Importance of ozone precursor's measurements in modelling urban surface ozone variability using machine learning model

Vigneshkumar Balamurugan (vigneshkumar.balamurugan@tum.de)
Technical University of Munich (TUM)
Vinothkumar Balamurugan
St. Joseph's Institute of Technology

Research Article

Keywords:

Posted Date: February 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1312561/v1

Importance of ozone precursor's measurements in modelling urban surface ozone variability using machine learning model

Vigneshkumar Balamurugan[1,*] **and Vinothkumar Balamurugan**[2]

[1]Environmental Sensing and Modeling, Technical University of Munich (TUM), Munich, 80333, Germany.
[2]Mechanical Engineering, St. Joseph's Institute of Technology, Chennai, 600119, India.
[*]vigneshkumar.balamurugan@tum.de

ABSTRACT

Surface ozone (O_3) is primarily formed through complex photo-chemical reactions in the atmosphere, which are non-linearly dependent on precursors (NO_X and VOC). Exploring the potential of machine learning (ML) in modeling surface ozone has received little attention, particularly when it comes to the inclusion of limited available ozone precursors information in the ML model. The ML model with past O_3, meteorology (relative humidity, temperature, boundary layer height, wind direction), season type and in-situ NO information explains 87 % (R^2 = 0.87) of the ozone variability over Munich, a German metropolitan area. The ML model trained for the urban measurement station in Munich can also explain the ozone variability of the other three stations in the same city, with R^2 = 0.88, 0.91, 0.63. While the same model robustly explains the ozone variability of two other German cities' (Berlin and Hamburg) measurement stations, with R^2 ranges from 0.72 to 0.84, giving confidence to use the ML model trained for one location to other locations with sparse ozone measurements. In all cases, including coarse CAMS model O_3 simulations in the ML model slightly improves the ML model's performance in predicting surface ozone.

Introduction

In today's world, air quality is always a major concern, and various efforts are being made to address this issue. Despite the fact that anthropogenic emissions have decreased significantly as a result of stringent emission control measures implemented over the last two decades, air quality in many parts of Europe remains poor[1]. Particularly, secondary air pollutants (ozone, secondary particulate matter) formed by complex atmospheric photo-chemical reactions did not show the same trend of decreasing as primary air pollutants, which are emitted directly from primary sources[2]. Ozone (O_3) has a negative impact on both human health and the ecosystem[3,4]. The primary source of ozone in the troposphere is photolysis of nitrogen dioxide (NO_2). Volatile organic compounds (VOCs) play a larger role in ozone production through producing hydrogen oxide radicals (HO_X = OH+HO_2+RO_2) (catalytic cycle), which drive the conversion of NO to NO_2 (NO_X = NO + NO_2)[5,6]. Because of the termination reactions that occur during the catalytic cycle, ozone production is not always directly proportional to the precursor's emission or concentration (NO_X and VOC)[7,8]. As a result, ozone production is widely classified into three regimes: NO_X limited (low NO_X and high VOC), NO_X saturated (high NO_X and low VOC), and transitional[9,10]. Ozone production can be controlled by lowering NO_X in a NO_X limited regime, whereas lowering NO_X can increase ozone production in a NO_X saturated regime. Recent ozone enhancements in urban areas during the COVID-19 lockdown period demonstrate the NO_X saturated regime's ozone production chemistry[11]. Chemical transport models (CTM) are widely used to study the ozone variability[12–16]. However, CTMs have a large bias in resolving complex topography and chemistry mechanisms due to coarser resolution[17,18], for example, urban areas are typically in a NO_X saturated regime, whereas rural areas are being in a NO_X limited regime. In addition, the bias in CTM is exacerbated when emission inventories are uncertain[19]. CTM, on the other hand, necessitate massive computational power.

Machine learning (ML) is gaining traction as an alternative modeling tool to complement CTM in Earth system science fields[20–26]. Because photo-chemical processes have a significant impact on ozone, ML models are trained using a wide range of meteorological variables, many of which drive photo-chemical processes[27–33]. The variability of surface ozone is well explained by the ML model with meteorological information alone[34–36]. Temperature is identified as a key factor in explaining ozone variability in the ML model[37]. Temperature is also a driver of biogenic VOC emissions (a precursor to O_3) in addition to being a driver of photo-chemical processes[7,8]. In the NO_X saturated regime, ozone production is directly proportional to VOC emission (temperature), but in the NO_X limited regime, ozone dependency on VOC shifts to NO_X[38]. Given that many urban areas are currently in a NO_X saturated regime, it is reasonable to expect that ML model trained solely on meteorology will be

able to explain ozone variability. After transitioning to a NO_X limited regime, the ML model trained solely on meteorology may fail to reproduce the surface ozone variability. Previous studies have also shown that the ozone response to temperature has been decreasing in recent years, as urban regions are transitioning to NO_X limited regime[39,40]. However, only a few studies have focused on the inclusion of precursor information into the ML model[30,31,33]. Since stratospheric ozone is highly variable, total column ozone measurements from satellites are unsuitable for studying surface ozone. Satellites, on the other hand, measure the tropospheric column of ozone precursors (NO_2 and HCHO (formaldehyde)), which can be used to study the surface ozone chemistry[41–43]. As CTMs resolve the physical-chemical processes, whereas ML models do not, a hybrid modelling approach that incorporates the CTM prediction as a predictor variable into the ML model may improve the performance of ML model. The objective of this study is to investigate the importance of limited available (in-situ and satellite) ozone precursor information and coarse CTM ozone simulations in modeling urban surface ozone variability using ML model. In addition, we examine the potential of ML model's transfer-ability; how well the ML model trained for one location explains ozone variability of other locations.

