac{Cov(X_{i_{2035}}, X_{i_{2019}})}{\mu_{x_{i_{2035}}}\mu_{x_{i_{2019}}}} + \frac{Var\left(\mu_{x_{i_{2019}}}\right)}{\left(\mu_{x_{i_{2019}}}\right)^2}\right] \quad \text{Eq. 5}$$

$Var(\mu_{x_{i_{2035}}})$ is the variance of random variable X in month i of 2035. According to equation 3, in month i the average difference percentage ratio between 2019 and 2035 for all models can be summarized as:

$$E(X_{i_{2035}}/X_{i_{2019}} - 1) \approx \left[\frac{\mu_{x_{i_{2035}}}}{\mu_{x_{i_{2019}}}} - \frac{Cov(X_{i_{2035}}, X_{i_{2019}})}{\left(\mu_{x_{i_{2019}}}\right)^2} + \frac{Var\left(\mu_{x_{i_{2019}}}\right)\mu_{x_{i_{2035}}}}{\left(\mu_{x_{i_{2019}}}\right)^3} - 1\right] \times 100\% \quad \text{Eq. 6}$$

In month i the variance of difference percentage ratio between 2019 and 2035 for all models can be summarized as:

$$Var(X_{i_{2035}}/X_{i_{2019}} - 1) \approx \frac{(\mu_{x_{i_{2035}}})^2}{(\mu_{x_{i_{2019}}})^2}\left[\frac{Var\left(\mu_{x_{i_{2035}}}\right)}{\left(\mu_{x_{i_{205}}}\right)^2} - 2\frac{Cov(X_{i_{2035}},\ X_{i_{2019}})}{\mu_{x_{i_{2035}}}\mu_{x_{i_{2019}}}} + \frac{Var\left(\mu_{x_{i_{2019}}}\right)}{\left(\mu_{x_{i_{2019}}}\right)^2}\right] \times 10\% \quad \text{Eq. 7}$$

The calculation result for equations 6 and 7 can be found at Fig S9.

For future traffic profiles, we follow the RF model framework by using the change of traffic VMT to approximate the change of traffic emissions (truck and non-truck separately). The California Air Resources Board (CARB)'s official traffic emission model, EMFAC (11), provides the on-road emission inventories of four LA Basin counties (Los Angeles, Orange, San Bernardino and Riverside) from calendar year 2000 to 2050. The EMFAC model has specific emission estimation for more detailed vehicle category and aggregated the emission results into truck and non-truck fleets. Nitrogen oxides (NO_X) and total organic gases (TOG) are known as two of the most important traffic-related pollutants in the atmospheric chemical reactions (12). Of note, the default EMFAC emission inventories have mild ratios of zero-emission vehicles for future years (e.g., 4.4% for non-truck and none for truck in 2035). In addition, CARB has adopted the new low-NO_x omnibus regulation that will reduce NO_x emission limit by 90% for new trucks from 2027. However, due to the large presence of out-of-state trucks (more than 60% in 2017) operating in California and a considerable fraction of low-load driving conditions (e.g., speed below 25 mile per hour), CARB estimates that the low-NO_x omnibus regulation would reduce by 17% and 29% heavy-duty vehicle NO_x emissions in 2035 and 2050, respectively, compared with the original 2017 version of EMFAC emission inventories (13). We have updated the truck emission inventories with considerations of the future heavy-duty vehicle NOx emission standard. Based on the updated result of EMFAC inventories, we found the relative emission ratio of NO_X and TOG are very similar in the future years (Fig. S5 A and B). Thus, using the 2019 emissions as baseline, the change of traffic emissions in the future years was calculated as the average relative emissions of NO_X and TOG for truck and non-truck fleets respectively (Fig. S5 C).

In addition, we introduced three degrees of fleet electrification (also include other zero-emission vehicles like hydrogen fuel cell vehicles) based on the default EMFAC emission inventories (see Table S1 for scenario description). The moderate electrification scenario (E1) for trucks was determined based on CARB's Advanced Clean Truck (ACT) Program (14). CARB's estimate suggests the ACT regulation would result in approximately 5% and 10% NO_x emission reductions in 2035 and 2050 compared with the original EMFAC results, which would be considered as the overall percentages of electrified mileage of truck fleets. We note that the highest electrification scenario (E3, 80% for non-truck and 40% for truck in 2050) projects less mileage penetration of zero-emission vehicles (mostly non-truck) proposed by the L.A.'s Green New Deal (15), because higher electrification scenarios might be beyond the application boundary of air quality prediction models which is determined by the range of training dataset. However, in 2050, the electrification rate of non-truck mileage under E3 (80%) is not far away from CARB's projection for Advanced Clean Cars program (i.e., 80% population ratio for battery electric vehicles and hydrogen fuel cell vehicles, and 13% for plug-in hybrid electric vehicles) (16). The relative emission ratios of truck and non-truck emissions in 2035 and 2050 comparing with 2019 for three electrification scenarios are shown in Fig S7.
We used Monte Carlo simulations to analyze the probable uncertainty due to the variability in predicting meteorological and traffic profiles. As there is no prior distribution of future meteorology and emissions, we assumed normal distributions for all parameters. Random sampling was repeated for 100 times according to the normal distribution in each scenario.

The means of repeating predictions were calculated as final results and the standard deviations are shown as error bars.

