Dear Reviewer,

We appreciate your comments and suggestions, which have helped us improve our manuscript further. We have made the necessary changes to the manuscript, which can be found in the attached file (Track Changes). The following is a response to your comments and suggestions. Corresponding changes in the revised manuscript are also made available below, if applicable, at the appropriate places.

Sincerely,

On behalf of all co-authors,

Vigneshkumar Balamurugan

Response to Reviewer-1:

The authors explored the gradient boosted tree approach for spatial-temporal modelling of NO2 and O3 and applied it to the case in Germany. There are some issues to address in the revised version:

Thank you so much for reading and reviewing our manuscript! We carefully reviewed and considered your comments/suggestions, and made improvements in the revised manuscript.

Validations:

Table 1 lists the types of datasets used in this study. May you clarify which dataset was used for the ground-truth data?

In the revised manuscript (Table 1), we now included the purpose of the data.

Table 1. Data sets and related information used in this study.

Data source	Data (purpose)	Temporal resolution	Spatial resolution
Governmental in situ measurements	Near-surface NO_2 and O_3 (Ground-truth data)	1 hr	-
TROPOMI satellite measurements	Tropospheric column NO_2, total column O_3 and total column HCHO (Input features)	Daily	7 km*3.5 km (5.5 km*3.5 km, after 6 August 2019)
ERA5 (ECMWF reanalysis)	Temperature, relative humidity, wind speed, wind direction, downwind UV solar radiation at surface, boundary layer height, surface pressure and temperature of air at 2m above the surface (Input features)	1 hr	0.25*0.25-degree
U.S. Geological Survey	Surface elevation (Input features)	-	1*1-km
GRIP global roads database	Road density (Input features)	-	8*8-km
CAMS European air quality forecasts	Near-surface NO_2 and O_3 (for validation)	1 hr	0.1*0.1-degree

GEOS-Chem chemical transport model	Near-surface NO_2 and O_3 (for disentangling meteorology impacts)	1 hr	0.5*0.625-degree

Figures 5-6 show the spatial distribution of the averaged NO2 and O3 during the study period. Is the study period between 2019-07-17 and 2020-01-31? May you specify which months were used for Summer, Spring, Autumn, and Winter?

In figure 5 (and 6), the averaged NO_2 (and O_3) concentrations are between 2018-04-30 and 2021-07-01. We have updated the figure captions in both Figure 5 and Figure 6 to include the study period as well as the specific months used to calculate the seasonal averages.

Figure 5. (a) Averaged GBT-simulated daily near-surface NO_2 concentrations over the study domain for the study period between 2018-04-30 and 2021-07-01. (b-e) Averaged GBT-simulated daily near-surface NO_2 concentrations for each season during the study period. Winter: December, January and February. Spring: March, April and May. Summer: June, July and August. Autumn: September, October and November.
Figure 6. (a) Averaged GBT-simulated daily near-surface O_3 concentrations over the study domain for the study period between 2018-04-30 and 2021-07-01. (b-e) Averaged GBT-simulated daily near-surface O_3 concentrations for each season during the study period. Winter: December, January and February. Spring: March, April and May. Summer: June, July and August. Autumn: September, October and November.

The data sets were pre-processed in daily scale. Could you please generate a spatial map illustrating the average daily concentrations of NO2 and O3 during Summer and Winter, instead of considering the seasonal averages? Furthermore, may you compare these results with reanalysis from CAMS?

For Figures 5 and 6, the seasonally average NO_2 (and O_3) values were not simulated. The Machine learning model was used to simulate daily NO_2 and O_3 concentrations spatial map, and daily maps were averaged for each season, as shown in Figure 5 (and 6). We also modified the figure 5 and 6 captions to make it clearer to the reader. We hope this clarifies your comment.

Figure 5. (a) Averaged GBT-simulated daily near-surface NO_2 concentrations over the study domain for the study period between 2018-04-30 and 2021-07-01. (b-e) Averaged GBT-simulated daily near-surface NO_2 concentrations for each season during the study period. Winter: December, January and February. Spring: March, April and May. Summer: June, July and August. Autumn: September, October and November.
Figure 6. (a) Averaged GBT-simulated daily near-surface O_3 concentrations over the study domain for the study period between 2018-04-30 and 2021-07-01. (b-e) Averaged GBT-simulated daily near-surface O_3 concentrations for each season during the study period. Winter: December, January and February. Spring: March, April and May. Summer: June, July and August. Autumn: September, October and November.

CAMS European air quality forecasts are only available for three years in the rolling archive. Therefore, we only compare the CAMS product for the period between 2019-07-17 and 2020-31-01 (Figure A5 and A6).

Line 131, “we also included “Near-surface NO2” modeled from NO2 ML model as a feature variable in the O3 ML model.” However, in Figure 3 (d), the Near-surface NO2” modeled from NO2 ML model is not listed. I guess the Near-surface NO2” modeled from NO2 ML model will be top one affecting the O3 predive results. Is this case? Maybe you can use the ML model to get the direct relationship between O3 and Near-surface NO2” modeled from NO2 ML model.