Study region, Datasets and Model

This study focuses on Munich, a southern German metropolitan area where air pollutants are currently measured at five different locations. Given the long-term availability of all pollutants data, we chose an urban measurement station (Lothstrasse) to train and test the ML model, which continuously measured O_3, NO_2, NO, and CO from 2001 to 2017. In our study, we also used data (2003 to 2017) from other three stations in Munich (Johanneskirchen-suburban, Allach-suburban, and Stachus-urban) to assess the transfer-ability of the ML model. We also tested the ML model's transfer-ability using data (2015 to 2019) from measurement stations in other German cities, including Berlin (Neukollen-urban, Wedding-urban, and Buch-suburban) and Hamburg (Bramfeld-suburban, Neugraben-suburban, and Sternschanze-urban).

Meteorological variables (temperature, boundary layer height, relative humidity, wind speed and wind direction) are obtained from the ERA 5 reanalysis dataset, with spatial and temporal resolutions of 0.25 degree and 1 hour, respectively[44]. Surface ozone simulations of CAMS (Copernicus Atmosphere Monitoring Service) global reanalysis dataset (EAC4) are also obtained from CAMS data store, which has a spatial resolution of 0.75 degree and a temporal resolution of 3 hour.

The tropospheric column NO_2 and HCHO data from the NASA Aura satellite's OMI (ozone monitoring instrument) are also used[45]. OMI data has a spatial resolution of 13*24 km and a daily temporal resolution. The OMI local overpass occurs between 1 p.m. and 2 p.m., matching with the diurnal maximum ozone. OMI data are available beginning in October of 2004. We filtered the OMI data before using it to include only data with no processing errors, less than 10 % snow or ice cover, a solar zenith angle of less than 80 degree for NO_2 (70 degree for HCHO), and a cloud radiance fraction of less than 0.5. At the end, we only had 689 days of OMI data out of 4809 days (October, 2004 to December, 2017) for Lothstrasse station.

The Extreme Gradient Boosting (XGBoost) algorithm, a supervised learning-gradient boosting algorithm of the ML model[46], is used in this study to model surface ozone concentrations. We train the XGBoost ML model with different predictor categories or combinations of predictor categories (Table 1), and then compare its performance in terms of correlation (R^2), root mean square error (RMSE) and slope of linear fit. The predictor categories are broadly classified into meteorology (temperature, relative humidity, boundary layer height, wind speed and wind direction), in-situ ozone precursors (NO, NO_2 and CO), satellite ozone precursors (column NO_2 and HCHO) and CTM simulations (CAMS model surface O_3). Additionally, we consider two more predictors (day of the week and season), which we include in the meteorology category. We also discuss the predictor variable (feature) importance in the ML model using the results derived from sklearn python library's "feature_importance" function, which calculates feature importance by taking the average gain across all splits (https://scikit-learn.org/stable/auto_examples/ensemble/plot_gradient_boosting_regression.html). For this study, we focus on the afternoon (1 to 2 pm) when ozone levels are at their highest (diurnal maximum). We also performed a similar analysis with the Random Forest (RF) ML model and compare its performance to that of XGBoost.

Results

Performance of ML Model in predicting the urban surface ozone

For the Lothstrasse station in Munich, all in-situ measurements, meteorological variables and CAMS data are available for 5375 days from 2003 to 2017. We divided the 5375 days of measurements into two parts: first 3800 days (70 %) for training, and remaining 1575 days (30 %) for testing the ML predictions. The k-fold cross validation (CV) is used to evaluate the performance of the ML model for different dataset combinations for training and testing. Here we choose k as 10, i.e., 5375 days of data split into 10 parts. The first nine parts are used to train the model, and the final one is used to test the model; this process is repeated ten times for the remaining combinations. The mean of R^2 derived from the k(10)-fold cross validation is then computed. The ML model that was trained solely on meteorology ("ML_met") explains 77 percent of the variance ($R^2 = 0.77$) in measured O_3, with RMSE of 16 μg m^{-3} and a slope of linear fit of about 0.83 (Figure 1 (a)). The mean R^2 of

k(10)-fold CV is 0.77. Wind speed and wind direction have a low importance in the fitted model when compared to other meteorological variables (relative humidity, boundary layer height, and temperature) (Figure S1). In addition, including the day of the week and season in the training dataset ("ML_met_ds") improves the ML model's performance (R^2 = 0.81, RMSE = 14.6 μg m^{-3}, slope = 0.87 and mean R^2 of k(10)-fold CV = 0.81) (Figure 1 (b)). This performance improvement could be attributed to the pronounced seasonal cycle of ozone and weekday-weekend differences. The ozone reaches its maximum in summer and minimum in winter, and due to being in a NO_X saturated regime, weekend ozone levels are higher than the weekdays[11]. The ML model trained solely with CAMS ("ML_cams") or in-situ precursors ("ML_insitu") show poor performance in all terms when compared to ML model trained with the meteorology category alone ("ML_met_ds") (Figure 1 (c,d)).