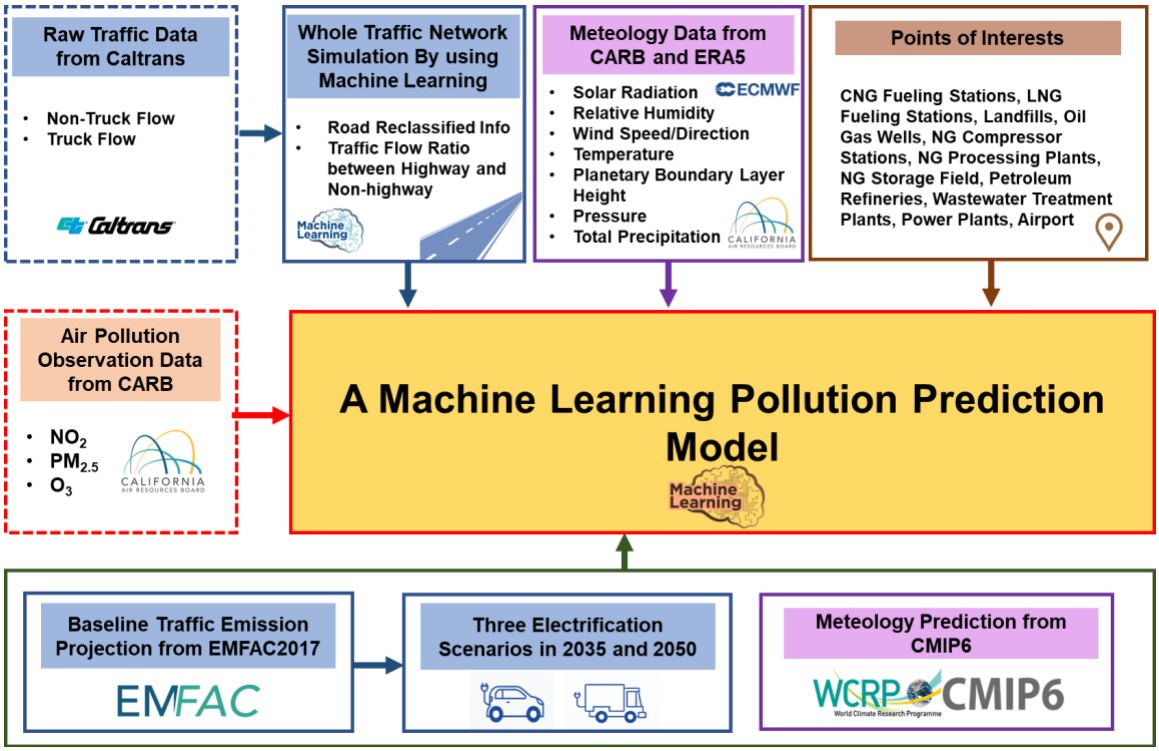

Fig. S1. A flow chart of the methodology in this study.

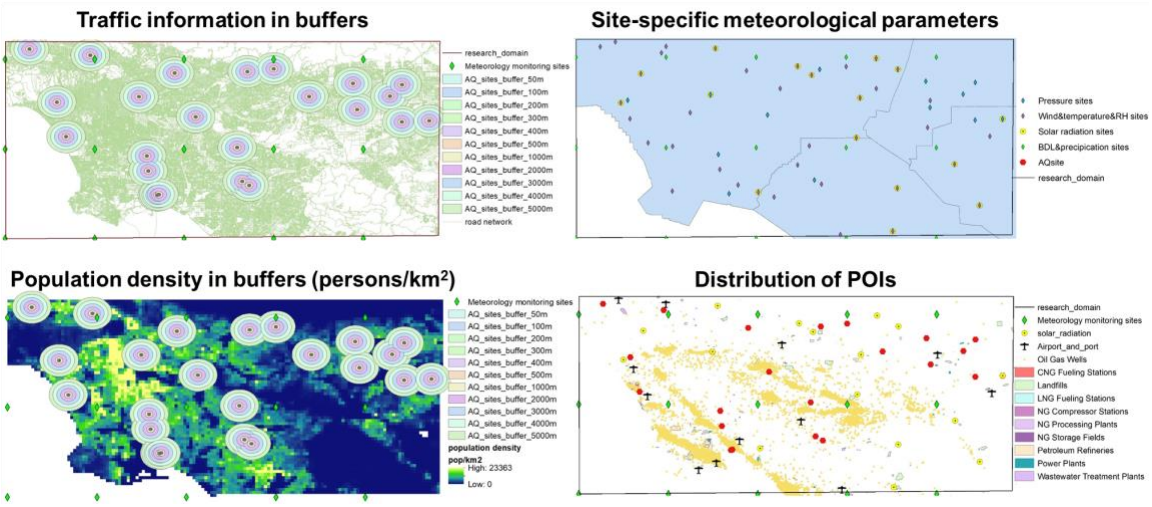

Fig. S2. Spatial distribution of factors contributing to the RF model.