Yes. We agree with the reviewer that ML modeled near-surface NO_2 is one of the most important factors influencing O_3 predictive results. Based on our results, it is the sixth most important feature. In figure 3(d), "ML modeled near-surface NO_2" is given as "in-situ NO_2". This is changed in the revised manuscript (Figure 3).

When using machine learning models, the direct relationship between variables, such as NO_2 and O_3, cannot be obtained as deterministic equations. Instead, one can analyze the feature importance or variable importance provided by the model.

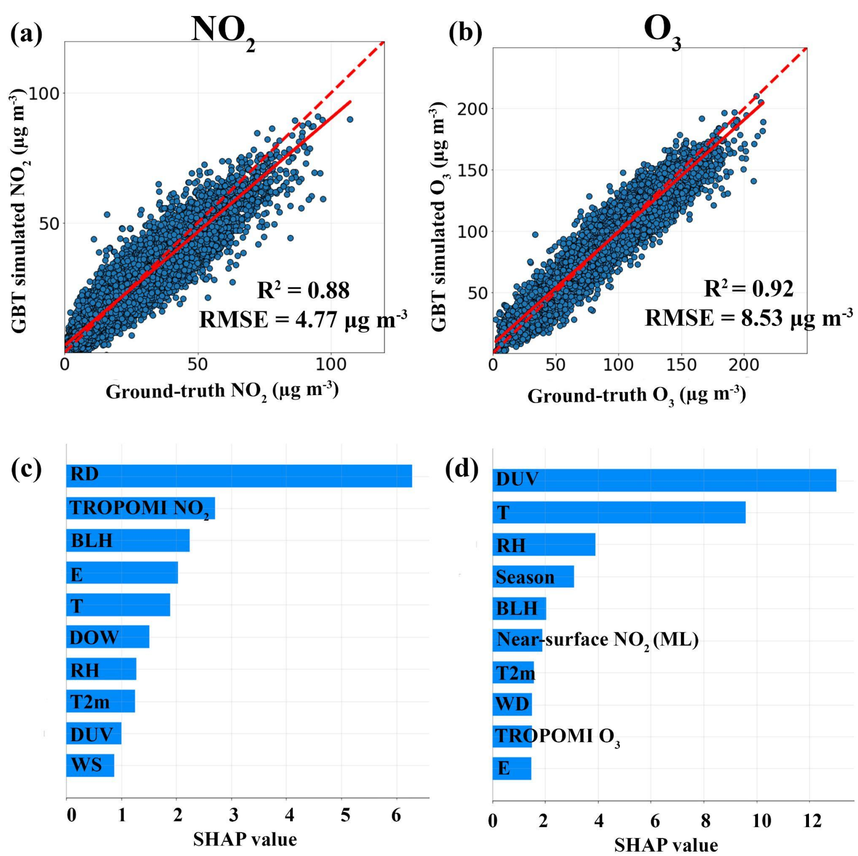

Figure 3. Comparison between ground-truth and GBT-simulated near-surface NO_2 (a) and O_3 (b). Feature importance (top 10) calculated based on SHAP (SHapley Additive exPlanations)

values for NO_2 (c) and O_3 (d) GBT model. RD: Road Density, BLH: Boundary Layer Height, E: Surface Elevation, T: Temperature, DOW: Day of the week, RH: Relative Humidity, T2m: Temperature at 2 meter height, DUV: Downwind UV radiation, WS: Wind speed, WD: Wind Direction.

Line 243, “After the discussed model evaluation, we trained the GBT model using 100% of the data and modeled the near-surface NO2 and O3 concentrations over the study domain at 0.1 degree resolution and daily”, It is not clear here. Are you re-train the model? How do you validate your model?

Yes. We trained the model using 100% of data after performing different ML model evaluations, which is a common practice in machine learning to leverage all available information and avoid losing any valuable data. After using 100% of the data for training, the model can only be evaluated with ground-truth data beyond the study period. However, in our ML model evaluations, we followed different validation approaches that involved more than just training and evaluating a single model. For example, we employed evaluation strategies such as the five-fold random/time/location-leave-out method. These methods enabled us to train five ML models by systematically leaving out different subsets of the whole dataset during each fold validations. Therefore, we believe our ML model would perform similarly on future data, as the models' performance on unseen data yielded robust estimates of their generalization ability during the different evaluation strategies.

Response to Reviewer-2:

General comments

The authors develop a machine learning framework for modeling NO2 and O3 concentrations in Germany, and based on that, they analyze human exposure to the two air pollutants and the effects of COVID quarantine. The authors also discuss the transferability of their model.

The manuscript is well organized and in particular the methodology is thoroughly described. However, before it can be published, I believe the authors should address the comments below.

Thank you so much for taking the time to read and review our manuscript! We carefully reviewed and considered your comments/suggestions, and as a result, we improved the manuscript.

Specific comments

Line 129: Does the "season" (season of the year) information in the ML model have only 4 values? In my opinion, "day of the year" would be a more ideal feature to help the model learn the daily variability of air pollutants. The author should try or clarify this.