The ML model trained with meteorology and in-situ precursors category ("ML_met_ds_insitu") performs better than "ML_met_ds", with R^2, RMSE and slope values are about 0.87, 12 μg m^{-3} and 0.87, respectively (Figure 1 (e)). The scatter of predicted O_3 by "ML_met_ds" is largely reduced in "ML_met_ds_insitu", resulting in a lower RMSE. The mean R^2 of k(10)-fold CV is 0.88, which is a 15 % increase over "ML_met". The important feature in "ML_met_ds_insitu" is derived to be in-situ NO measurements, followed by boundary layer height, temperature, and relative humidity. The improvement in performance from "ML_met_ds_insitu" is thus due to the inclusion of NO measurements in the model. The addition of CAMS O_3 simulations with meteorology and in-situ precursors ("ML_met_ds_insitu_cams") further improves the model performance (R^2 = 0.89, RMSE = 10.9 μg m^{-3}, slope = 0.87 and mean R^2 of k(10)-fold CV = 0.9), which is slightly higher than that of "ML_met_ds_insitu" (Figure 1 (f)), with CAMS O_3 simulations being the most important feature (Figure S1). We also performed a similar analysis using Random Forest ML model with a split of 5375 dataset into 70 % / 30 % (training/testing). The overall performance of RF is comparable to that of XGBoost, with modest improvements in some cases (Table S1). Also, when compared to "ML_met_ds" in RF model simulations, the performance of "ML_met_ds_insitu" is improved (in all terms). This supports our earlier findings that including in-situ precursor information is not redundant when modeling surface ozone with ML model.

For 689 days between 2003 and 2017, all in-situ and satellite measurements, meteorological variables, and CAMS data are available. Similarly, we use the first 70 % of data (480 days) for training and remaining 30 % (209 days) for testing the model. Also, we performed the k(10)-fold CV for 689 days of dataset. The performance of the ML model trained with meteorology and satellite precursors ("ML_met_ds_satellite") is, however, equal to the performance of the ML model trained with meteorology alone (Figure 2 (a-d)). This implies that including satellite ozone precursor data had less effect on model performance. In terms of mean R^2 of k(10)-fold CV, the ML model with meteorology, satellite precursors, and the CAMS category provides slightly better results. However, it is poor than that of the ML model trained with meteorology, in-situ precursors, and the CAMS category. The performance difference between ML model with a high (5375) and low (698) number of days is marginal. In all cases, the performance of the ML model with fewer days (698 days) is slightly worse than the performance of the ML model with 5375 days for training and testing. To see how the availability of training dataset affects performance, we train and test the "ML_met_ds_insitu" for varying percentages of data for the 5375 days case (Figure S2). The difference between different dataset combinations for training and testing is also marginal; the 80 % / 20 % (training/testing) dataset performs slightly better than the 20 % / 80 % dataset (lower RMSE by 1.5 μg m^{-3} and higher R^2 by 0.03). However, in this case, 20 % of data equates to nearly three years of data, which may be sufficient to capture all ozone variability by ML model.

We investigated the sensitivity of each predictor variable in the ML model. This is done by excluding the particular predictor variable from the "ML_met_ds_insitu" (Table S2). Temperature is the important feature fitted in model. When temperature is excluded from "ML_met_ds_insitu", the RMSE increases by 1.9 μg m^{-3} and the R^2 decreases by 0.04 compared to all variables included in "ML_met_ds_insitu". Furthermore, at each case, when variable such as season, relative humidity, wind direction, boundary layer height, and in-situ NO is excluded, RMSE increases and R^2 decreases. There are no changes in RMSE and R^2 when the day of the week or wind speed is removed. When in-situ NO_2 or CO is removed, the RMSE decreases in comparison to "ML_met_ds_insitu", indicating that the model is over-fitted when these variables are included. Therefore, we train the ML model only with season, relative humidity, temperature, wind direction, boundary layer height and in-situ NO variables ("ML_s_rh_t_wd_blh_no"), which show slightly better performance in-terms of RMSE decreases by 0.4 μg m^{-3} compared to "ML_met_ds_insitu".

ML Model's transfer-ability

First, we use the "ML_met_ds" trained for "Lothstrasse" station (5375 days) to predict the ozone concentrations of other three stations in Munich, two (Johanneskirchen, Allach) of which are sub-urban and remaining one (Stachus) is urban station. When compared to ground-truth, the performance of "ML_met_ds" for two sub-urban station is better (R^2 = 0.86, 0.81 and RMSE = 12.6, 15.1 μg m^{-3}) than for the urban station (R^2 = 0.5 and RMSE = 20.3 μg m^{-3}) (Figure S3). The predictions are better in all terms when we use "ML_s_rh_t_wd_blh_no", compared to "ML_met_ds", indicating that including precursor information plays an important role in explaining ozone variability of other locations (Figure 3). When including CAMS with "ML_s_rh_t_wd_blh_no" ("ML_s_rh_t_wd_blh_no_cams"), ML model show slightly better performance (Figure S4).