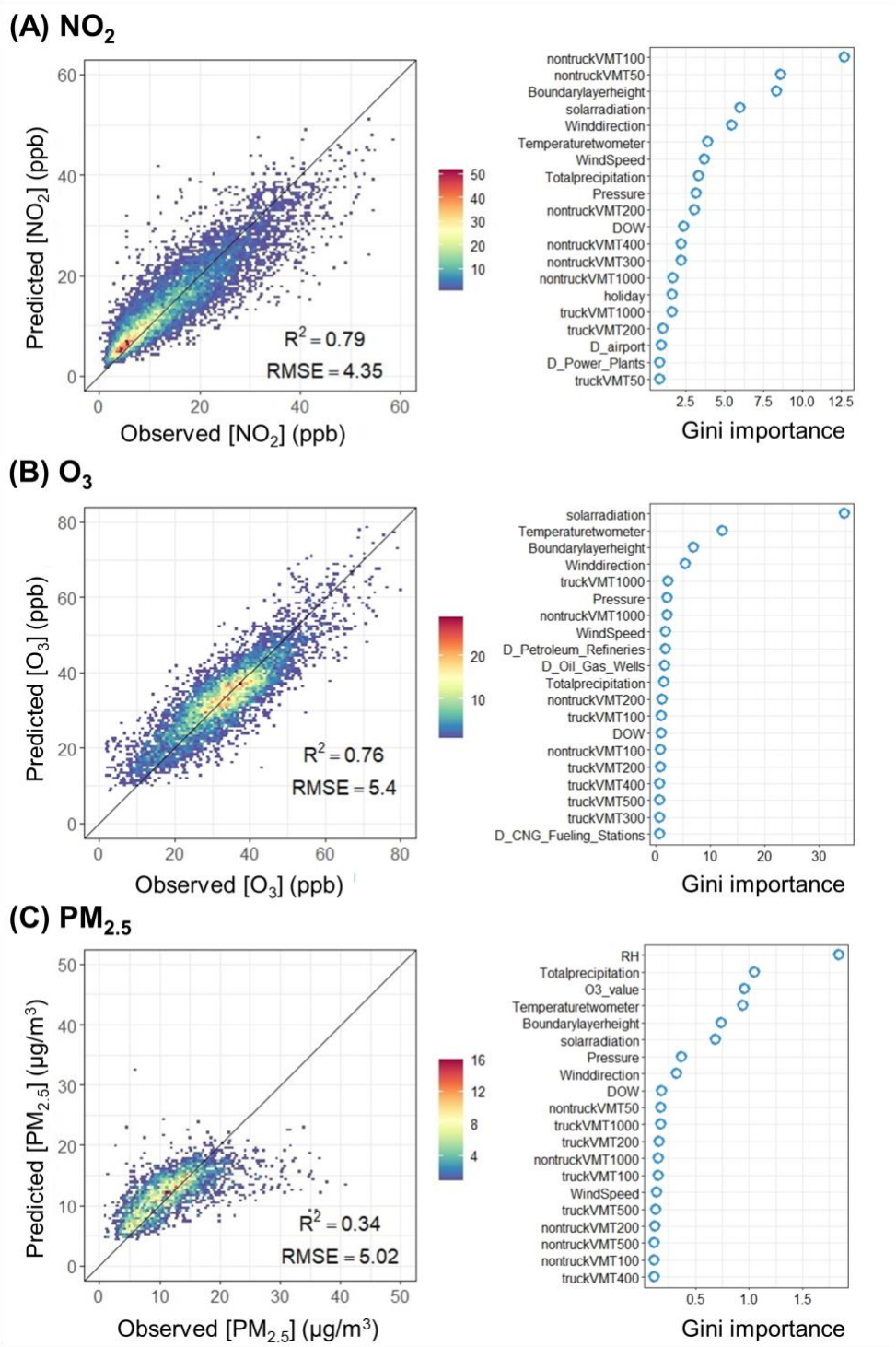

Fig. S3. Similar as Fig. 1 but based on daily average data.

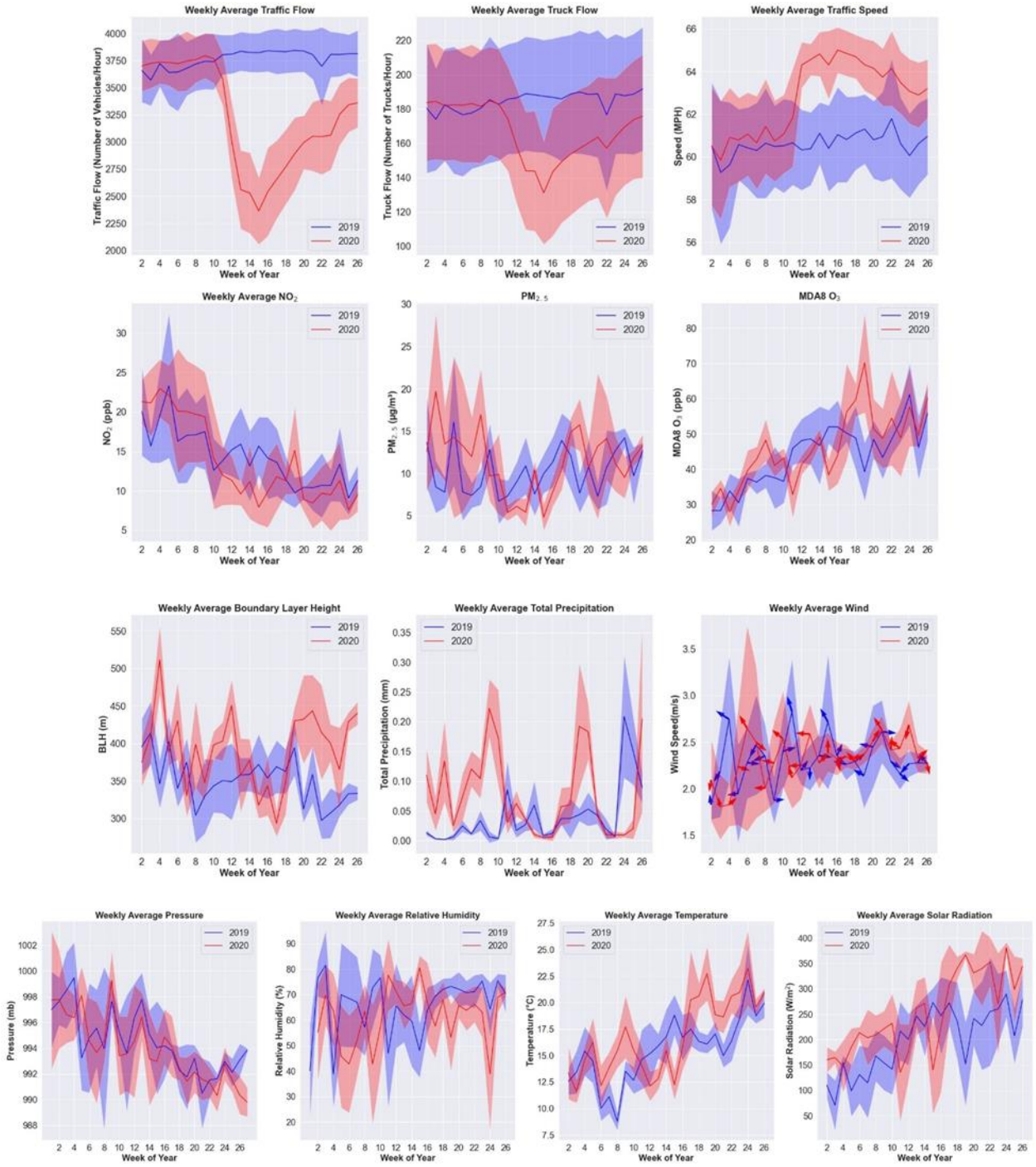

Fig. S4. Observed changes in LA traffic and air pollution during the COVID-19.