Thank you for your suggestion! We have evaluated the ML model using both "Day of the Year" and "season of the year" as features in all our evaluation strategies. We noted that there is a slightly worse performance in both the NO_2 and O_3 GBT model (Table R1 and R2), when using "Day of the year" as a feature instead of "season of the year". Therefore, we decided to use "season of the year" instead of "Day of the Year" in our study.

Table R1. Evaluation metrics of our GBT model in different testing strategies (using "Season of the Year" as a feature).

		Random (1-fold)	Random (5-fold)	Time-leave-out (5-fold)	Location-leave -out (5-fold)
NO_2 GBT model	**R^2**	0.88	0.89±0.002	0.74±0.07	0.68±0.12
	RMSE (μg m^{-3})	4.77	4.65±0.034	6.77±0.7	8.67±1
O_3 GBT model	**R^2**	0.92	0.92±0.001	0.74±0.09	0.8±0.06
	RMSE (μg m^{-3})	8.53	9.36±0.068	13.2±1.01	12.45±1.26

Table R2. Evaluation metrics of our GBT model in different testing strategies (using "Day of the Year" as a feature).

		Random (1-fold)	Random (5-fold)	Time-leave-out (5-fold)	Location-leave -out (5-fold)
NO_2 GBT model	**R^2**	0.88	0.89±0.002	0.74±0.061	0.68±0.14
	RMSE (μg m^{-3})	4.76	4.67±0.05	6.76±0.68	8.74±1.3
O_3 GBT model	**R^2**	0.91	0.90±0.001	0.72±0.09	0.78±0.06
	RMSE (μg m^{-3})	8.60	9.82±0.054	13.6±1.16	12.96±1.21

Line 131: Given the coupled nature of NO2 and ozone, I would suggest the authors try to include O3 as a feature in the NO2 ML model, like why they did the same way for O3 model, or please clarify why they didn't do so.

If we include the ML-modeled O_3 in the NO_2 ML model iteratively, we believe the ML model may suffer from overfitting. For example, O_3 could become an important feature as it already contains information about NO_2 (it is important to note that the ML-modeled NO_2 is the sixth most important feature). Additionally, the errors from both the NO_2 and O_3 ML models in the first iteration would propagate and potentially amplify the errors.

Line 148: 24h-mean of ERA-5 data makes sense for NO2 model, but I would suggest the authors to test daytime-mean or daily-max for O3 model, as ozone is calculated as MDA8. This is especially the case for daily-max 2m temperature, which has been shown to be well correlated with MDA8 ozone.

Thanks for the suggestion! Before deciding on the 24-hr mean of meteorology as a feature for the O_3 GBT model, we also conducted a test on the maximum O_3-time (10 - 6 local time), when maximum 8-hr O_3 concentration occurs (Figure R1). When we used maximum O_3-time mean as a feature, we noted a similar performance, compared to 24-hr mean as a feature (Table R3 and R4). Therefore, we chose a 24-hour mean for both the NO_2 and O_3 models.

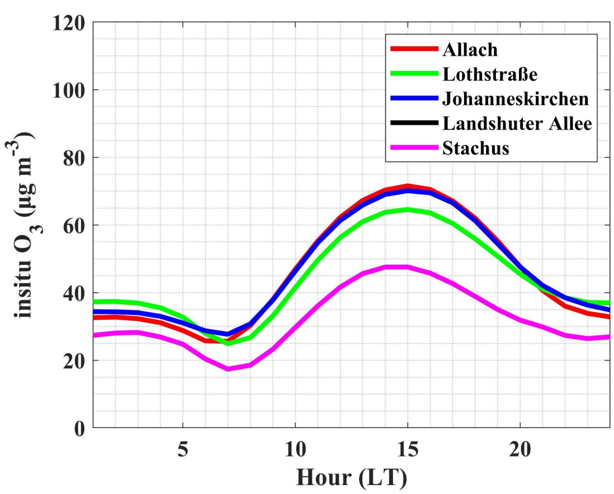

Figure R1. The diurnal mean O_3 averaged between 2010 and 2019.

Table R3. Evaluation metrics of our O_3 GBT model in different testing strategies (using “24-hr mean of meteorology” as a feature).

		Random (1-fold)	Random (5-fold)	Time-leave-out (5-fold)	Location-leave-out (5-fold)
O_3 GBT model	**R^2**	**0.92**	**0.92±0.001**	**0.74±0.09**	**0.8±0.06**
	RMSE (μg m^{-3})	**8.53**	**9.36±0.068**	**13.2±1.01**	**12.45±1.26**

Table R4. Evaluation metrics of our O_3 GBT model in different testing strategies (using “maximum O_3-time of meteorology” as a feature).

		Random (1-fold)	Random (5-fold)	Time-leave-out (5-fold)	Location-leave-out (5-fold)
O_3 GBT model	**R^2**	**0.91**	**0.92±0.001**	**0.75±0.09**	**0.79±0.07**
	RMSE (μg m^{-3})	**8.6**	**8.8±0.054**	**13.16±1.16**	**12.65±1.47**

Line 160: Authors should give the exact size of data samples (both training and testing set), as text or labelled on the figure.

