Environment International 158 (2022) 106917

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

A machine learning model to estimate ground-level ozone concentrations in California using TROPOMI data and high-resolution meteorology

Wenhao Wang [a], Xiong Liu [b], Jianzhao Bi [c], Yang Liu [a,*]

[a] Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA, USA
[b] Harvard–Smithsonian Center for Astrophysics, Cambridge, MA, USA
[c] Department of Environmental & Occupational Health Sciences, School of Public Health, University of Washington, Seattle, WA, USA

ARTICLE INFO

Handling Editor: Thanh Nguyen

Keywords:
Surface ozone
TROPOMI
OMI
Ozone profile
HRRR
Random forest
Spatiotemporal distribution

ABSTRACT

Estimating ground-level ozone concentrations is crucial to study the adverse health effects of ozone exposure and better understand the impacts of ground-level ozone on biodiversity and vegetation. However, few studies have attempted to use satellite retrieved ozone as an indicator given their low sensitivity in the boundary layer. Using the Troposphere Monitoring Instrument (TROPOMI)'s total ozone column together with the ozone profile information retrieved by the Ozone Monitoring Instrument (OMI), as TROPOMI ozone profile product has not been released, we developed a machine learning model to estimate daily maximum 8-hour average ground-level ozone concentration at 10 km spatial resolution in California. In addition to satellite parameters, we included meteorological fields from the High-Resolution Rapid Refresh (HRRR) system at 3 km resolution and land-use information as predictors. Our model achieved an overall 10-fold cross-validation (CV) R^2 of 0.84 with root mean square error (RMSE) of 0.0059 ppm, indicating a good agreement between model predictions and observations. Model predictions showed that the suburb of Los Angeles Metropolitan area had the highest ozone levels, while the Bay Area and the Pacific coast had the lowest. High ozone levels are also seen in Southern California and along the east side of the Central Valley. TROPOMI data improved the estimate of extreme values when compared to a similar model without it. Our study demonstrates the feasibility and value of using TROPOMI data in the spatiotemporal characterization of ground-level ozone concentration.

1. Introduction

Ground-level ozone is a secondary air pollutant produced by photochemical reactions involving nitrogen oxides (NO_x) and volatile organic compounds (VOCs) (Sicard et al., 2013). Many epidemiological studies worldwide have associated exposure to ground-level ozone with adverse respiratory and cardiovascular morbidity and mortality (Day et al., 2017; Nuvolone et al., 2018; Tian et al., 2020). It was estimated that globally long-term ozone exposure caused additional 254,000 deaths from chronic obstructive pulmonary disease (Cohen et al., 2017). Ground-level ozone pollution is also projected to increase in a warming climate, further exacerbating its public health burden (Orru et al., 2017; Stowell et al., 2017). To date, most ozone health effects research has relied on ground measurements from central monitors for exposure assessment. However, the limited spatial coverage of the ground monitoring networks may not fully characterize the spatial contrast of ground-level ozone concentration (Huang et al., 2019).

Given its complex formation mechanism, atmospheric chemical transport models (CTM) were often used to simulate the spatiotemporal patterns of ground-level ozone (Bey et al., 2001; Emmons et al., 2010; Hu et al., 2016). For example, Liu et al. applied Community Multiscale Air Quality Model (CMAQ) to evaluate daily maximum 1-hour average surface ozone concentrations during specific air pollution episodes in China (X.-H. Liu et al., 2010). However, the spatial resolution of the model is 36 km, insufficient in understanding the fine-scale spatial variations of ground-level ozone concentrations. Since CTMs rely on emissions inventories and consume large computing resources, spatial resolutions of these models are often coarse (Di et al., 2017). Sicard et al. used the Weather Research and Forecasting model with Chemistry (WRF-Chem) model to estimate the spatial and seasonal variations of ground-level ozone and other air pollutants in east Asia at a finer spatial resolution of 8 km (Sicard et al., 2021). However, the correlation

* Corresponding author at: Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA.
E-mail address: yang.liu@emory.edu (Y. Liu).

https://doi.org/10.1016/j.envint.2021.106917
Received 29 June 2021; Received in revised form 30 September 2021; Accepted 1 October 2021
Available online 5 October 2021

coefficient of the model predictions with ground observations is 0.51, suggesting that 75% of the variability in ground-level ozone concentrations was not explained. For the past two decades, due to its wide spatial coverage and long data records, satellite remote sensing has been extensively used to better characterize various air pollutants as well as to assess the adverse health outcomes associated with air pollution exposure (Diao et al., 2019; Fioletov et al., 2013; Larkin et al., 2017; Liang et al., 2020; Shi et al., 2016). However, studies that incorporated satellite-based ozone products to estimate ground-level ozone concentrations are very limited. Liang et al. retrieved ground ozone concentrations from Ozone Measurement Instrument (OMI) Level 2 Ozone Profile product, fusing ground monitors to estimate ambient ozone exposure in China (Liang et al., 2019). Although ground measurements were used to calibrate satellite retrievals, model performance in some regions still needs improvement. Di et al. integrated OMI OMO3PR vertical profiles with convolutional layers of ground monitors in a neural network to predict daily ozone concentration in the Contiguous United States (Di et al., 2017). Requia et al. combined seven OMI data products and three machine learning models to predict daily ozone concentrations from 2000 to 2016 in the contiguous United States (Requia et al., 2020). While these two studies achieve high performance with R^2 values of 0.76 and 0.91, respectively, the massive input data, fusion of ground measurements in the prediction, and the multi-model ensemble make result interpretation a challenge. Given the complexities of their models, it is also difficult to evaluate the utility of satellite-retrieved ozone information.

The Tropospheric Monitoring Instrument (TROPOMI), launched in October 2017 aboard the European Space Agency's Sentinel-5P, measures the reflected solar radiation in the ultraviolet and visible (270–500 nm), near-infrared (675–775 nm) and shortwave infrared (2305–2385 nm) range. Compared to previous launched Ozone Monitoring Instrument (OMI), TROPOMI requirements are much higher in several aspects (Veefkind et al., 2012). TROPOMI can provide abundance information of a wide range of pollutant gases including NO_2, ozone, formaldehyde, SO_2 and CO at a spatial resolution as high as 3.5 km × 7.0 km (increased to 3.5 km × 5.5 km since August 6, 2019) and a swath width of 2600 km. TROPOMI's extended near-infrared band provides better cloud correction in trace gas retrievals. Several studies have indicated the feasibility of TROPOMI in measuring ambient air pollution. For example, Goldberg et al. (2019) applied TROPOMI tropospheric NO_2 columns to estimate NO_x emission from three North America cities during May to September 2018. Zhao et al. retrieved the tropospheric ozone column using TROPOMI ultraviolet radiances, and assessed the variation of tropospheric ozone concentrations during the COVID-19 pandemic in China (Zhao et al., 2021). However, no validation against ground ozone measurements was reported in their study.