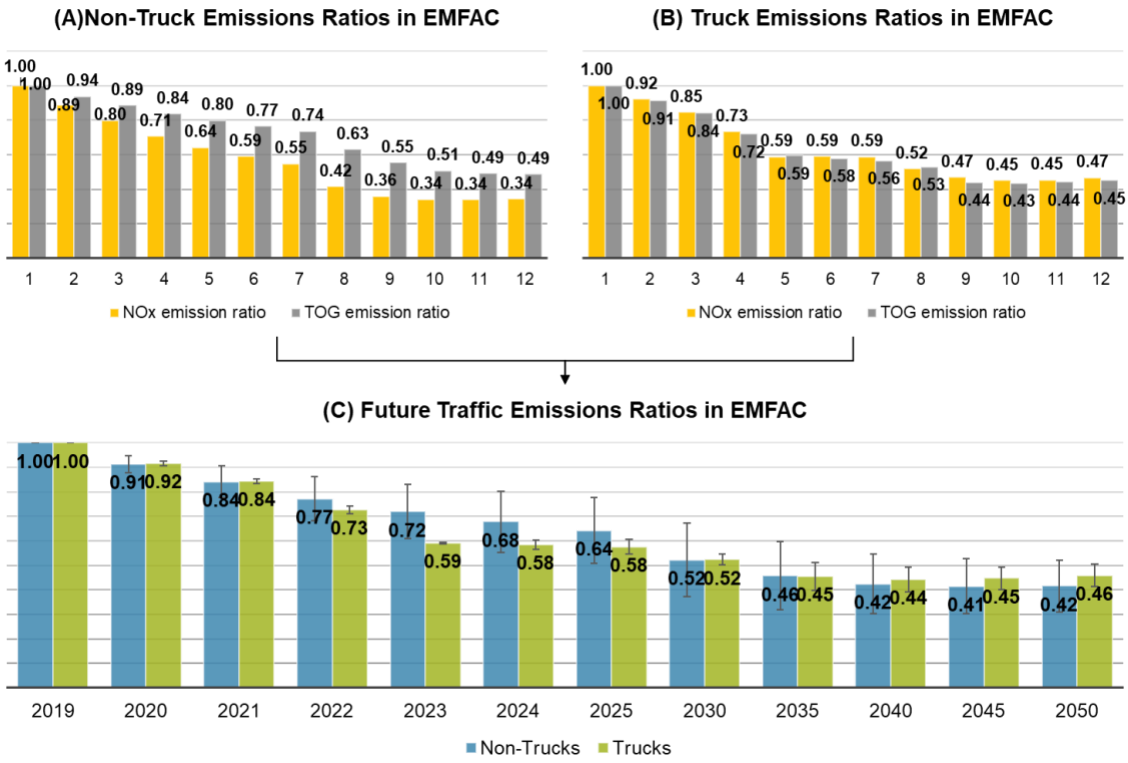

Fig. S5. Prediction of future emission ratios of truck and non-truck fleets relative to 2019 based on the result of EMFAC inventories. Columns in (C) are average relative emission ratios of most important traffic-related pollutants (i.e., NO_X and TOG) and the error bars represent standard variations among pollutants. For future truck emissions, the estimated benefits from the low-NOx omnibus regulation by CARB are considered in panels (B) and (C)**.**