Similarly, we use the “ML_met_ds”, “ML_s_rh_t_wd_blh_no” and “ML_s_rh_t_wd_blh_no_cams” trained for “Lothstrasse” station to predict the ozone concentration of two major German cities (3 stations for each city) (Figure 3, S3, S4). Here, as well, the performance of “ML_s_rh_t_wd_blh_no” is better than “ML_met_ds” in all terms, with R^2 ranges from 0.72 to 0.84 and RMSE ranges from 13.1 to 17.2 μg m^{-3}. When using “ML_s_rh_t_wd_blh_no_cams”, the performance is slightly better than “ML_s_rh_t_wd_blh_no" in terms of R^2 and RMSE, but mixed effects on slope of linear fit. We also performed a ML simulation for the days that have OMI data for all nine stations in Munich, Berlin and Hamburg (Table S3-S5). In all cases, “ML_met_ds_satellite” trained for lothstrasee station performs slightly better than “ML_met_ds” in predicting the ozone concentrations of other locations.

Discussion

In this study, the potential of a machine learning model in simulating urban surface ozone has been demonstrated. As ozone is primarily produced by complex photo-chemical reactions in the atmosphere, the performance of the ML model with meteorology information alone is promising; however, including the precursor emission (NO_X), particularly NO concentration, information further enhance the ML model's performance in predicting the surface ozone. It could be because NO is an important scavenger of O_3 in the urban environment. Due to the scarcity of measurements, we did not use another important ozone precursor (VOC) information in this study, but instead used satellite column HCHO information in the ML model. Because HCHO is an intermediate gas-product of VOC oxidation, it can be used as a proxy for VOC emissions. The addition of a satellite ozone precursor (column NO_2, HCHO) information as a new feature has little effect on the ML model performance. This could be because satellite column NO_2 and HCHO measurements are less sensitive to surface emissions. Furthermore, the coarser resolution of satellite measurements might limit its applicability. This study also reveals that ML model, with O_3, meteorology and precursor information (NO), trained for one location can be used to suitably model the surface ozone concentrations of different locations with sparse ozone measurements. However, the performance of ML model vary by location because other factors (such as VOC emissions and aerosol load) also influence ozone production. Therefore, we advocate for additional research that focuses on specific campaigns that measure all other factors influencing ozone formation and use an ML model to simulate the ozone variability of other locations.

Data availability

The satellite OMI NO_2 and HCHO data can be found at https://disc.gsfc.nasa.gov/. Hourly NO_2, NO, CO and O_3 concentrations are downloaded from European Environment Agency (EEA) website (https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm). Hourly ERA 5 meteorological data are freely available at https://cds.climate.copernicus.eu/. CAMS global reanalysis surface ozone simulations are obtained from CAMS data store (https://ads.atmosphere.copernicus.eu/).

References

1. Air quality in europe 2021. https://www.eea.europa.eu/publications/air-quality-in-europe-2021 (2021).
2. Sicard, P., Agathokleous, E., De Marco, A., Paoletti, E. & Calatayud, V. Urban population exposure to air pollution in europe over the last decades. *Environ. Sci. Eur.* **33**, 1–12 (2021).
3. Zhang, J. *et al.* The acute health effects of ozone and $pm_{2.5}$ on daily cardiovascular disease mortality: a multi-center time series study in china. *Ecotoxicol. environmental safety* **174**, 218–223 (2019).
4. Xie, X. *et al.* Numerical modeling of ozone damage to plants and its effects on atmospheric co_2 in china. *Atmospheric Environ.* **217**, 116970 (2019).
5. Jacob, D. J. *Introduction to atmospheric chemistry* (Princeton University Press, 1999).
6. Jacobson, M. Z. *Fundamentals of atmospheric modeling* (Cambridge university press, 1999).
7. Pusede, S. & Cohen, R. On the observed response of ozone to no x and voc reactivity reductions in san joaquin valley california 1995–present. *Atmospheric Chem. Phys.* **12**, 8323–8339 (2012).
8. Pusede, S. *et al.* On the temperature dependence of organic reactivity, nitrogen oxides, ozone production, and the impact of emission controls in san joaquin valley, california. *Atmospheric Chem. Phys.* **14**, 3373–3395 (2014).
9. Sillman, S., Logan, J. A. & Wofsy, S. C. The sensitivity of ozone to nitrogen oxides and hydrocarbons in regional ozone episodes. *J. Geophys. Res. Atmospheres* **95**, 1837–1851 (1990).
10. Sillman, S. The relation between ozone, no_x and hydrocarbons in urban and polluted rural environments. *Atmospheric Environ.* **33**, 1821–1845 (1999).

11. Balamurugan, V. *et al.* Tropospheric no_2 and o_3 response to covid-19 lockdown restrictions at the national and urban scales in germany. *J. Geophys. Res. Atmospheres* **126**, e2021JD035440 (2021).

12. Bell, M. L. The use of ambient air quality modeling to estimate individual and population exposure for human health research: a case study of ozone in the northern georgia region of the united states. *Environ. international* **32**, 586–593 (2006).

13. Brauer, M. *et al.* Ambient air pollution exposure estimation for the global burden of disease 2013. *Environ. science & technology* **50**, 79–88 (2016).