In this study, we developed a random forest model incorporating TROPOMI tropospheric NO_2 column and TROPOMI total ozone column data to estimate daily maximum 8-hour average ground-level ozone concentration in California at 10 km spatial resolution. We used OMI ozone profile data to estimate boundary layer ozone column from the TROPOMI total ozone column. We also included high-resolution meteorology to address the high sensitivity of ozone reaction to meteorological conditions. In addition, we conducted sensitivity analyses and comparisons to assess the advantages of including TROPOMI data and high-resolution meteorology.

2. Data and methods

2.1. Study design

Our study domain is the state of California, an area of ~423,900 km^2 and home to nearly 40 million residents (Fig. 1). Our one-year study period starts from May 1st, 2018, the first available day for TROPOMI data products, to April 30th, 2019. We designed a modeling grid in 10 × 10 km^2 with a total of 4207 grid cells to integrate all modeling parameters in our study domain. The overall study workflow is presented in Fig. S1.

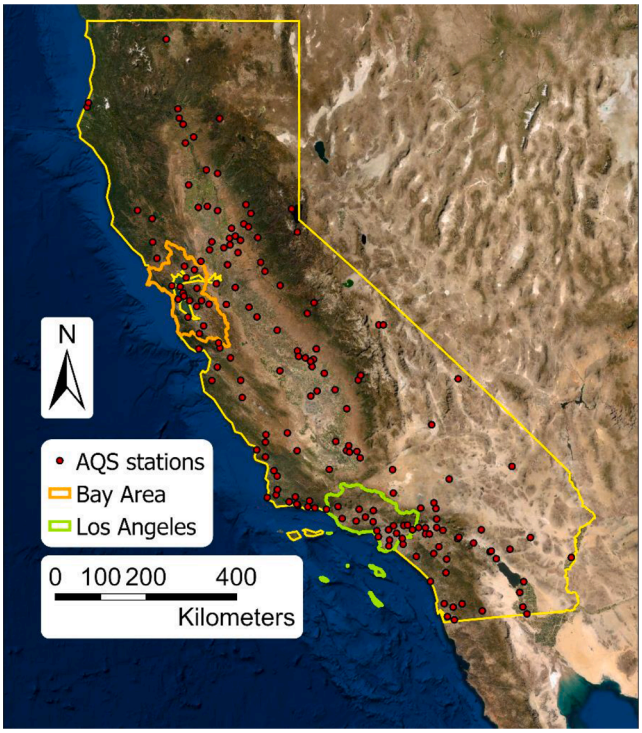

Fig. 1. The study domain with the locations of AQS stations.

2.2. Ground ozone data

Hourly ground ozone measurements were acquired from EPA's Air Quality System (https://www.epa.gov/aqs). There are 177 stations providing measurements in California during our study period as shown in Fig. 1. We calculated the daily maximum 8-hour average ozone concentrations (MDA8) on all station-days with more than 16 h of continuous measurements.

2.3. Remote sensing data

We acquired TROPOMI level-2 total ozone column and TROPOMI level-2 NO_2 tropospheric column to represent this ozone precursor (Garane et al., 2019). TROPOMI data was downloaded from NASA's Earth Science Data System (https://earthdata.nasa.gov/). We then resampled the TROPOMI satellite data products from their original spatial resolution of 3.5 km × 7.0 km to the designed 10 km grid by averaging the daily values of ozone and NO_2 retrievals from the TROPOMI pixels whose centroids fall in a given grid cell.

Because TROPOMI ozone profile product is not currently available, we introduced OMI ozone profile product OMPROFOZ developed at the Smithsonian Astrophysical Observatory to estimate the boundary layer fraction of the TROPOMI total ozone column. We downloaded the OMPROFOZ from NASA Aura Validation Data Center (AVDC) (https://avdc.gsfc.nasa.gov/pub/data/satellite/Aura/OMI/V03/L2/OMPROFOZ/) and gridded the data to 0.5° x 0.5° spatial resolution. OMPROFOZ ozone profiles include 24 vertical layers (~2.5 km thick per layer) from the surface to ~65 km above the ground. The retrieval errors due to instrument random-noise and smoothing errors are estimated at 6–35% in the troposphere (X. Liu et al., 2010). The ozone vertical profiles were assigned to each modeling grid cell using inverse distance weighting. Finally, we derived the satellite-based boundary layer ozone column by multiplying the surface layer fraction (i.e., from surface to about 2.5 km above surface) of OMI ozone column by the TROPOMI total ozone

column. The combination of TROPOMI total ozone column and OMPROFOZ ozone profile takes advantage of TROPOMI's higher spatial resolution and retrieval accuracy and OMI's ozone vertical profile information (Veefkind et al., 2012).

2.4. Meteorological fields

The High-Resolution Rapid Refresh (HRRR) is a cloud-resolving and convection-allowing atmospheric model at 3 km resolution from NOAA's Earth System Research Laboratory (https://rapidrefresh.noaa.gov/hrrr/). Previous validation studies suggested that HRRR can accurately predict near-surface temperature and precipitation (Lee et al., 2019; Yue & Gebremichael, 2020). We extracted 19 meteorological parameters from HRRR, including u-wind (an east-west component of wind vector) and v-wind (a north-south component of the wind vector) at 10 m above the ground and 250 hPa pressure level (m/s), surface air temperature (K), pressure (Pa), moisture availability (%), relative humidity (%), surface roughness (m), ground heat flux (W/m^2), convective available potential energy (J/kg), medium cloud cover (%), visible diffuse downward solar flux (W/m^2), incoming short-wave radiation (W/m^2), maximum upward velocity from 1000 to 400 hPa (m/s), and maximum downward velocity from 1000 to 400 hPa (m/s). Hourly data between 10 a.m. to 4 p.m. local standard time were averaged to represent the average weather condition at the Sentinel-5P satellite overpass time of 1:30 p.m. local solar time. We then aggregated the daily value of those meteorological parameters by taking the average of all 3 km HRRR pixels falling into a given 10 km model grid cell. Correlation tests among all meteorological parameters were performed to ensure that none of them were strongly correlated with a Pearson correlation coefficient greater than 0.5 (Hu et al., 2017) (Table S1).

2.5. Land-use variables

We obtained elevation information from the National Elevation Dataset (NED, http://ned.usgs.gov) at 30 m resolution. Land-use parameters including area fraction of water, forest, developed, and barren land were derived from 2016 National Land Cover Database (NLCD) (http://www.mrlc.gov) at 30 m resolution. Elevation and the land use parameters were resampled to the 10 km grid cells. Finally, we calculated the distance of each model grid cell centroid to the Pacific coastline as an indicator of the impact of sea breezes on air pollution dispersion (Stowell et al., 2020). GIS operations were done using ArcGIS Pro with the USA Contiguous Albers Equal Area Conic projection coordination system.