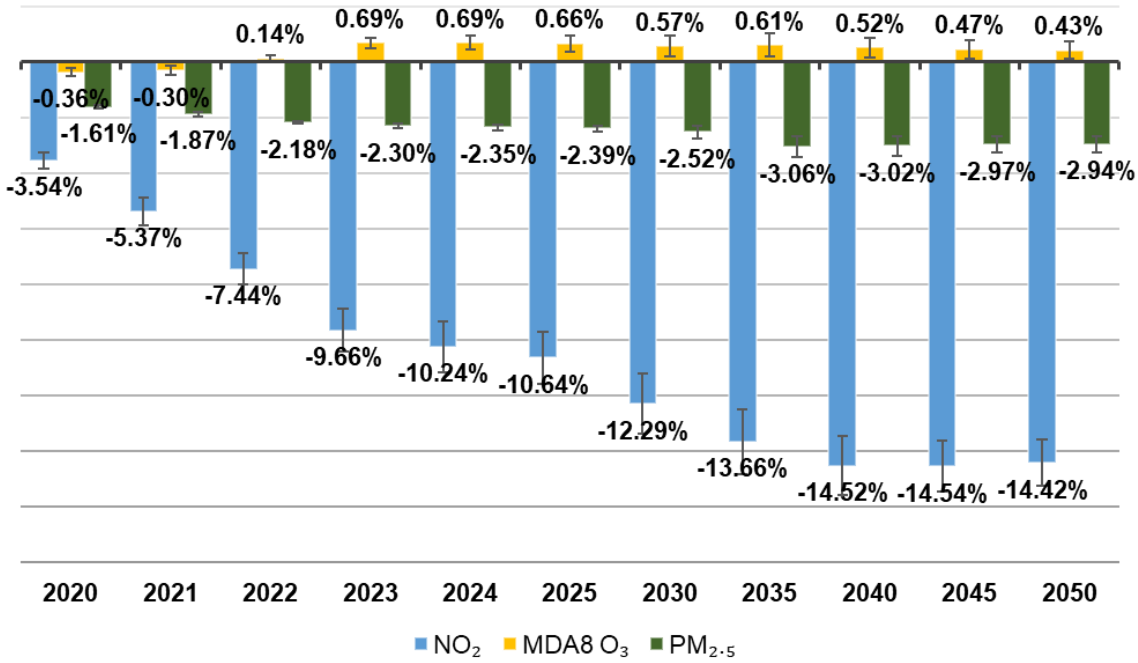

Fig. S6. Reduction ratios of NO_2, MDA8 O_3 and $PM_{2.5}$ concentrations in different calendar years relative to 2019 based on baseline traffic emissions from EMFAC. The error bars represent uncertainty of model predictions calculated by the Monte Carlo Method. Random sampling was repeated for 100 times considering uncertainty of traffic emissions in prediction of each calendar year.

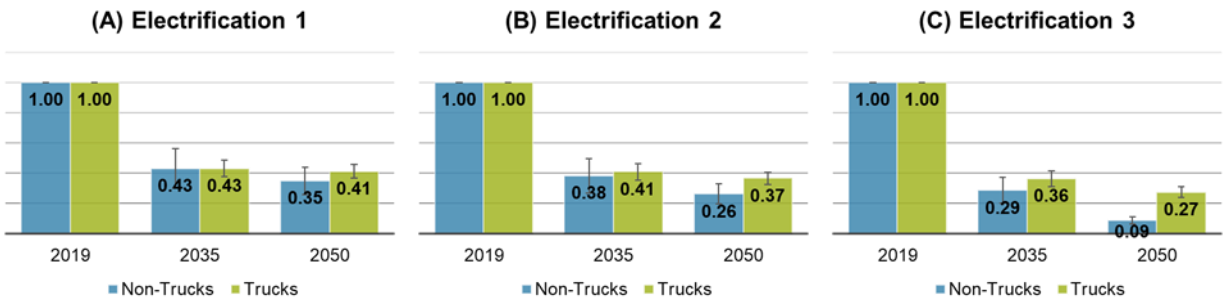

Fig. S7. Future emission ratios in 2035 and 2050 relative to 2019 under different electrification scenarios. columns are average relative emission ratios of most important traffic-related pollutants (i.e., NO_X and TOG) and the error bars represent standard variations among pollutants**.**

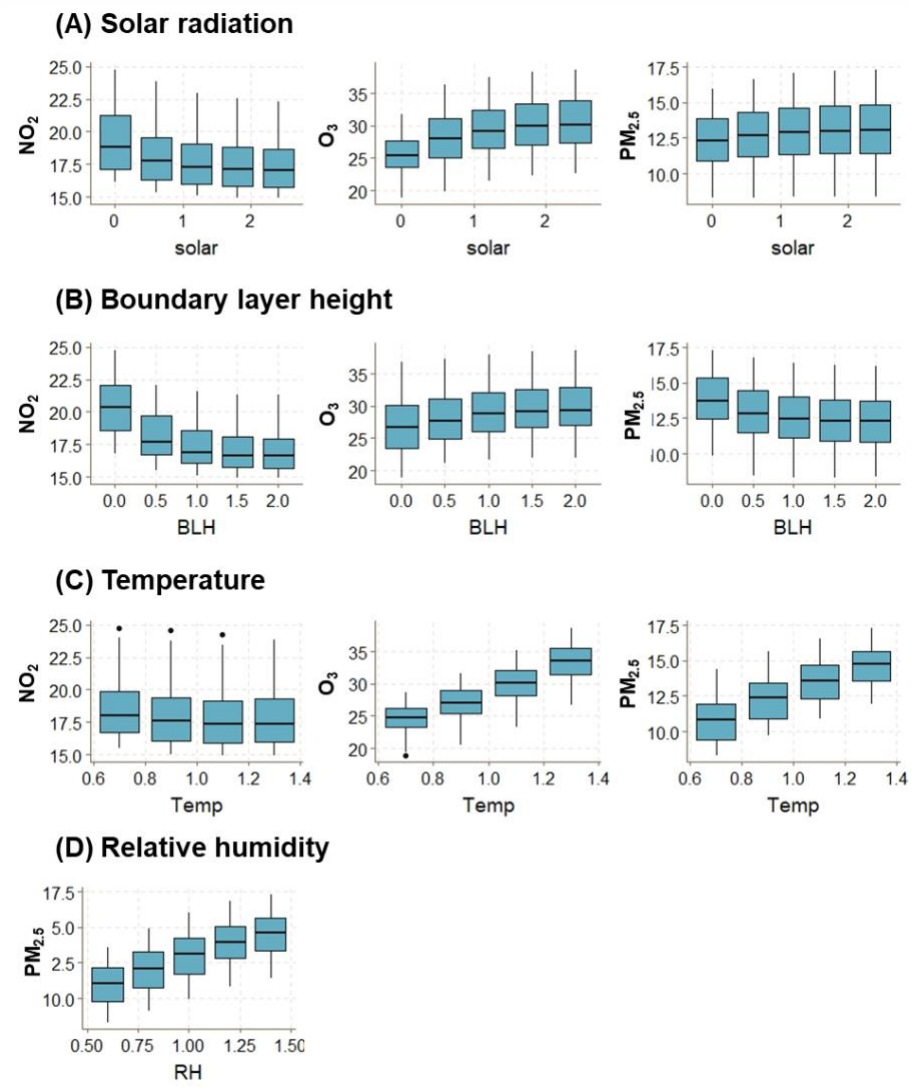

Fig. S8. Responses of pollutants to the perturbations on four key meteorological factors in the RF model.