14. Hu, J., Chen, J., Ying, Q. & Zhang, H. One-year simulation of ozone and particulate matter in china using wrf/cmaq modeling system. *Atmospheric Chem. Phys.* **16**, 10333–10350 (2016).

15. Lou, S., Liao, H., Yang, Y. & Mu, Q. Simulation of the interannual variations of tropospheric ozone over china: Roles of variations in meteorological parameters and anthropogenic emissions. *Atmospheric Environ.* **122**, 839–851 (2015).

16. Wang, Y., Zhang, Y., Hao, J. & Luo, M. Seasonal and spatial variability of surface ozone over china: contributions from background and domestic pollution. *Atmospheric Chem. Phys.* **11**, 3511–3525 (2011).

17. Kumar, R. *et al.* Simulations over south asia using the weather research and forecasting model with chemistry (wrf-chem): chemistry evaluation and initial results. *Geosci. Model. Dev.* **5**, 619–648 (2012).

18. Singh, J. *et al.* Effects of spatial resolution on wrf v3. 8.1 simulated meteorology over the central himalaya. *Geosci. Model. Dev.* **14**, 1427–1443 (2021).

19. Sharma, A. *et al.* Wrf-chem simulated surface ozone over south asia during the pre-monsoon: effects of emission inventories and chemical mechanisms. *Atmospheric Chem. Phys.* **17**, 14393–14413 (2017).

20. Betancourt, C., Stomberg, T., Roscher, R., Schultz, M. G. & Stadtler, S. Aq-bench: a benchmark dataset for machine learning on global air quality metrics. *Earth Syst. Sci. Data* **13**, 3013–3033 (2021).

21. Amato, F., Guignard, F., Robert, S. & Kanevski, M. A novel framework for spatio-temporal prediction of environmental data using deep learning. *Sci. reports* **10**, 1–11 (2020).

22. Gensheimer, J., Turner, A. J., Köhler, P., Frankenberg, C. & Chen, J. A convolutional neural network for spatial downscaling of satellite-based solar-induced chlorophyll fluorescence (sifnet). *Biogeosciences Discuss.* 1–25 (2021).

23. de Hoogh, K. *et al.* Predicting fine-scale daily no2 for 2005–2016 incorporating omi satellite data across switzerland. *Environ. science & technology* **53**, 10279–10287 (2019).

24. Chan, K. L., Khorsandi, E., Liu, S., Baier, F. & Valks, P. Estimation of surface no2 concentrations over germany from tropomi satellite observations using a machine learning method. *Remote. Sens.* **13**, 969 (2021).

25. Zhan, Y. *et al.* Satellite-based estimates of daily no2 exposure in china using hybrid random forest and spatiotemporal kriging model. *Environ. science & technology* **52**, 4180–4189 (2018).

26. Gu, K., Zhou, Y., Sun, H., Zhao, L. & Liu, S. Prediction of air quality in shenzhen based on neural network algorithm. *Neural Comput. Appl.* **32**, 1879–1892 (2020).

27. Liang, Y.-C., Maimury, Y., Chen, A. H.-L. & Juarez, J. R. C. Machine learning-based prediction of air quality. *Appl. Sci.* **10**, 9151 (2020).

28. Amuthadevi, C., Vijayan, D. & Ramachandran, V. Development of air quality monitoring (aqm) models using different machine learning approaches. *J. Ambient Intell. Humaniz. Comput.* 1–13 (2021).

29. Zhang, X., Zhao, L., Cheng, M. & Chen, D. Estimating ground-level ozone concentrations in eastern china using satellite-based precursors. *IEEE Transactions on Geosci. Remote. Sens.* **58**, 4754–4763 (2020).

30. Juarez, E. K. & Petersen, M. R. A comparison of machine learning methods to forecast tropospheric ozone levels in delhi. *Atmosphere* **13**, 46 (2022).

31. Ojha, N. *et al.* Exploring the potential of machine learning for simulations of urban ozone variability. *Sci. reports* **11**, 1–7 (2021).

32. Zhan, Y. *et al.* Spatiotemporal prediction of daily ambient ozone levels across china using random forest for human exposure assessment. *Environ. Pollut.* **233**, 464–473 (2018).

33. Di, Q., Rowland, S., Koutrakis, P. & Schwartz, J. A hybrid model for spatially and temporally resolved ozone exposures in the continental united states. *J. Air & Waste Manag. Assoc.* **67**, 39–52 (2017).

	Predictor variables					Target variable
	Meteorology		In-situ ozone precursors measurement	Satellite ozone precursors measurement	CTMs simulation	In-situ ozone measurement
ML simulation name	T, RH, BLH, WS, WD	DW, S	Surface NO, NO_2, CO	Tropospheric column NO_2, HCHO	CAMS surface O_3 simulations	Surface O_3
ML_met (1)	X					X
ML_met_ds (2)	X	X				X
ML_cams (3)					X	X
ML_insitu (4)			X			X
ML_met_ds_insitu (5)	X	X	X			X
ML_met_ds_insitu_cams (6)	X	X	X		X	X
ML_satellite (7)				X		X
ML_met_ds_satellite (8)	X	X		X		X
ML_met_ds_satellite_cams (9)	X	X		X	X	X

Table 1. Different ML simulation type and associated training data (marked as X). T-Temperature, RH-Relative Humidity, BLH-Boundary Layer Height, WS-Wind Speed, WD-Wind Direction, DW-Day of Week and S-Season. The index of different ML simulation types is given in brackets in the first column, to which we refer in Figure 2.