2.6. Model development

The random forest machine learning algorithm is a set of decision trees that can capture non-linear relationships between variables, requiring no assumptions on independence or probability distributions. Random forest averages the predictions from individual decision trees generated based on randomly selected subsets of predictors and samples. There are two primary hyperparameters: m_{try} which is the number of predictors sampled for splitting at each node, and n_{tree} which is the number of trees in the forest. We trained our model by each subset of the combination of the m_{try} and n_{tree} to get the model with best prediction accuracy (Breiman, 2001). Many previous studies have applied the random forest model in predicting air pollution levels (Hu et al., 2017; Zhan et al., 2018). These studies presented robust performance, and the random forest model provides predictor importance rankings to help with result interpretation which is not available from neural network models. Previous research also suggested that tree-based ensembles machine learning model perform well at predicting ground-level ozone in California during wildfire events (Watson et al., 2019). The importance ranking of predictors was assessed according to the out-of-bag samples as results of each predictor variable being permuted which is expressed as an increase of mean square errors (IncMSE%) of predictions (Altmann et al., 2010). We developed our random forest model with the MDA8 ozone concentration as the dependent variable and 27 predictors including remote sensing data, land-use variables, and meteorological fields. To capture the diverse relationships among predictors and ozone concentration in different seasons across California, we include all available stations-days in the training dataset.

We adopted three cross-validation strategies to evaluate the model performance. First, we randomly split the training set into 10 subsets, each of which contains about 10% of the training data. We used 9 subsets to do the model training, then set the remaining subset as the testing data to compare with model predictions. This process was repeated 10 times with 10 different choices of testing dataset. Similarly, we conducted a temporal CV by partitioning the training data set by day of year. These 10-fold CV methods are widely applied in previous machine learning models (Di et al., 2017; Requia et al., 2020). To assess how model performance varies in space, we also conducted leave-one-location-out CV (LOLO CV). For each fold of the LOLO CV, the model was trained on a subset of data from all but one ground monitor, while the prediction errors were calculated on the excluded ground monitor (Watson et al., 2019). The process was repeated on all ground monitors. The LOLO CV ensures the absence of training observations in validation sets and appears more accurately estimates prediction error than 10-fold CV. The CV R^2values and root mean squared errors (RMSE) were used to assess the prediction accuracy in all three CV settings. All the modeling processes and data analysis were done using the *ranger* package in R software (version 3.6.0), while the map visualizations were produced using ArcGIS Pro (version 2.7).

3. Results

3.1. Descriptive statistics and model performance

The descriptive statistics for all prediction parameters are presented in Table S1. Our training dataset included 34,336 daily-level samples while the prediction datasets had 858,386 incidences after removing all missing values (out of a total of 1,495,654 possible pixel-days during our study period). The predicted annual mean (standard deviation) MDA8 ozone concentration was 0.047 (0.0052) ppm as compared to 0.044 ppm calculated from all ground monitors.

Fig. 2 shows the random, temporal, and LOLO CV results with R^2, RMSE as well as the linear slope between predictions and observations. The overall random CV had a R^2 value of 0.84 and a RMSE of 0.0059 ppm, while the temporal CV had a R^2 value of 0.80 and a RMSE of 0.0067 ppm. In terms of the LOLO CV, we had a lower R^2 of 0.73 and a larger RMSE of 0.0077 ppm.

3.2. Predictor importance ranking

Fig. 3 illustrates the importance ranking of all predictors. TROPOMI tropospheric NO_2 column and the satellite derived boundary layer ozone column were among top three most important predictors. Among the meteorological variables, two vertical mixing indicators, i.e., maximum downward velocity and maximum upward velocity ranked the highest. Other surface level meteorological parameters, such as surface temperature, relative humidity, and planetary boundary layer height, were also important in predicting ground ozone concentrations. Distance to coast was the most significant spatial predictor although all NLCD land use variables and elevation had relatively high importance.

3.3. Model estimated spatial and temporal trends

Fig. 4 presents the time trend of weekly mean predicted and observed MDA8 ozone concentrations. As expected, ozone levels in California exhibited a clear season pattern with high concentrations in the summer (up to 0.108 ppm) and low concentrations in winter (0.006 ppm). When

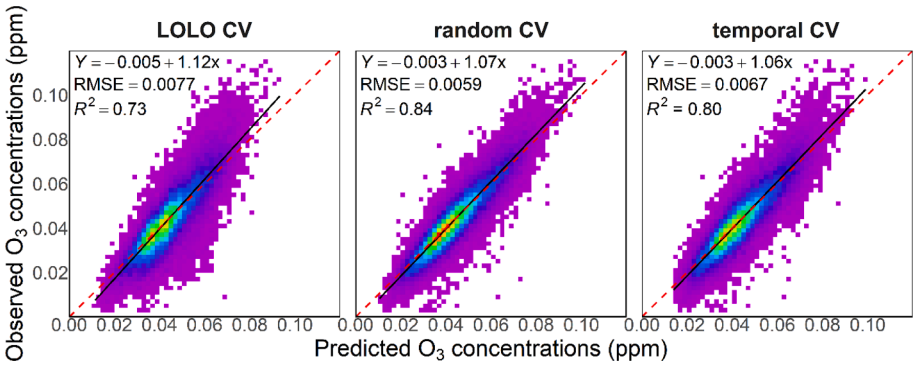

Fig. 2. CV results.

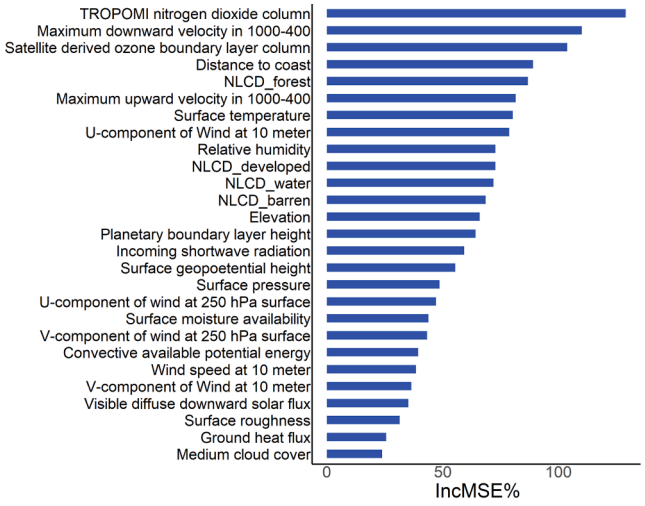

Fig. 3. Importance ranking of variables according to mean square errors (MSE).