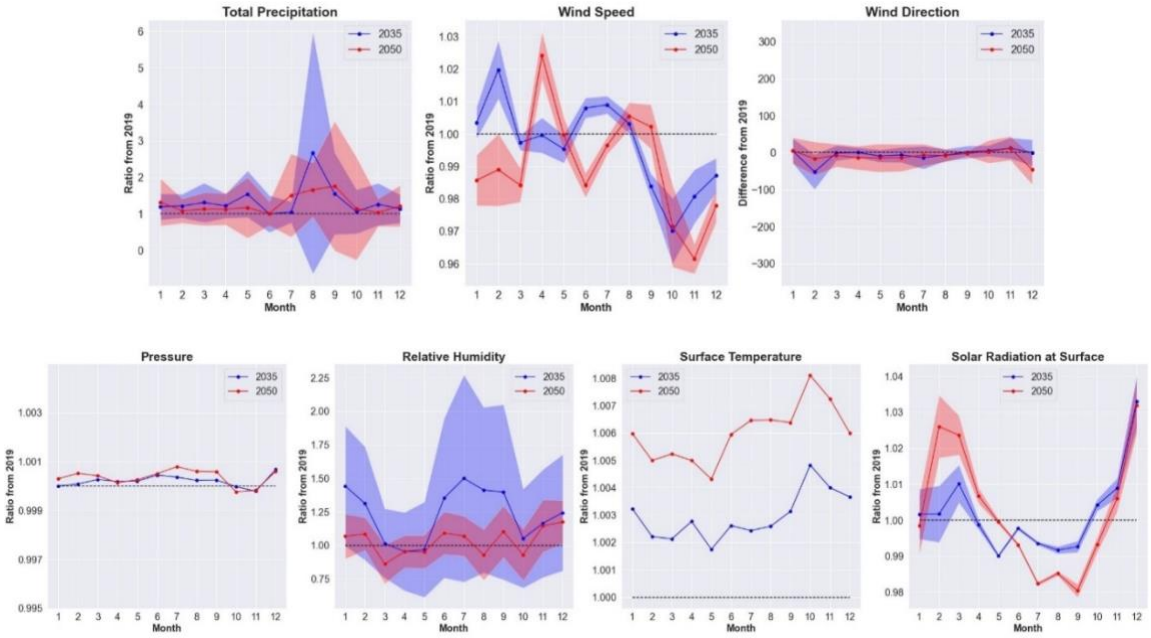

Fig. S9. Relative changes of meteorological conditions in 2035 and 2050 compared to 2019 based on the CMIP6 SSP585 future projection.

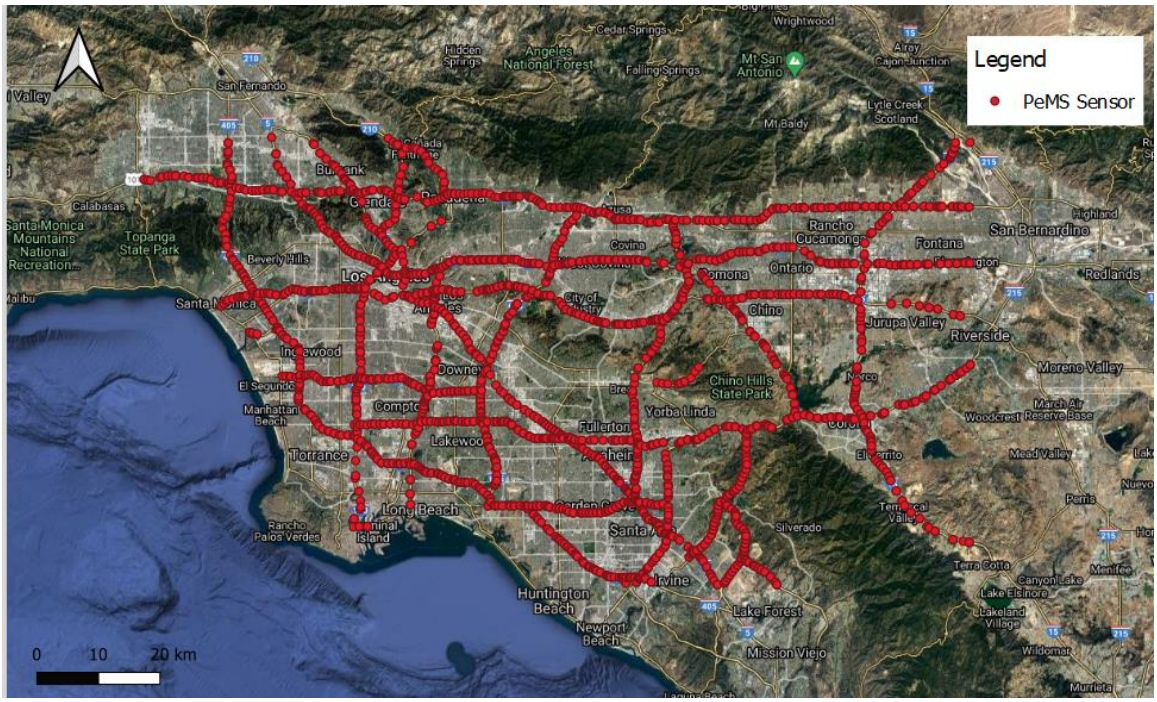

Fig. S10. The distribution of PeMS sensors located in LA Basin.