34. Gong, X., Hong, S., Jaffe, D. A. *et al.* Ozone in china: Spatial distribution and leading meteorological factors controlling o3 in 16 chinese cities. *Aerosol Air Qual. Res.* **18**, 2287–2300 (2018).

35. Hu, C. *et al.* Understanding the impact of meteorology on ozone in 334 cities of china. *Atmospheric Environ.* **248**, 118221 (2021).

36. Brancher, M. Increased ozone pollution alongside reduced nitrogen dioxide concentrations during vienna's first covid-19 lockdown: Significance for air quality management. *Environ. Pollut.* **284**, 117153 (2021).

37. Kovač-Andrić, E., Brana, J. & Gvozdić, V. Impact of meteorological factors on ozone concentrations modelled by time series analysis and multivariate statistical methods. *Ecol. Informatics* **4**, 117–122 (2009).

38. Pusede, S. E., Steiner, A. L. & Cohen, R. C. Temperature and recent trends in the chemistry of continental surface ozone. *Chem. reviews* **115**, 3898–3918 (2015).

39. Otero, N., Rust, H. W. & Butler, T. Temperature dependence of tropospheric ozone under nox reductions over germany. *Atmospheric Environ.* **253**, 118334 (2021).

40. Nussbaumer, C. M. & Cohen, R. C. The role of temperature and no x in ozone trends in the los angeles basin. *Environ. Sci. & Technol.* **54**, 15652–15659 (2020).

41. Jin, X. *et al.* Evaluating a space-based indicator of surface ozone-nox-voc sensitivity over midlatitude source regions and application to decadal trends. *J. Geophys. Res. Atmospheres* **122**, 10–439 (2017).

42. Wang, W., van der A, R., Ding, J., van Weele, M. & Cheng, T. Spatial and temporal changes of the ozone sensitivity in china based on satellite and ground-based observations. *Atmospheric Chem. Phys.* **21**, 7253–7269 (2021).

43. Jin, X., Fiore, A., Boersma, K. F., Smedt, I. D. & Valin, L. Inferring changes in summertime surface ozone–no x–voc chemistry over us urban areas from two decades of satellite and ground-based observations. *Environ. science & technology* **54**, 6518–6529 (2020).

44. Hersbach, H. *et al.* The era5 global reanalysis. *Q. J. Royal Meteorol. Soc.* **146**, 1999–2049 (2020).

45. Levelt, P. F. *et al.* The ozone monitoring instrument. *IEEE Transactions on geoscience remote sensing* **44**, 1093–1101 (2006).

46. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).

Author contributions statement

Vigneshkumar. B. and Vinothkumar. B. conceived the idea of the study and performed the modelling work. Vigneshkumar. B. wrote the manuscript. Vinothkumar. B. reviewed and edited the manuscript.

Competing interests The authors declare no conflicts of interest relevant to this study.