In the revised manuscript, we added the training and test sample size in the corresponding locations.

Line 196-197	The trained GBT model with 70% of the data (78433) for NO_2 reproduced the observed NO_2 concentration well in the test case (33615), with an R^2 of 0.88 and RMSE of 4.77 μg m^{-3}.

Line 215-216	The GBT model trained with 70% of the data (65705) for O_3 also well reproduced the observed O_3 concentrations in the test case (28160), with an R^2 of 0.92 and RMSE of 8.53 μg m^{-3}.

Line 205: It is interesting to see that road density is the most important feature, given that it has constant values which don't show temporal variations. Can the authors explain this further?

We agree with the reviewer that road density doesn't show temporal variation for a particular location. However, in our study, we developed a ML model for the whole Germany domain, in which spatial variation in road density explains the majority of the near-surface NO_2 variation. Therefore, road density is the most important feature in our ML model.

Line 229 (and also line 153): The fact that MLP is worse than GBT can be interesting or maybe controversial here, as people now tend to believe that deep learning techniques should outperform light-weight algorithms such as GBT. The authors should explain more about this, as it is an important and perhaps new finding. Personally, I can think of a few questions below that might help clarify this.

- **What is tabular/structured data and what is non-tabular/structured data? Is the data we use for air pollutants prediction usually of the former type?**

  In this study, we prepared the data as structured data format. Tabular/Structured Data and Non-Tabular/Unstructured Data are the terms used to categorize different types of data based on their format. Tabular/structured data refers to data that is organized in a tabular format, similar to a table or spreadsheet. Most ML models, such as decision trees, SVR, and feedforward neural networks, take this type of input.

  Non-tabular/unstructured data refers to data that does not have a predefined structure and does not necessarily fit into rows and columns. It can include text, images, audio, video, or other formats that do not conform to a table-like structure. Typically, ML models such as CNN and GAN are used to handle these types of inputs.

- **Is the use of tabular/structured data the only reason why GBT outperforms MLP in this study? Is it possible that the size of the data**

samples limits the capability of MLP, given that it is a deep learning technique after all?

The use of tabular data could be one of the reasons for the better performance of GBT compared to MLP. The GBT algorithm is known for its ability to capture feature interactions effectively, which can be particularly advantageous when dealing with tabular data. On the other hand, the MLP algorithm might require a larger number of hidden layers and neurons to achieve similar performance. Additionally, the performance of MLP can also be affected by the sample size. Deep learning algorithms, including MLP, are known to be data-hungry and often require a large amount of data to generalize well. We have included a discussion on the sample size and other neural network algorithms in the revised manuscript.

Line 231-235	It is important to note that deep learning models are data-intensive, and their performance and generalization capabilities tend to improve with larger amounts of data. In our study, we utilized the simplest deep learning algorithm known as MLP. However, it is essential to explore the capabilities of other deep learning algorithms, such as CNN and LSTM, in future studies to gain further insights. Additionally, employing multiple ML models through bagging techniques could potentially lead to improved performance, despite the computational expense involved.

- **In addition to the work of Heaton and Lundberg et al, can the authors find any other studies that have focused on the prediction of air pollutants that can support the results of this study?**

  There have been numerous studies conducted on deep learning models (Chan et. al., 2021) and traditional machine learning models (Zhu et. al., 2022) like Random Forest, XGBoost, etc., individually, to model air pollutant concentrations. However, we have not come across any studies that support our findings regarding the comparison between deep learning and tree-based models.

- **What about other neural network techniques? The author may not need to try them, but at least give a brief discussion, as MLP is one of the simplest deep learning algorithms.**

Thanks for the suggestion. In the revised manuscript, we have included a discussion on the sample size and other neural network algorithms in the revised manuscript, as given below.

Line 231-235	It is important to note that deep learning models are data-intensive, and their performance and generalization capabilities tend to improve with larger amounts of data. In our study, we utilized the simplest deep learning algorithm known as MLP. However, it is essential to explore the capabilities of other deep learning algorithms, such as CNN and LSTM, in future studies to gain further insights. Additionally, employing multiple ML models through bagging techniques could potentially lead to improved performance, despite the computational expense involved.

Section 3.3: In this section, the authors discuss the exceedances of NO2 and O3 using data produced by the GBT model, but the model's ability to capture extreme pollution is hardly evaluated in the validation section above. In fact, the scatter plot of Figure 4 indicates model does have a weakness in reproducing large NO2/O3 values. Therefore, I would suggest that the authors add this uncertainty discussion when analyzing people living beyond the WHO limit.

Thanks for the suggestion. We agree with the reviewer that our model has some difficulty in capturing the extreme pollution events, as shown in figure 4. In order to evaluate the model capability in capturing the exceedance events (above WHO limit), we used the time-leave-out evaluation strategy. This approach is chosen because comparing the ML model simulations (after training with 100 % of data) with ground-truth is questionable as it was already used during the training process. In time-leave-out strategy, the exceedances of NO_2 and O_3 values simulated by GBT model are compared with Ground-truth exceedance events in each iteration. This allows us to assess the model's ability to reproduce the exceedance data that has not been used in the training process.