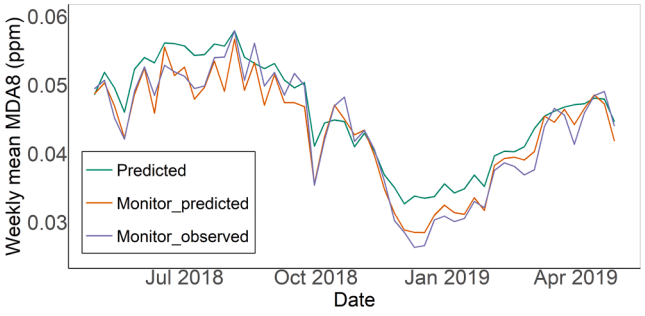

Fig. 4. Time-series plots of state-wide model estimated MDA8 ozone concentration, MDA8 ozone concentration averaged over pixels coincided with ground monitors, and state-wide averaged ground observations.

using predictions from model grid cells coincided with ground monitors, the predicted ozone temporal patterns agreed with ground observations well. On the other hand, the state-wide weekly mean MDA8 values (green line in Fig. 4) were significantly higher in January and February of 2019 (0.035 ppm) than ground observations (0.031 ppm). One potential explanation is that the AQS stations, mostly located in populated areas with high NOx emissions, do not represent ozone levels in the entire state. With the low photochemical activity in winter, the relatively high NO_x emission will lead to ozone titration, which reduces ozone concentration in these areas with monitors (Li et al., 2021). In addition, we compared the predicted ozone pattern in January and February of 2019 and HRRR-estimated snow coverage in the period. We found that the high ozone concentrations in the Sierra Nevada forests area are consistent with the high snow coverage in the area (Fig. S2). We also observed an inconsistency in July and Summer of 2019. During summer months, the ozone concentrations in desert areas in southern California is high where the AQS stations are not able to capture these peaks.

Fig. 5 shows the spatial distribution of the predicted annual mean MDA8 ozone levels as well as ozone levels during the ozone season (May to October) and the non-ozone season (rest of the year) (Malig et al., 2016). High ozone levels were observed in Southern California and the east side of the Central Valley through entire study period. The inland suburb of Greater Los Angeles Area saw the highest concentration, where the industrial sources emit ozone precursors CH4 and VOCs into the air (Jeričević et al., 2014). We also observed higher ozone concentrations in the desert areas. The Bay area had the lowest concentration in California, and ozone levels were also significantly lower in urbanized areas along the coastline including low elevation areas around Oakland. Model prediction shows higher ozone levels in areas such as east suburbs of Los Angeles downwind of emissions sources of ozone precursors due to transport of ozone as well as 'reservoir' species such as peroxyacetyl nitrate (PAN) (Monks et al., 2015; Sicard et al., 2016). Although ozone levels followed similar spatial patterns across different periods, MDA8 ozone concentrations can be 50% greater in the ozone season than the non-ozone season. The spatial contrast of ozone levels was also more dramatic in the ozone season. For example, the highest MDA8 ozone level during the ozone season was above 0.070 ppm in San Bernardino, east to the Los Angeles Metropolitan while the lowest value was 0.013 ppm in north of the Bay Area.

3.4. Sensitivity analysis

We compared the LOLO CV results of different model configurations in order to assess the importance of various predictors (Table 1). Compared to the REF model (full model with all predictors included), the M2 model with HRRR meteorological parameters excluded had the lowest R^2 (0.417) and highest RMSE (0.0112 ppm), indicating that meteorological conditions are crucial to the spatiotemporal variability of ground-level ozone. M1 (satellite parameters excluded) and M3 (land use parameters excluded) have similar CV R^2 and RMSE values to the REF model. However, a closer examination indicated that the reference model predicted more extreme values of MDA8 ozone concentrations with the help of the satellite-based predictors. For example, Fig. 6 showed that in August 2018, compared to the M1, the change of predicted monthly mean MDA8 between REF and M1 (i.e., REF – M1) ranged −6.9% in forests regions such as the Mendocino and Six Rivers National Forests and Pfeiffer Big Sur State Park to 13.3% around Los Angeles and Riverside.

To further access the contribution of satellite data, we evaluated the REF and M1 models at the top 20% of all station-day observations and their corresponding training dataset. The LOLO CV R^2 and RMSE of the

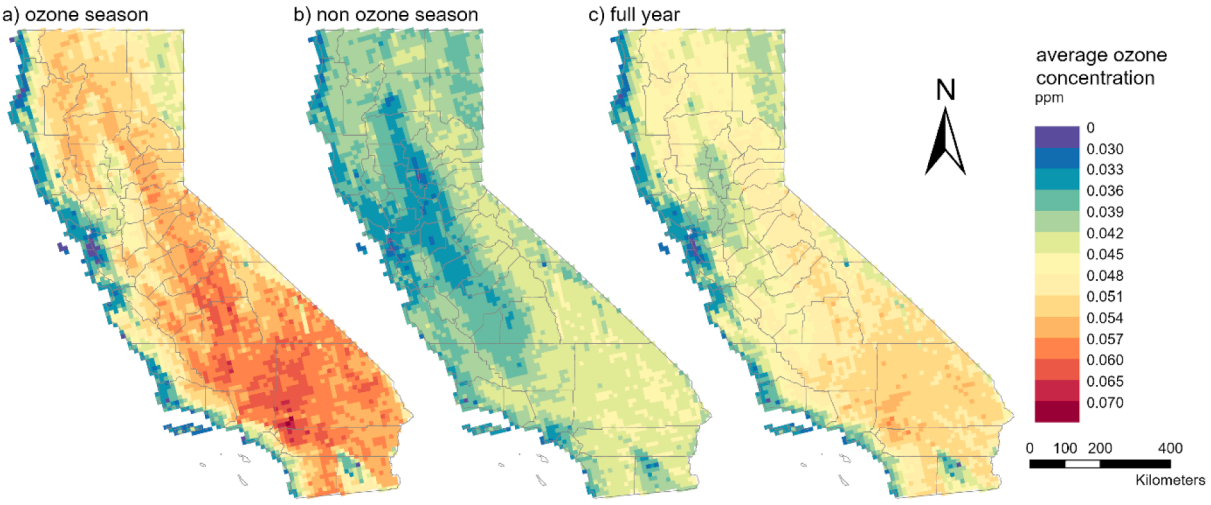

Fig. 5. Mean prediction maps of daily maximum 8-hour mean ground-level ozone concentrations in California for different periods: a) 5/1/2018 to 10/31/2018; b) 11/1/2018 to 4/30/2019; c) 5/1/2018 to 4/30/2019.

Table 1
R^2 and RMSE values of sensitivity test result for different model settings.

Case	Settings	CV R^2	RMSE (ppm)
REF	Satellite + Meteorology + Land Use	0.726	0.0077
M1	No Satellite	0.696	0.0081
M2	No Meteorology	0.417	0.0112
M3	No Land use	0.718	0.0078

REF high concentration model are 0.444 and 0.0834 ppm, respectively, while the LOLO CV R^2 and RMSE of the M1 high concentration model are 0.377 and 0.0858 ppm, respectively (Fig. S3).