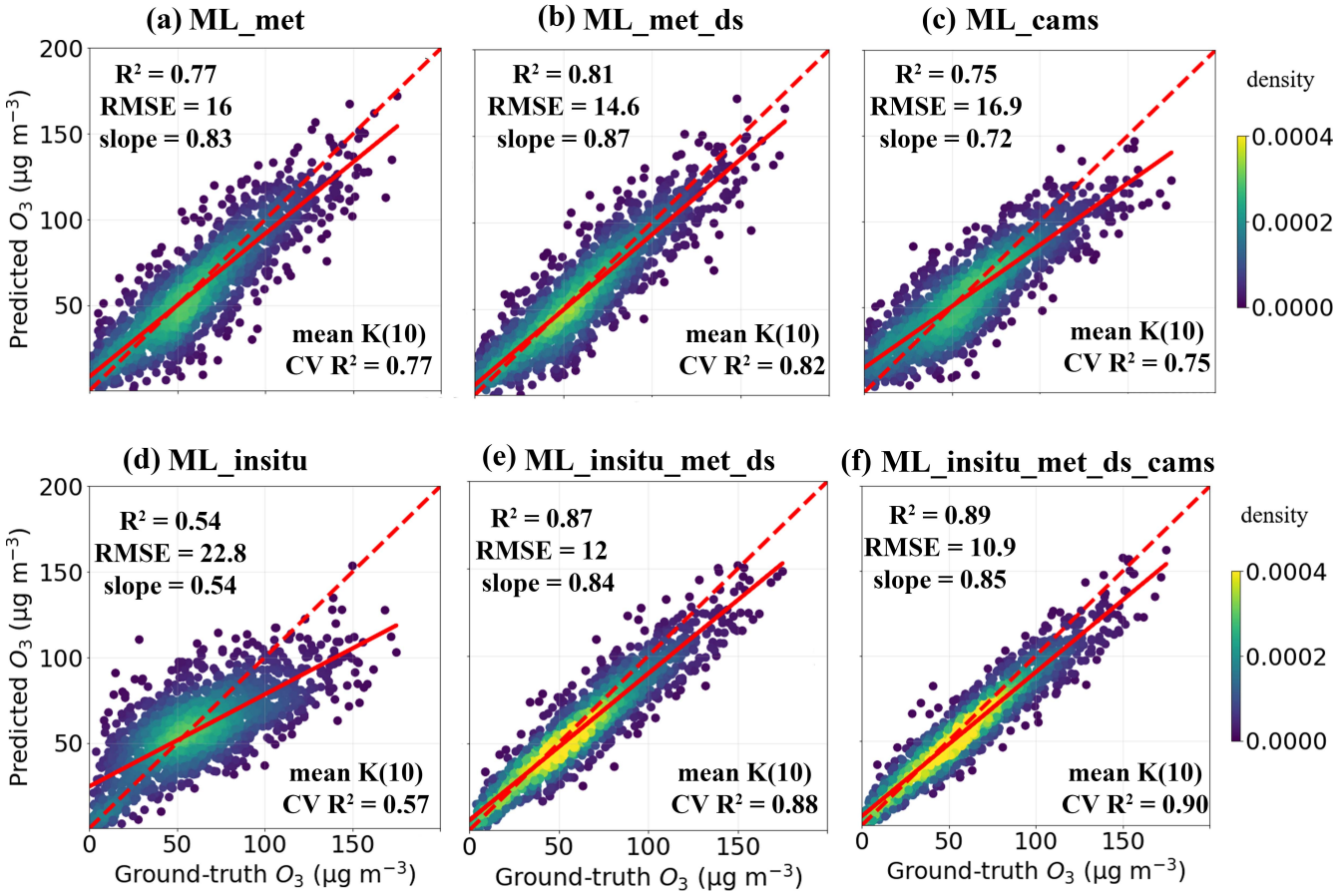

Figure 1. Density scatter plots of predicted ozone by different ML simulation type vs ground-truth ozone at Lothstrasse station at Munich. In a total of 5375 days (between 2003 to 2017), first 3800 days used for training and remaining 1575 days used for testing. Mean R^2 of k(10)-fold cross validation is also given at bottom of figure panels at each case. Red solid line represents the linear fit and red dotted line represents 1:1 line.