In both the NO_2 and O_3 GBT models, 82% of the WHO NO_2 and O_3 exceedance data in the whole dataset (Ground-truth) were correctly identified as WHO NO_2 and O_3 exceedance (True Positives), meaning 18% of actual WHO exceedances have not been identified as such by our GBT models (False Negatives). However, we also noted that 6.6% and 7.3% (False Positives) of the whole data were incorrectly identified as exceedance data by our NO_2 and O_3 GBT models, respectively (Table A6).

This discussion and table are included in the revised manuscript, as given below.

Line 269-276	We also evaluated the model capability in capturing the exceedance events (above WHO limit) using time-leave-out evaluation strategy. The exceedances of NO_2 and O_3 events simulated by GBT model compared with Ground-truth events in each iteration. This allows us to assess the model's ability to reproduce the exceedance events that have not been used in the training process. The 82% of the WHO NO_2 and O_3 exceedance events in the whole dataset (Ground-truth) were correctly identified as WHO NO_2 and O_3 exceedance events (True Positives) in both the NO_2 and O_3 GBT models (Table A5). However, we also noted that 6.6% and 7.3% of the whole data were incorrectly identified as exceedance events by our NO_2 and O_3 GBT models, respectively (False Positives). This indicates that our GBT model might slightly underestimate the exceedance events for both NO_2 and O_3. This could be due to unknown drivers that are not included in the model.

Table A6. Comparison between WHO NO_2 and O_3 exceedance events in the ground-truth dataset and GBT simulated WHO NO_2 and O_3 exceedance events using time-leave-out testing strategy.

	Ground-truth WHO exceedance	**Correct detection as exceedance by GBT model (True Positives)**	**Incorrect detection as exceedance by GBT model (False Positives)**
Near-surface NO_2	36772	30125	7439
Near-surface O_3	35860	29396	6924

In addition, a temporal evaluation of the daily time-series (CAMS/GBT versus ground-truth O3) may be meaningful, such as using the temporal correlation coefficient.

As discussed above, it is questionable to compare the ground-truth O_3 values to the model predictions (after training with 100 % of ground-truth data). This is because the model is fitted based on the ground-truth O_3. However, we compared CAMS vs Ground-truth and GBT vs Ground-truth for the period between 17-07-2019 and

31-01-2020 (this time period was not used for training the GBT model for this comparison). This evaluation strategy involves comparing the model predictions with the ground-truth O_3 for a particular period, which is not included in the training dataset (Figure 4). The outcome of this evaluation, along with the results of the time-leave-out evaluation strategy results, provides valuable insight into the model's temporal correlation coefficient.

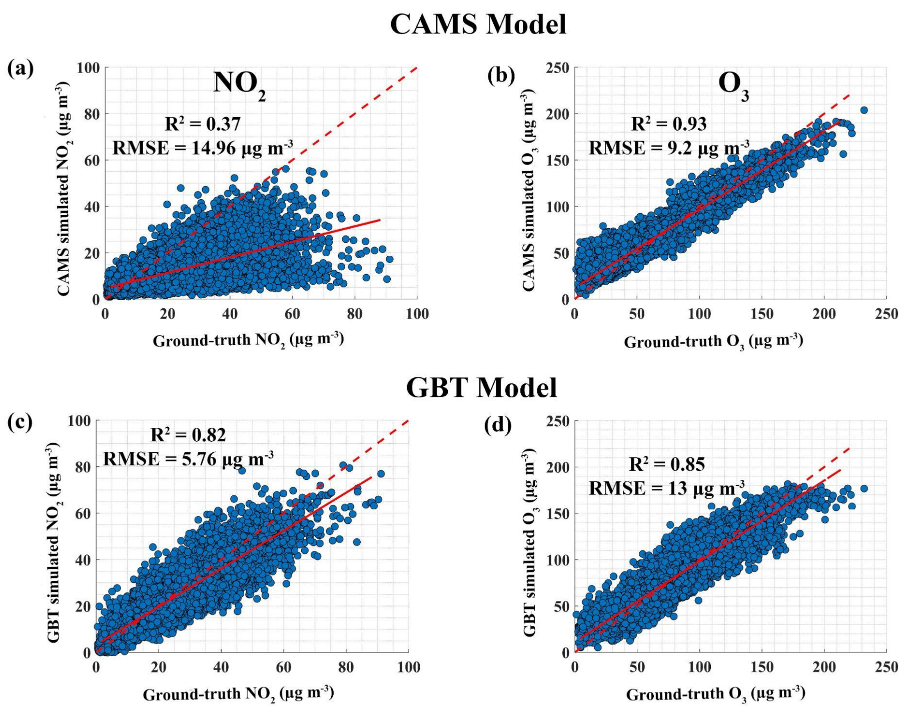

Figure 4. Top: Comparison between ground-truth near-surface NO_2 and CAMS reanalysis near-surface NO_2 (a) and O_3 (b) for the period between 17-07-2019 and 31-01-2020. Bottom: Comparison between ground-truth near-surface NO_2 and GBT-simulated near-surface NO_2 (c) and O_3 (d) for the period between 17-07-2019 and 31-01-2020. The dotted line represents a 1:1 line, while the solid line represents a linear fit.