4. Discussion

In a previous study in California using spatiotemporal statistical models, Bogaert et al. performed Bayesian Maximum Entropy analysis corresponding to ground measurements to achieve monthly average MDA8 predictions in a spatial resolution of 20 km (Bogaert et al., 2009). In comparison, our model has finer spatial and temporal resolutions and better performance. An important reason is that we included TROPOMI and OMI satellite data to characterize ozone and NO_2 abundance in our machine learning model. Compared to statistical models, machine learning algorithms are more capable of handling the complex and nonlinear interactions among precursors such as NOx and volatile organic compounds and meteorological conditions. When compared with a previously reported national scale hybrid ozone model (Di et al., 2017) and an ensemble machine learning model (Requia et al., 2020), our model's performance is commensurate to their CV R^2 and RMSE

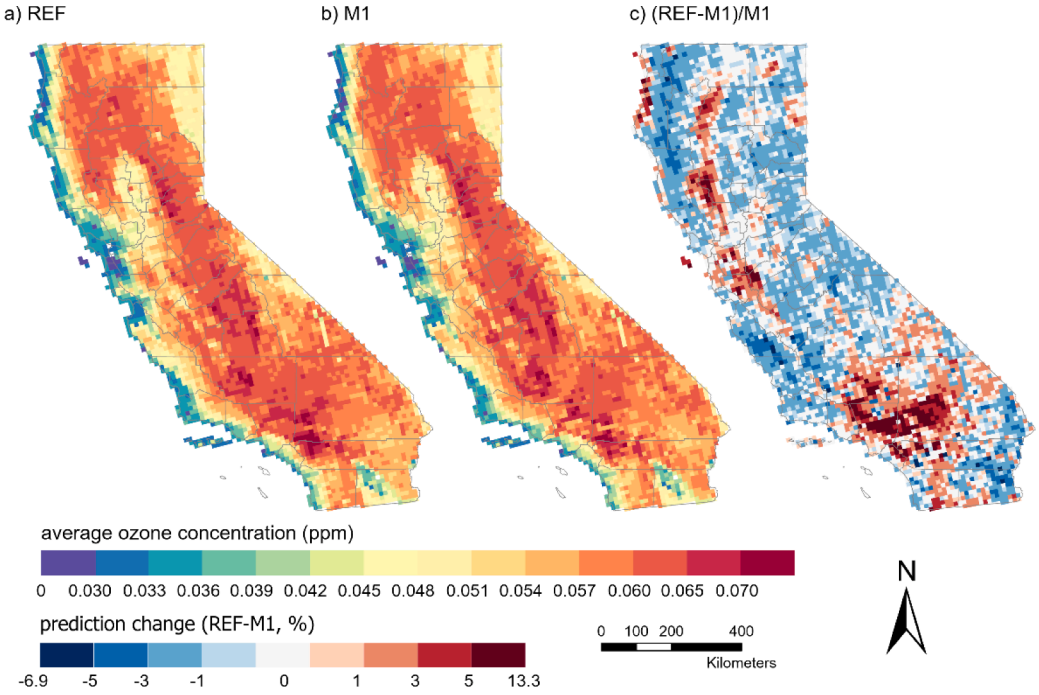

Fig. 6. Difference in prediction of mean daily maximum 8-hour average concentrations in August 2018. a) The monthly predicting value under model REF, b) The monthly predicting value under model M1, c) Change of average MD8 ozone concentration in percentage from M1 to REF ((REF-M1)/M1).

estimated in the west coast. While these models and our model used similar groups of predictors including satellite derived data, land-use terms, and meteorological data, our model does not include spatially interpolated ground observations as a predictor to allow potential generalization of our model in regions with more sparse ground monitors. Unlike the complex and multiple-stage modeling strategies adopted by previous studies, our random forest model can provide an importance ranking of model predictors and allowed us to assess the contribution of satellite data products.

We compared our model predicted spatial distribution of ground-level ozone with a previous statistical model in California (Bogaert et al., 2009). Both studies predict high ground ozone in Southern California with the highest level in the eastern suburb of Los Angeles, likely due to large traffic emissions of precursor gases as well as a warm weather and strong solar radiation in this region (Levy II et al., 1997; Niu et al., 2018). The higher ozone level in the suburb as compared to the neighboring urban center was also reported by a previous study (Zhao et al., 2018). The relative high ozone concentrations in the Central Valley was likely due to reactive organic gas emissions from livestock feed (Howard et al., 2010). The low ozone concentrations in coastal California predicted by our model are consistent with previous studies (Bogaert et al., 2009; Requia et al., 2020), and could be because there is no or little ozone deposition on water.

In terms of predictor importance, TROPOMI level-2 NO_2 column is the most important predictor in our model, and this reflects the importance of NOx as an ozone precursor. Several previous studies reported that TROPOMI NO_2 retrievals can provide highly accurate estimation of boundary layer NO_2 concentrations at a fine spatial resolution (Goldberg et al., 2019; Griffin et al., 2019; Ialongo et al., 2020). The slightly lower ranking of satellite-derived boundary layer ozone column could be due to the greater uncertainty in estimating the boundary layer ozone fraction of the TROPOMI ozone column using the OMI ozone profile. Without TROPOMI ozone profiles, the combination of TROPOMI and OMI data products is a compromise. However, it still helped to strengthen the correlation of satellite-retrieved ozone column with ground measurements. It is interesting to note that the vertical mixing indicators including maximum upward and downward velocity present significant importance in our model. One potential explanation is the inclusion of vertical mixing indicators modified the lack of sensitivities in the OMI ozone profile retrievals. Another interesting finding is that land use variables are somewhat more important than other predictors and forest cover is the most important land use variable. Previous research suggested that forest cover represents the natural source of VOCs which is an important precursor of ground level ozone (Hu et al., 2018). This variable may also indirectly reflect the potential of wildland fires that contributed to sudden changes in ozone concentration in the western United States (Jaffe et al., 2013).

When compared with a reduced form model without satellite data, our full model predicted higher ozone concentrations in Los Angeles in a somewhat smoother spatial distribution in August 2018 (Fig. 6). In terms of model performance at high concentrations levels, the inclusion of satellite data predictors improved model LOLO CV R^2 from 0.377 to 0.444, while the LOLO CV R^2 of the REF and M1 models on full training dataset are 0.726 and 0.696. The greater improvement at the high MDA8 values indicates the satellite data may increase the accuracy in predicting high ozone concentrations which is consistent with our analysis on the spatial distribution of predicted MDA8 ozone concentrations. Our sensitivity analysis also revealed the large impact of the HRRR meteorological data on model performance, i.e., increasing the LOLO CV R^2 value from 0.42 (model M2) to 0.73 (model REF). Given the strong dependence of ozone formation on weather conditions, our finding is not surprising (Liu et al., 2020). It suggests the current satellite data alone is insufficient to fully capture the fine-scale spatiotemporal trend of surface ozone concentrations. To our knowledge, this is the first time HRRR data was applied in statistical models to estimate ground ozone distributions. Further research is needed to investigate how high-resolution meteorological data from models such as HRRR may benefit quantitative air pollution exposure assessment. Our study has a few limitations. First, the spatial resolution of our model can be improved. For our study period, the spatial resolution of TROPOMI ozone and NO_2 products was 3.5×7.0 km^2, which has been improved to 3.5×5.5 km^2 since August 2019. Given the 3 km resolution of HRRR meteorology and higher resolution land use data, future models covering time periods after August 2019 should be able to achieve a 5 km or even higher spatial resolution. Second, we resampled the OMI profile data from the original spatial resolution to the prediction grid by inverse distance weighted method, while the OMI ozone profile was retrieved based on OMI's observations with missing data. The missing OMI data also accounted for the incomplete training dataset. Thus, scaling TROPOMI total column to its boundary layer fraction using OMI ozone vertical profile inevitably introduced uncertainty, which could be the reason for this parameter not be the most important predictor of surface ozone concentrations. As TROPOMI tropospheric ozone column or ozone vertical profile product becomes available, the performance of our random forest model may be further improved. Third, the resampling methods of integrating different datasets with various original spatial resolutions to the designed 10 km spatial resolution may bring uncertainty. Since many observations were analyzed as the value at pixel centroids, the distance-based resampling methods may not accurately calculate those values in the prediction grid cells. Finally, more research is needed to develop gap-filling techniques to eliminate the data gap left by missing TROPOMI retrievals following the examples of aerosol optical depth gap filling (Bi et al., 2019; Xiao et al., 2017).