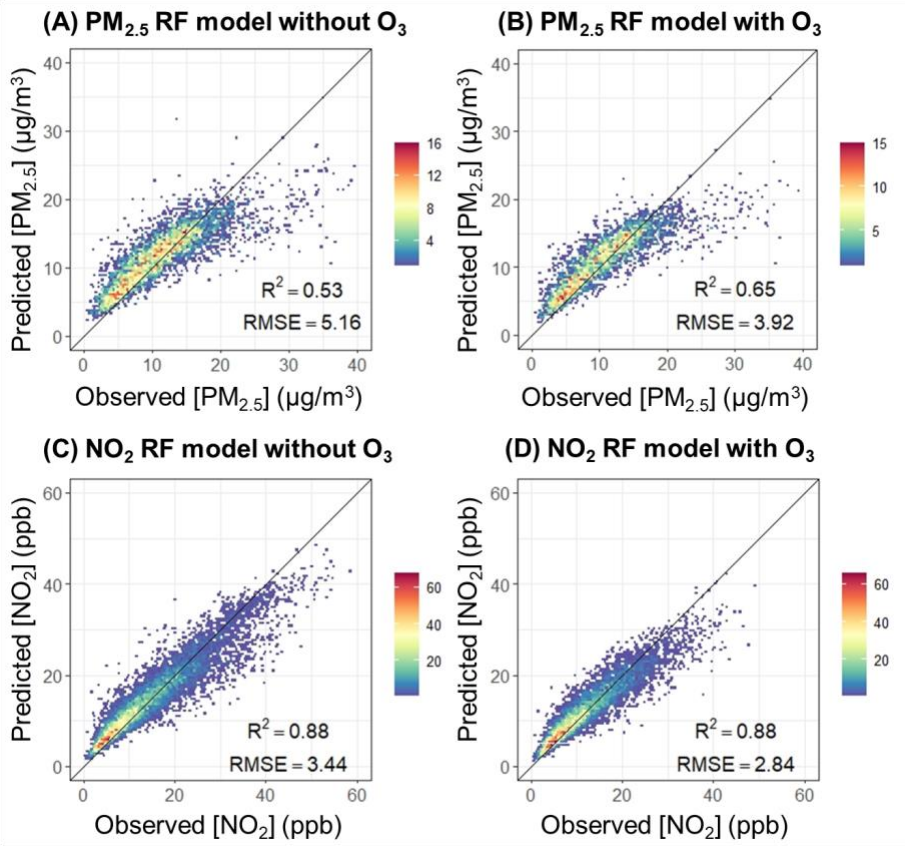

Fig. S11. Model performance of $PM_{2.5}$ and NO_2 RF models with and without O_3 as an indicator.

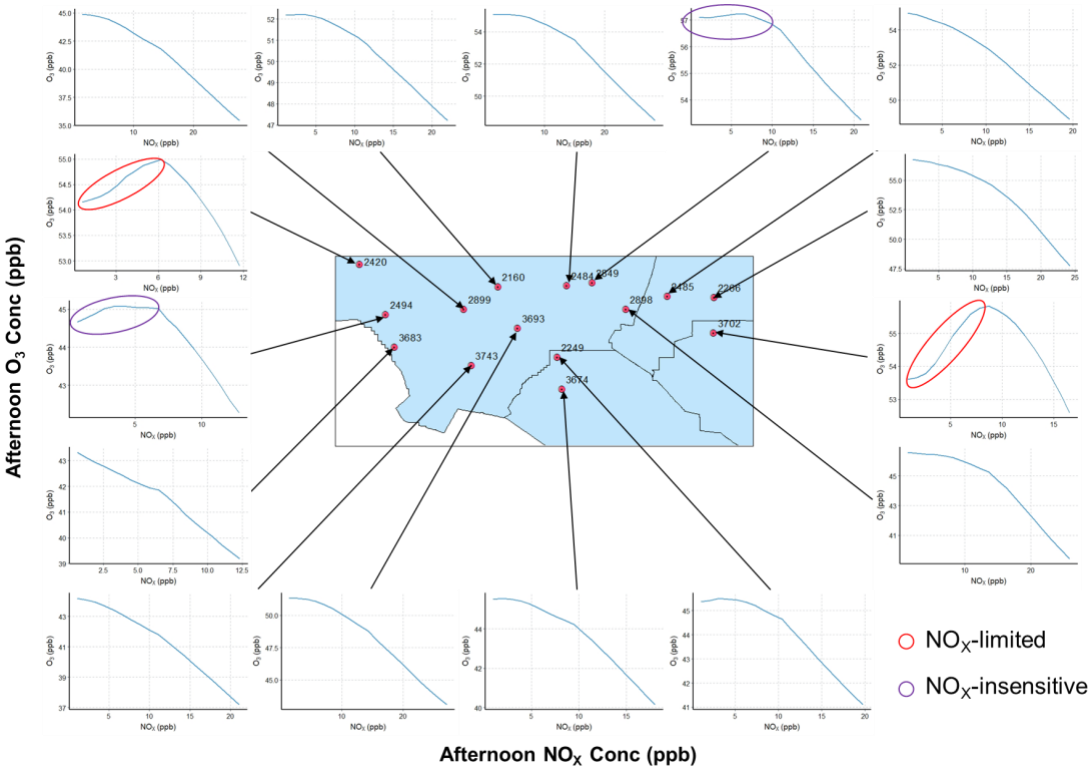

Fig. S12. Relationship between afternoon NO_x and O_3 in a new RF model in which O3 is a function of NOx and all other factors are fixed constant. The RF model is trained by the same observational dataset in 2019.