References:

Chan, K. L., Khorsandi, E., Liu, S., Baier, F., and Valks, P.: Estimation of surface NO2 concentrations over Germany from TROPOMI satellite observations using a machine learning method, Remote Sensing, 13, 969, 2021.

Zhu, Q., Bi, J., Liu, X., Li, S., Wang, W., Zhao, Y., and Liu, Y.: Satellite-Based Long-Term Spatiotemporal Patterns of Surface Ozone Concentrations in China: 2005–2019, Environmental health perspectives, 130, 027 004, 2022.

Spatio-temporal modeling of air pollutant concentrations in Germany using machine learning

Vigneshkumar Balamurugan[1], Jia Chen[1], Adrian Wenzel[1], and Frank N. Keutsch[2,3]

[1]Environmental Sensing and Modeling, Technical University of Munich (TUM), Munich, Germany.
[2]School of Engineering and Applied Science, Harvard University, Cambridge, MA, USA.
[3]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA.

Correspondence: Vigneshkumar Balamurugan (vigneshkumar.balamurugan@tum.de), Jia Chen (jia.chen@tum.de)

Abstract. Machine learning (ML) models are becoming a meaningful tool for modeling air pollutant concentrations. ML models are capable of learning and modeling complex non-linear interactions between variables, and they require less computational effort than chemical transport models (CTMs). In this study, we used gradient boosted tree (GBT) and multi-layer perceptron (MLP; neural network) algorithms to model near-surface nitrogen dioxide (NO_2) and ozone (O_3) concentrations over Germany at 0.1 degree spatial resolution and daily intervals.

We trained the ML models using TROPOMI satellite column measurements combined with information on emission sources, air pollutant precursors and meteorology as feature variables. We found that the trained GBT model for NO_2 and O_3 explained a major portion of the observed concentrations (R^2 = 0.68-0.88, RMSE = 4.77-8.67 μg m^{-3} and R^2 = 0.74-0.92, RMSE = 8.53-13.2 μg m^{-3}, respectively). The trained MLP model performed worse than the trained GBT model for both NO_2 and O_3 (R^2 = 0.46-0.82 and R^2 = 0.42-0.9, respectively).

Our NO_2 GBT model outperforms the CAMS model, a data-assimilated CTM, but slightly under-performs for O_3. However, our NO_2 and O_3 ML models require less computational effort than CTM. Therefore, we can analyze people's exposure to near-surface NO_2 and O_3 with significantly less effort. During the study period (2018-04-30 and 2021-07-01), it was found that around 36% of people lived in locations where the WHO NO_2 limit was exceeded for more than 25% of the days, while 90% of the population resided in areas where the WHO O_3 limit was surpassed for over 25% of days. Although metropolitan areas had high NO_2 concentrations, rural areas, particularly in southern Germany, had high O_3 concentrations.

Furthermore, our ML models can be used to evaluate the effectiveness of mitigation policies. Near-surface NO_2 and O_3 concentrations changes during the 2020 COVID-19 lockdown period over Germany were indeed reproduced by the GBT model, with meteorology-accounted for near-surface NO_2 significantly decreased (by 23$\pm$5.3%) and meteorology-accounted for near-surface O_3 slightly increased (by 1$\pm$4.6%) over ten major German metropolitan areas, compared to 2019. Finally, our O_3 GBT model is highly transferable to other countries, at least to neighboring countries and locations where no measurements are available (R^2 = 0.87-0.94), whereas our NO_2 GBT model is moderately transferable (R^2 = 0.32-0.64).

1 Introduction

Air pollution is a major threat to human health and impacts ecosystems (Bell et al., 2011; Lelieveld et al., 2015; Zhang et al., 2019; Xie et al., 2019). Based on the source, air pollutants are classified as primary (directly emitted from anthropogenic/-natural sources) or secondary (formed through complex atmospheric chemical reactions). Near-surface nitrogen oxide (NO_X= $NO+NO_2$) is a primary air pollutant emitted largely by fossil-fuel-consuming sectors such as vehicles, industries, power plants, etc., but there are also natural sources such as lightning, soil emissions, and biomass burning. Near-surface ozone (O_3) is a secondary air pollutant produced solely by the photolysis of NO_2 (nitrogen dioxide) in the presence of sunlight (Crutzen, 1988; Council et al., 1992).