5. Conclusion

In this study, we developed a machine learning model to incorporate TROPOMI level-2 satellite data with high resolution meteorology data to predict ground level ozone concentrations in California. This model predicted accurate ground level ozone concentrations in California with explainable spatial temporal distributions. Compared to previous ground level ozone machine learning models, our model can assess the contribution of satellite data products in a concise modeling framework. The finding from this study suggests that TROPOMI data improved the estimate of extreme values in ground level ozone modeling. It could also accelerate future research on applying satellite data product and high-resolution meteorological data to predict ground level ozone concentrations.

CRediT authorship contribution statement

Wenhao Wang: Methodology, Software, Formal analysis, Data curation, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Xiong Liu:** Resources, Writing – review & editing. **Jianzhao Bi:** Resources, Writing – review & editing. **Yang Liu:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The work of W. Wang and Y. Liu was supported by the NASA Applied Sciences Program (Grant # 80NSSC21K0507 and 80NSSC19K0191). The work of X. Liu was supported by the NASA Aura science team program (Grant # NNX17AI82G). The content is solely the responsibility of the authors and does not necessarily represent the official views of NASA.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2021.106917.

References

Altmann, A., Toloşi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26 (10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134.

Bey, I., Jacob, D.J., Yantosca, R.M., Logan, J.A., Field, B.D., Fiore, A.M., Li, Q., Liu, H.Y., Mickley, L.J., Schultz, M.G., 2001. Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. J. Geophys. Res. [Atmos.] 106 (D19), 23073–23095. https://doi.org/10.1029/2001JD000807.

Bi, J., Belle, J.H., Wang, Y., Lyapustin, A.I., Wildani, A., Liu, Y., 2019. Impacts of snow and cloud covers on satellite-derived PM2.5 levels. Remote Sens. Environ. 221, 665–674. https://doi.org/10.1016/j.rse.2018.12.002.

Bogaert, P., Christakos, G., Jerrett, M., Yu, H.L., 2009. Spatiotemporal modelling of ozone distribution in the State of California. Atmos. Environ. 43 (15), 2471–2480. https://doi.org/10.1016/j.atmosenv.2009.01.049.

Breiman, L., 2001. Random forests. *Machine Learning* 45 (1), 5–32.

Cohen, A.J., Brauer, M., Burnett, R., Anderson, H.R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., Pope, C.A., Shin, H., Straif, K., Shaddick, G., Thomas, M., van Dingenen, R., van Donkelaar, A., Vos, T., Murray, C.J.L., Forouzanfar, M.H., 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. The Lancet 389 (10082), 1907–1918. https://doi.org/10.1016/S0140-6736(17)30505-6.

Day, D.B., Xiang, J., Mo, J., Li, F., Chung, M., Gong, J., Weschler, C.J., Ohman-Strickland, P.A., Sundell, J., Weng, W., Zhang, Y., Zhang, J.J., 2017. Association of Ozone Exposure With Cardiorespiratory Pathophysiologic Mechanisms in Healthy Adults. JAMA Int. Med. 177 (9), 1344–1353. https://doi.org/10.1001/jamainternmed.2017.2842.

Di, Q., Rowland, S., Koutrakis, P., Schwartz, J., 2017. A hybrid model for spatially and temporally resolved ozone exposures in the continental United States. J. Air Waste Manag. Assoc. 67 (1), 39–52. https://doi.org/10.1080/10962247.2016.1200159.

Diao, M., Holloway, T., Choi, S., O'Neill, S.M., Al-Hamdan, M.Z., Van Donkelaar, A., Martin, R.V., Jin, X., Fiore, A.M., Henze, D.K., Lacey, F., Kinney, P.L., Freedman, F., Larkin, N.K., Zou, Y., Kelly, J.T., Vaidyanathan, A., 2019. Methods, availability, and applications of PM2.5 exposure estimates derived from ground measurements, satellite, and atmospheric models. J. Air Waste Manag. Assoc. 69 (12), 1391–1414. https://doi.org/10.1080/10962247.2019.1668498.

Emmons, L.K., Walters, S., Hess, P.G., Lamarque, J.F., Pfister, G.G., Fillmore, D., Granier, C., Guenther, A., Kinnison, D., Laepple, T., Orlando, J., Tie, X., Tyndall, G., Wiedinmyer, C., Baughcum, S.L., Kloster, S., 2010. Description and evaluation of the Model for Ozone and Related chemical Tracers, version 4 (MOZART-4). Geosci. Model Dev. 3 (1), 43–67. https://doi.org/10.5194/gmd-3-43-2010.

Fioletov, V.E., McLinden, C.A., Krotkov, N., Yang, K., Loyola, D.G., Valks, P., Theys, N., Van Roozendael, M., Nowlan, C.R., Chance, K., Liu, X., Lee, C., Martin, R.V., 2013. Application of OMI, SCIAMACHY, and GOME-2 satellite SO2 retrievals for detection of large emission sources [Article]. J. Geophys. Res.-Atmos. 118 (19), 11399–11418. https://doi.org/10.1002/jgrd.50826.

Garane, K., Koukouli, M.-E., Verhoelst, T., Lerot, C., Heue, K.-P., Fioletov, V., Balis, D., Bais, A., Bazureau, A., Dehn, A., Goutail, F., Granville, J., Griffin, D., Hubert, D., Keppens, A., Lambert, J.-C., Loyola, D., McLinden, C., Pazmino, A., Pommereau, J.-P., Redondas, A., Romahn, F., Valks, P., Van Roozendael, M., Xu, J., Zehner, C., Zerefos, C., Zimmer, W., 2019. TROPOMI/S5P total ozone column data: global ground-based validation and consistency with other satellite missions. Atmos. Meas. Tech. 12 (10), 5263–5287.