Table S1. Electrification ratios of total fleet mileage for truck and non-truck fleets in different future scenarios

Calendar Year	Fleet	Future traffic from EMFAC	Electrification 1	Electrification 2	Electrification 3
2035	Non-Trucks	4.41%	10%	20%	40%
2050	Non-Trucks	5.23%	20%	40%	80%
2035	Trucks	0.00%	5%	10%	20%
2050	Trucks	0.00%	10%	20%	40%

Table S2. Summary of predictors used to train RF models

Codes	Prediction variables	Units
Meteorological indicators		
Pressure	Atmospheric pressure from the nearest meteorological monitoring site	Pa
Wind direction	Wind direction from the nearest meteorological monitoring site	degree
Wind Speed	Wind speed from the nearest meteorological monitoring site	m/s
Temperature	Air temperature at 2 m from the nearest meteorological monitoring site	℃
Boundary layer height	Boundary layer height from the nearest meteorological monitoring site	m
Total precipitation	Total precipitation from the nearest meteorological monitoring site	m
Solar radiation	Solar radiation from the nearest meteorological monitoring site	
RH	Relative humidity	%
Traffic indicators		
Non-truck VMT*	Total Vehicle Mileage Traveled (VMT) ** of the non-truck fleet in buffers	miles
Truck VMT*	Total Vehicle Mileage Traveled (VMT) of the truck fleet in buffers	miles
Demographic indicators		
pop*	Population density in buffers	count/m^2
Distance to importance POIs		
D_CNG_Fueling_Stations	Distance to the nearest Compressed Natural Gas (CNG) fueling station	miles
D_Landfills	Distance to the nearest landfill	miles
D_LNG_Fueling_Stations	Distance to the nearest Liquefied Natural Gas (LNG) fueling station	miles
D_NG_Compressor_Stations	Distance to the nearest Natural Gas (NG) compressor station	miles
D_NG_Processing_Plants	Distance to the nearest Natural Gas (NG) processing plant	miles
D_NG_Storage_Fields	Distance to the nearest Natural Gas (NG) storage field	miles
D_Oil_Gas_Wells	Distance to the nearest oil gas well	miles
D_Petroleum_Refineries	Distance to the nearest petroleum refinery	miles
D_Power_Plants	Distance to the nearest power plant	miles
D_Wastewater_Treatment_ Plants	Distance to the nearest wastewater treatment plant	miles
D_airport	Distance to the nearest airport	miles
Temporal indicators		
DOW	Day of week (1-7)	n/a
holiday	1 for weekends and national holidays, otherwise is 0.	n/a

Note: * Buffer value variables (buffer radii 50 m, 100 m, 200 m, 300 m, 500 m, 1000 m, 2000 m, 3000 m, 4000 m, 5000 m)

** Vehicle Mileage Traveled (VMT) = $\sum \text{traffic volume} \times \text{road length}$

Table S3. Parameter setting of two prediction scenarios

Prediction scenarios	Meteorology indicators	Traffic indicators
Normal Traffic	Observed meteorology	Weekly average activity level based on Jan to Mar, 2020 (pre-COVID 19) for both non-truck and truck fleets
Normal Truck	Observed meteorology	Weekly average activity level based on Jan to Mar, 2020 (pre-COVID 19) for truck fleet + real activity level for non-truck fleet

Table S4. Primary features of the physical meteorological components participating models in this study

Modelling Group	ESM	Variable	Model Variant	Experiment ID	Nominal Resolution	Table ID	Frequent
CSIRO	ACCESS-ESM1-5	Surface Temperature (ts), Precipitation flux (pr), Surface Air Pressure (ps), Surface Downwelling Shortwave Flux in Air (rsds), Relative Humidity (hur), Wind Speed (sfcWind), Northward Wind (vas), Eastward Wind (uas)	r1i1p1f1, r2i1p1f1,r3i1p1f1,r4i1p1f1,r5i1p1f1	ssp585	250 km	Amon	Month
MPI-M	MPI-ESM1-2-LR		r1i1p1f1, r2i1p1f1,r3i1p1f1,r4i1p1f1,r5i1p1f1		250 km		
MOHC	HadGEM3-GC31-LL		r1i1p1f3, r2i1p1f3, r3i1p1f3, r4i1p1f3		250 km		
MOHC	UKESM1-0-LL		r1i1p1f2,r2i1p1f2,r3i1p1f2,r4i1p1f2,r8i1p1f2,		250 km		
NASA-GISS	GISS-E2-1-G		r1i1p3f1, r2i1p3f1, r3i1p3f1, r4i1p3f1, r5i1p3f1		250 km		
E3SM-Project	E3SM-1-1		r1i1p1f1		100 km		

SI References

1. Annual Monitoring Network Report | California Air Resources Board (December 22, 2020).

2. PeMS_Intro_User_Guide_v6.pdf (December 22, 2020).

3. V. Carranza, *et al.*, Vista-LA: Mapping methane-emitting infrastructure in the Los Angeles megacity. 24 (2018).

4. L. Breiman, Random Forests. *Machine Learning* **45**, 5–32 (2001).

5. P. Vizcaino, C. Lavalle, Development of European NO2 Land Use Regression Model for present and future exposure assessment: Implications for policy analysis. *Environmental Pollution* **240**, 140–154 (2018).

6. M. N. Wright, A. Ziegler, ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **77**, 1–17 (2017).

7. S. Nembrini, I. R. König, M. N. Wright, The revival of the Gini importance? *Bioinformatics* **34**, 3711–3718 (2018).

8. B. H. Menze, *et al.*, A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* **10**, 213 (2009).

9. Kendall's Advanced Theory of Statistics, Volume 1, Distribution Theory, 6th Edition | Wiley. *Wiley.com* (May 13, 2020).

10. R. C. Elandt-Johnson, N. L. Johnson, *Survival Models and Data Analysis* (John Wiley & Sons, 1980).

11. P. Samson, EMFAC2017 Volume III Technical Documentation V1.0.2 July 20, 2018. 253.

12. J. R. Horne, D. Dabdub, Impact of global climate change on ozone, particulate matter, and secondary organic aerosol concentrations in California: A model perturbation analysis. *Atmospheric Environment* **153**, 1–17 (2017).

13. Appendix D - Emissions Inventory Methods and Results for the Proposed Amendments. 20 (2020).

14. California's Advanced Clean Trucks regulation: Sales requirements for zero-emission heavy-duty trucks | International Council on Clean Transportation (December 21, 2020).

15. pLAn | L.A.'s Green New Deal | Sustainability pLAn 2019 (December 21, 2020).

16. Workshop Discussion Draft 2020 Mobile Source Strategy. 145 (2020).