Tropospheric NO_X and O_3 are chemically strongly coupled via complex atmospheric chemical reactions (Jacob, 1999). The majority of NO_X, from primary sources such as fossil-fuel combustion, is emitted in the form of nitric oxide (NO), which rapidly converts to NO_2 by reacting with O_3. In turn, O_3 and NO are generated again by photolysis of NO_2, forming a null cycle. Therefore, the amount of sunlight present and the total concentration of NO_X determine ozone production via this NO_X null cycle. However, the oxidation of volatile organic compounds (VOCs) via the hydroxyl (OH) radical results in the formation of hydro-peroxy radicals (HO_2) and organic-peroxy radicals (RO_2), which can alter the NO/NO_2 ratio. The presence of hydroxyl radical initiates the VOC oxidation process, followed by the formation of hydro- and organic peroxy radicals, which convert the NO to NO_2, which can form additional O_3, as well as converting HO_2 back to OH thus forming a catalytic cycle (HO_X catalytic cycle). However, ozone production is non-linear in relation to its precursors (NO_X and VOC) due to termination reactions that occur within the catalytic cycle (Lin et al., 1988; Nussbaumer and Cohen, 2020; Pusede and Cohen, 2012; Pusede et al., 2014). To that end, the response of ozone production is categorized into three regimes: NO_X-saturated (high NO_X with low VOC), NO_X-limited (low NO_X with high VOC), and transitional (Sillman et al., 1990; Sillman, 1999). In the NO_X-saturated regime (typically urban areas), ozone production is inversely proportional to NO_X concentration, whereas ozone production is directly proportional to VOC concentration. However, in NO_X-limited regimes (typically rural areas), ozone production is directly proportional to NO_X concentration, whereas VOCs have little effect on ozone production. This complex ozone production vs. precursor emission response is also evident in real-time observations, such as urban weekend ozone levels being higher than weekday levels (Sicard et al., 2020) and high ozone levels during public holidays and national shutdowns (e.g., the COVID-19 lockdown) due to low NO_X emissions (Balamurugan et al., 2021, 2022b).

Chemical transport models (CTMs) are commonly used to study air pollution and its drivers (Hu et al., 2016; Lou et al., 2015), but these models are dependent on emissions as represented in emission inventories (Pisoni et al., 2018). Emission inventories are typically developed using the bottom-up method, based on data such as economic activity, fuel consumption, traffic density, etc (McDuffie et al., 2020; Osses et al., 2022). However, bottom-up emission inventories can be highly uncertain due to inaccuracies in the data used in the bottom-up method, especially from unaccounted sources (Chen et al., 2020; Crippa et al., 2019; Trombetti et al., 2018). Because of the significant computational effort and storage space requirements, CTMs often perform at coarse spatial resolution, making it unable to solve fine transport and chemical mechanisms, particularly over complex topography (Singh et al., 2021). Machine learning (ML) models have been shown to be an effective complement to these computationally expensive CTMs (Vlasenko et al., 2021). The performance of machine learning models for modeling air pollutants is promising (Balamurugan et al., 2022a; Cheng et al., 2022; Lee et al., 2020; Li et al., 2022; Liang et al., 2020; Liu et al., 2022; Zaini et al., 2022; Zhao et al., 2023). Meteorological variables such as solar radiation and temperature have been shown to be important parameters in near-surface ozone modeling using machine learning (Diao et al., 2021; Hu et al., 2021). Meteorological conditions influence the concentration of O_3 both directly and indirectly. Solar UV radiation is responsible for the photolysis of O_3 precursors (NO_2 and VOCs). Temperature directly influences the photochemical reaction rate. Furthermore, meteorology influences biogenic and fuel-leak-related VOC emissions (exponentially proportional to temperature), which account for a significant portion of total VOC emissions (Guenther et al., 1993). In addition to meteorology, when emission source information is included, ML models predict near-surface NO_2 very well (Ghahremanloo et al., 2021; De Hoogh et al., 2019).

In-situ air quality measurements are sparse and concentrated primarily in urban areas. Recent advancements in satellite remote sensing allow us to analyze urban and non-urban air quality with adequate spatial and temporal coverage; however, they typically only measure the total or tropospheric column of specific air quality species, making it difficult to interpret people's exposure to near-surface air pollutants concentration. Therefore, in this study, we trained two ML models for near-surface NO_2 and O_3 concentrations over Germany using available information on proxies for near-surface air pollutants (satellite column measurements) and emission sources, precursors of air pollutants, as well as meteorology. Many recent studies, similar to ours, have attempted to model near-surface NO_2 and O_3 concentrations at national/regional spans ~~(De Hoogh et al., 2019; Kang et al., 2021; Li et al., 2020; Zhu et al., 2022)~~(De Hoogh et al., 2019; Kang et al., 2021; Kim et al., 2021; Li ; there are, however, very few attempts over Germany. To the best of the authors' knowledge, only one study (Chan et al., 2021) used TROPOMI satellite NO_2 tropospheric column measurements and other auxiliary information (e.g., meteorology) to model near-surface NO_2 concentrations over Germany using a MLP model. Furthermore, previous studies have focused on a single pollutant (e.g., NO_2), whereas in this study, we model and analyze the spatio-temporal variations in both NO_2 and O_3, which are chemically strongly coupled. In terms of anthropogenic emissions, we also evaluate the ML model performance of NO_2 and O_3 during the 2020 COVID-19 lockdown period, which serves as a natural experiment period with significantly lower primary anthropogenic emissions (Gensheimer et al., 2021).