Goldberg, D.L., Lu, Z., Streets, D.G., de Foy, B., Griffin, D., McLinden, C.A., Lamsal, L.N., Krotkov, N.A., Eskes, H., 2019. Enhanced Capabilities of TROPOMI NO2: Estimating NOX from North American Cities and Power Plants. Environ. Sci. Technol. 53 (21), 12594–12601. https://doi.org/10.1021/acs.est.9b0448810.1021/acs.est.9b04488.s001.

Griffin, D., Zhao, X., McLinden, C.A., Boersma, F., Bourassa, A., Dammers, E., Degenstein, D., Eskes, H., Fehr, L., Fioletov, V., Hayden, K., Kharol, S.K., Li, S.-M., Makar, P., Martin, R.V., Mihele, C., Mittermeier, R.L., Krotkov, N., Sneep, M., Lamsal, L.N., Linden, M.T., Geffen, J.V., Veefkind, P., Wolde, M., 2019. High-Resolution Mapping of Nitrogen Dioxide With TROPOMI: First Results and Validation Over the Canadian Oil Sands. Geophys. Res. Lett. 46 (2), 1049–1060. https://doi.org/10.1029/2018GL081095.

Howard, C.J., Kumar, A., Malkina, I., Mitloehner, F., Green, P.G., Flocchini, R.G., Kleeman, M.J., 2010. Reactive Organic Gas Emissions from Livestock Feed Contribute Significantly to Ozone Production in Central California. Environ. Sci. Technol. 44 (7), 2309–2314. https://doi.org/10.1021/es902864u.

Hu, B., Jarosch, A.-M., Gauder, M., Graeff-Hönninger, S., Schnitzler, J.-P., Grote, R., Rennenberg, H., Kreuzwieser, J., 2018. VOC emissions and carbon balance of two bioenergy plantations in response to nitrogen fertilization: A comparison of Miscanthus and Salix. Environ. Pollut. 237, 205–217. https://doi.org/10.1016/j.envpol.2018.02.034.

Hu, J., Chen, J., Ying, Q., Zhang, H., 2016. One-year simulation of ozone and particulate matter in China using WRF/CMAQ modeling system. Atmos. Chem. Phys. 16 (16), 10333–10350.

Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L.A., Strickland, M.J., Liu, Y., 2017. Estimating PM2.5 Concentrations in the Conterminous United States Using the Random Forest Approach. Environ. Sci. Technol. 51 (12), 6936–6944. https://doi.org/10.1021/acs.est.7b0121010.1021/acs.est.7b01210.s001.

Huang, X.G., Zhao, J.B., Cao, J.J., Song, Y.Y., 2019. Spatial-temporal Variation of Ozone Concentration and Its Driving Factors in China. Huan Jing Ke Xue 40 (3), 1120–1131. https://doi.org/10.13227/j.hjkx.201807038.

Ialongo, I., Virta, H., Eskes, H., Hovila, J., Douros, J., 2020. Comparison of TROPOMI/Sentinel-5 Precursor NO2 observations with ground-based measurements in Helsinki. Atmos. Meas. Tech. 13 (1), 205–218. https://doi.org/10.5194/amt-13-205-2020.

Jaffe, D.A., Wigder, N., Downey, N., Pfister, G., Boynard, A., Reid, S.B., 2013. Impact of wildfires on ozone exceptional events in the Western u.s. Environ. Sci. Technol. 47 (19), 11065–11072. https://doi.org/10.1021/es402164f.

Jeričević, A., Koračin, D., Jiang, J., Chow, J., Watson, J., Fujita, E., Minoura, H., 2014. Air Quality Study of High Ozone Levels in South California. Springer Netherlands, pp. 629-633. Doi: 10.1007/978-94-007-5577-2_106.

Larkin, A., Geddes, J.A., Martin, R.V., Xiao, Q., Liu, Y., Marshall, J.D., Brauer, M., Hystad, P., 2017. Global Land Use Regression Model for Nitrogen Dioxide Air Pollution. Environ. Sci. Technol. 51 (12), 6957–6964. https://doi.org/10.1021/acs.est.7b0114810.1021/acs.est.7b01148.s00110.1021/acs.est.7b01148.s002.

Lee, T.R., Buban, M., Turner, D.D., Meyers, T.P., Baker, C.B., 2019. Evaluation of the High-Resolution Rapid Refresh (HRRR) Model Using Near-Surface Meteorological and Flux Observations from Northern Alabama. Weather Forecasting 34 (3), 635–663. https://doi.org/10.1175/waf-d-18-0184.1.

Levy II, H., Kasibhatla, P.S., Moxim, W.J., Klonecki, A.A., Hirsch, A.I., Oltmans, S.J., Chameides, W.L., 1997. The global impact of human activity on tropospheric ozone. Geophys. Res. Lett. 24 (7), 791–794. https://doi.org/10.1029/97gl00599.

Li, K.e., Jacob, D.J., Liao, H., Qiu, Y., Shen, L.u., Zhai, S., Bates, K.H., Sulprizio, M.P., Song, S., Lu, X., Zhang, Q., Zheng, B.o., Zhang, Y., Zhang, J., Lee, H.C., Kuk, S.K., 2021. Ozone pollution in the North China Plain spreading into the late-winter haze season. Proc. Natl. Acad. Sci. 118 (10) https://doi.org/10.1073/pnas.2015797118.

Liang, F., Xiao, Q., Huang, K., Yang, X., Liu, F., Li, J., Lu, X., Liu, Y., Gu, D., 2020. The 17-y spatiotemporal trend of PM_{2.5} and its mortality burden in China. Proc. National Acad. Sci. 201919641. https://doi.org/10.1073/pnas.1919641117.

Liang, S., Li, X., Teng, Y., Fu, H., Chen, L., Mao, J., Zhang, H., Gao, S., Sun, Y., Ma, Z., Azzi, M., 2019. Estimation of health and economic benefits based on ozone exposure level with high spatial-temporal resolution by fusing satellite and station observations. Environ. Pollut. 255 (Pt 2), 113267 https://doi.org/10.1016/j.envpol.2019.113267.

Liu, P., Song, H., Wang, T., Wang, F., Li, X., Miao, C., Zhao, H., 2020. Effects of meteorological conditions and anthropogenic precursors on ground-level ozone concentrations in Chinese cities. Environ. Pollut. 262, 114366. https://doi.org/10.1016/j.envpol.2020.114366.

Liu, X.-H., Zhang, Y., Xing, J., Zhang, Q., Wang, K., Streets, D.G., Jang, C., Wang, W.-X., Hao, J.-M., 2010a. Understanding of regional air pollution over China using CMAQ, part II. Process analysis and sensitivity of ozone and particulate matter to precursor emissions. Atmos. Environ. 44 (30), 3719–3727. https://doi.org/10.1016/j.atmosenv.2010.03.036.