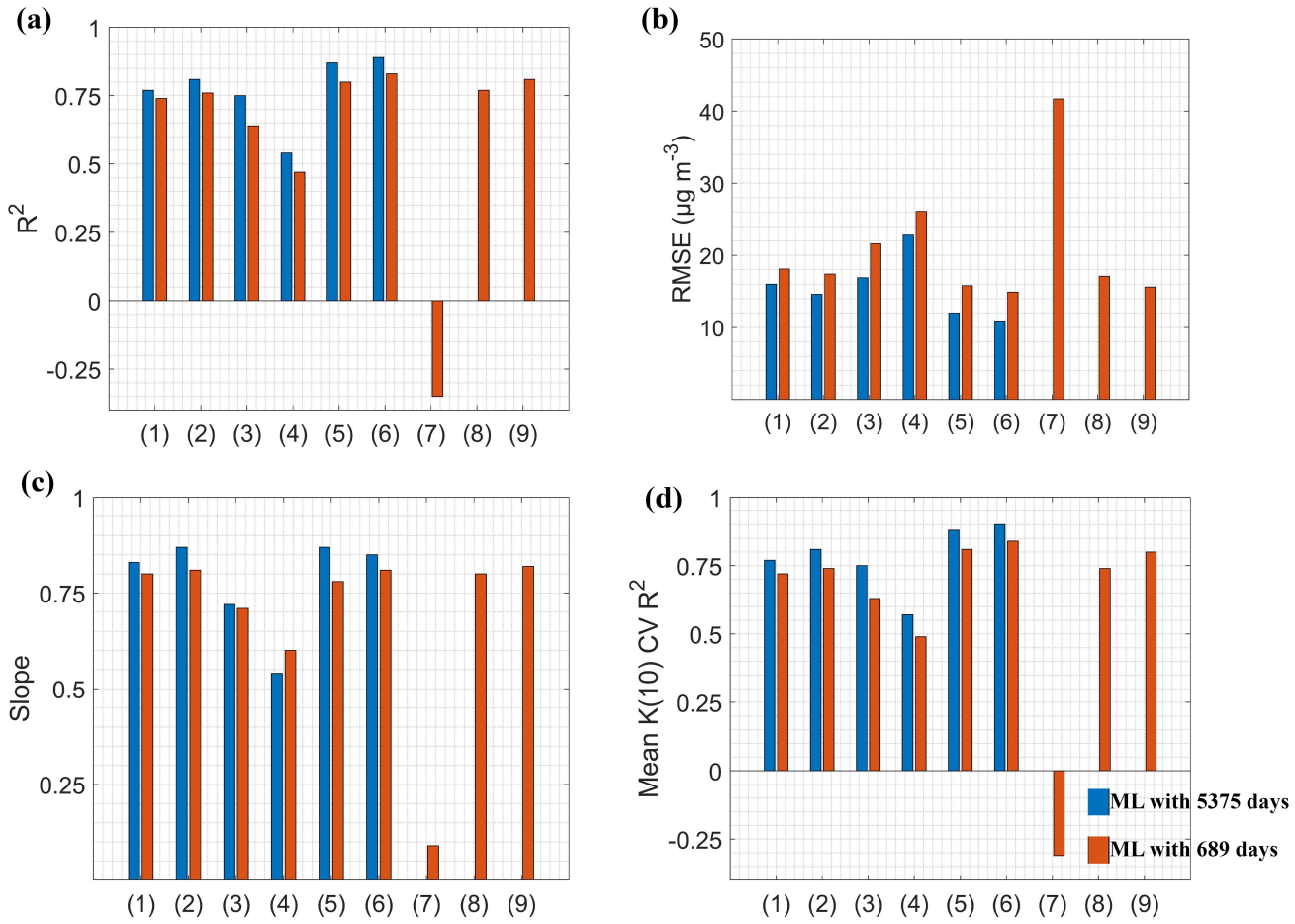

Figure 2. Performance comparison of different ML simulation types with 5375 days (blue) and 689 days (red) for training and testing (70 % / 30 %). X axis indexes refer to the index of different ML simulation type (Table 1).

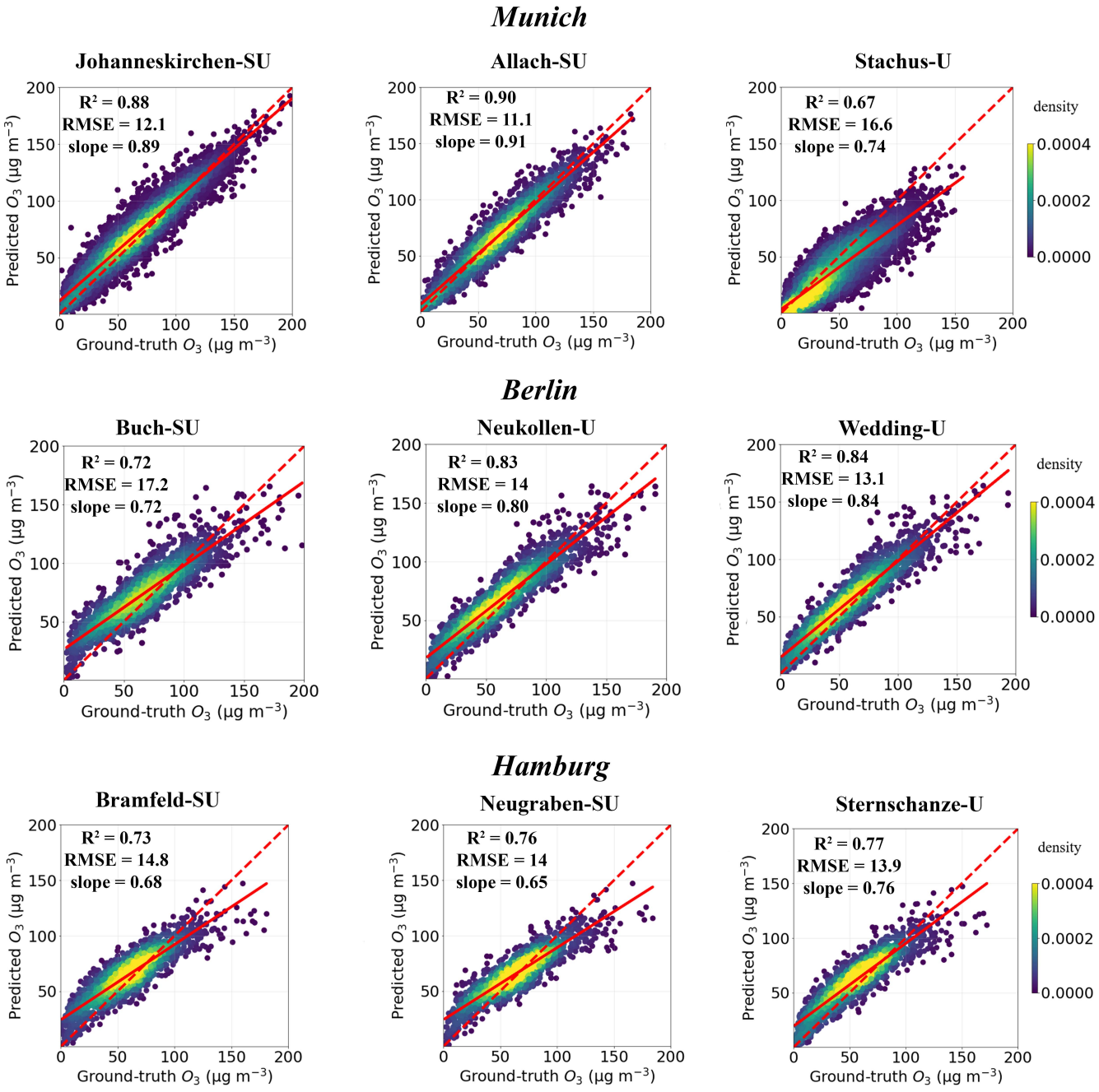

Figure 3. Density scatter plots of predicted ozone by "ML_s_rh_t_wd_blh_no" trained for Lothstrasse station at Munich vs ground-truth ozone measurements for different locations. First row shows the stations for Munich, second row for Berlin and third row for Hamburg stations. U represents urban station and SU represents suburban station.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial.pdf