Liu, X., Bhartia, P., Chance, K., Spurr, R., Kurosu, T., 2010b. Ozone profile retrievals from the Ozone Monitoring Instrument. Atmos. Chem. Phys. 10 (5), 2521–2537.

Malig, B.J., Pearson, D.L., Chang, Y.B., Broadwin, R., Basu, R., Green, R.S., Ostro, B., 2016. A time-stratified case-crossover study of ambient ozone exposure and emergency department visits for specific respiratory diagnoses in California (2005–2008). Environ. Health Perspect. 124 (6), 745–753.

Monks, P.S., Archibald, A.T., Colette, A., Cooper, O., Coyle, M., Derwent, R., Fowler, D., Granier, C., Law, K.S., Mills, G.E., Stevenson, D.S., Tarasova, O., Thouret, V., Von Schneidemesser, E., Sommariva, R., Wild, O., Williams, M.L., 2015. Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. Atmos. Chem. Phys. 15 (15), 8889–8973. https://doi.org/10.5194/acp-15-8889-2015.

Niu, Y., Cai, J., Xia, Y., Yu, H., Chen, R., Lin, Z., Liu, C., Chen, C., Wang, W., Peng, L., Xia, X., Fu, Q., Kan, H., 2018. Estimation of personal ozone exposure using ambient concentrations and influencing factors. Environ. Int. 117, 237–242. https://doi.org/10.1016/j.envint.2018.05.017.

Nuvolone, D., Petri, D., Voller, F., 2018. The effects of ozone on human health. Environ. Sci. Pollut. Res. Int. 25 (9), 8074–8088. https://doi.org/10.1007/s11356-017-9239-3.

Orru, H., Ebi, K.L., Forsberg, B., 2017. The Interplay of Climate Change and Air Pollution on Health. Curr. Environ. Health Rep. 4 (4), 504–513. https://doi.org/10.1007/s40572-017-0168-6.

Requia, W.J., Di, Q., Silvern, R., Kelly, J.T., Koutrakis, P., Mickley, L.J., Sulprizio, M.P., Amini, H., Shi, L., Schwartz, J., 2020. An Ensemble Learning Approach for Estimating High Spatiotemporal Resolution of Ground-Level Ozone in the Contiguous United States. Environ. Sci. Technol. 54 (18), 11037–11047. https://doi.org/10.1021/acs.est.0c0179110.1021/acs.est.0c01791.s001.

Shi, L., Zanobetti, A., Kloog, I., Coull, B.A., Koutrakis, P., Melly, S.J., Schwartz, J.D., 2016. Low-Concentration PM_{2.5} and Mortality: Estimating Acute and Chronic Effects in a Population-Based Study. Environ. Health Perspect. 124 (1), 46–52. https://doi.org/10.1289/ehp.1409111.

Sicard, P., Crippa, P., De Marco, A., Castruccio, S., Giani, P., Cuesta, J., Paoletti, E., Feng, Z., Anav, A., 2021. High spatial resolution WRF-Chem model over Asia: Physics and chemistry evaluation. Atmos. Environ. 244, 118004. https://doi.org/10.1016/j.atmosenv.2020.118004.

Sicard, P., De Marco, A., Troussier, F., Renou, C., Vas, N., Paoletti, E., 2013. Decrease in surface ozone concentrations at Mediterranean remote sites and increase in the cities. Atmos. Environ. 79, 705–715.

Sicard, P., Serra, R., Rossello, P., 2016. Spatiotemporal trends in ground-level ozone concentrations and metrics in France over the time period 1999–2012. Environ. Res. 149, 122–144. https://doi.org/10.1016/j.envres.2016.05.014.

Stowell, J.D., Bi, J., Al-Hamdan, M.Z., Lee, H.J., Lee, S.-M., Freedman, F., Kinney, P.L., Liu, Y., 2020. Estimating PM2.5 in Southern California using satellite data: factors that affect model performance. Environ. Res. Lett. 15 (9), 094004. https://doi.org/10.1088/1748-9326/ab9334.

Stowell, J.D., Kim, Y.M., Gao, Y., Fu, J.S., Chang, H.H., Liu, Y., 2017. The impact of climate change and emissions control on future ozone levels: Implications for human health. Environ. Int. 108, 41–50. https://doi.org/10.1016/j.envint.2017.08.001.

Tian, Y., Wu, Y., Liu, H., Si, Y., Wu, Y., Wang, X., Wang, M., Wu, J., Chen, L., Wei, C., Wu, T., Gao, P., Hu, Y., 2020. The impact of ambient ozone pollution on pneumonia: A nationwide time-series analysis. Environ. Int. 136, 105498. https://doi.org/10.1016/j.envint.2020.105498.

Veefkind, J.P., Aben, I., McMullan, K., Förster, H., de Vries, J., Otter, G., Claas, J., Eskes, H.J., de Haan, J.F., Kleipool, Q., van Weele, M., Hasekamp, O., Hoogeveen, R., Landgraf, J., Snel, R., Tol, P., Ingmann, P., Voors, R., Kruizinga, B., Vink, R., Visser, H., Levelt, P.F., 2012. TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. Remote Sens. Environ. 120, 70–83.

Watson, G.L., Telesca, D., Reid, C.E., Pfister, G.G., Jerrett, M., 2019. Machine learning models accurately predict ozone exposure during wildfire events. Environ. Pollut. 254, 112792. https://doi.org/10.1016/j.envpol.2019.06.088.

Xiao, Q., Wang, Y., Chang, H.H., Meng, X., Geng, G., Lyapustin, A., Liu, Y., 2017. Full-coverage high-resolution daily PM2.5 estimation using MAIAC AOD in the Yangtze River Delta of China. Remote Sens. Environ. 199, 437–446. https://doi.org/10.1016/j.rse.2017.07.023.

Yue, H., Gebremichael, M., 2020. Evaluation of high-resolution rapid refresh (HRRR) forecasts for extreme precipitation. Environ. Res. Commun. 2 (6), 065004 https://doi.org/10.1088/2515-7620/ab9002.

Zhan, Y., Luo, Y., Deng, X., Grieneisen, M.L., Zhang, M., Di, B., 2018. Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. Environ. Pollut. 233, 464–473. https://doi.org/10.1016/j.envpol.2017.10.029.

Zhao, F., Liu, C., Cai, Z., Liu, X., Bak, J., Kim, J., Hu, Q., Xia, C., Zhang, C., Sun, Y., Wang, W., Liu, J., 2021. Ozone profile retrievals from TROPOMI: Implication for the variation of tropospheric ozone during the outbreak of COVID-19 in China. Sci. Total Environ. 764, 142886. https://doi.org/10.1016/j.scitotenv.2020.142886.

Zhao, H., Zheng, Y., Li, T., Wei, L., Guan, Q., 2018. Temporal and Spatial Variation in, and Population Exposure to, Summertime Ground-Level Ozone in Beijing. Int. J. Environ. Res. Public Health 15 (4), 628. https://doi.org/10.3390/ijerph15040628.