2 Study region, Data sets, Model, and Method

All data sets used in this study, as well as their spatial and temporal resolutions, are summarized in Table 1.

2.1 Study region and near-surface NO_2 and O_3 measurements

We focused on the spatial domain of 5-15°E and 47-55.5°N, particularly over Germany. Near-surface NO_2 and O_3 data from measurement stations across Germany were used in this study. However, not all measuring stations collect data on both pollutants; there are less stations measuring O_3 than those measuring NO_2. There were also temporal gaps in the measurement data. Therefore, we only considered stations that had more than 80% data coverage during the study period. In the end, we

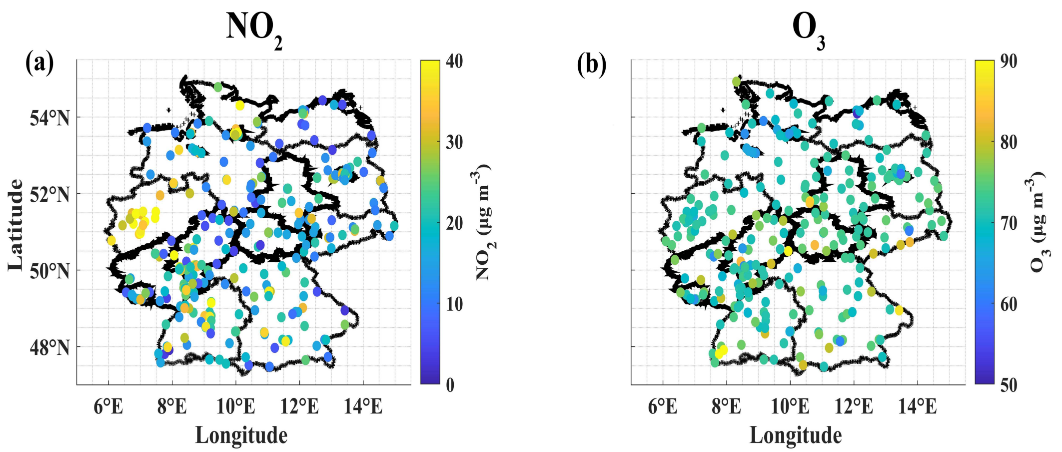

Figure 1. Locations of near-surface NO_2 (a) and O_3 (b) measurement stations considered in this study. The color bar depicts the mean of near-surface NO_2 and O_3 for each measurement station during the study period.

considered 321 stations for modeling NO_2 and 256 stations for modeling O_3. The selected measurement stations are located throughout the entire country and are situated in high-traffic, industrial, and background locations (Fig. 1 & Table A1).

2.2 Predictor variables of ML model

Predictor variables or input features for the ML models include satellite column measurements of air pollutants, meteorology and auxiliary data containing information on the area of interest.

2.2.1 Satellite column measurements

Tropospheric column NO_2, total column O_3, and troposheric column HCHO data are used, which are level-2 retrieval products from TROPOMI (TROPOspheric Monitoring Instrument) aboard the Sentinel-5P satellite. Sentinel-5P overpasses the study area between 13:00 and 14:00 local standard time. The spatial resolution of TROPOMI data is 7*3.5 km (increased to 5.5*3.5 km after August 6, 2019). We applied the data quality filtering described in the product manual to each data product (S5P (2022b) for NO_2, S5P (2022c) for O_3, and S5P (2022a) for HCHO). Tropospheric column NO_2 is used in the NO_2 ML model because it can be considered as proxy for near-surface NO_2. Since NO_2 is the precursor for O_3, we also included the tropospheric column NO_2 in the O_3 ML model. Because formaldehyde (HCHO) is an intermediate gas-product of VOC oxidation, it can be used as a proxy for VOC-oxidation (Jin et al., 2017). Therefore, we included tropospheric column HCHO in the O_3 model. We also considered the "TROPOMI FNR" (ratio of "TROPOMI HCHO" and "TROPOMI NO_2") in the O_3 ML model, which in previous studies has been shown to be a useful indicator of ozone production regime (Jin et al., 2020; Wang et al., 2021). We included total column O_3 in the O_3 ML model by considering total column O_3 as a proxy for near-surface O_3.

Table 1. Data sets and related information used in this study.

Data source	Data (purpose)	Temporal resolution	Spatial resolution
Governmental in situ measurements	Near-surface NO_2 and O_3 (Ground-truth data)	1 hr	-
TROPOMI satellite measurements	Tropospheric column NO_2, total column O_3 and total column HCHO (Input features)	Daily	7 km*3.5 km (5.5 km*3.5 km, after 6 August 2019)
ERA5 (ECMWF reanalysis)	Temperature, relative humidity, wind speed, wind direction, downwind UV solar radiation at surface, boundary layer height, surface pressure and temperature of air at 2m above the surface (Input features)	1 hr	0.25*0.25-degree
U.S. Geological Survey	Surface elevation (Input features)	-	1*1-km
GRIP global roads database	Road density (Input features)	-	8*8-km
CAMS European air quality forecasts	Near-surface NO_2 and O_3 (for validation)	1 hr	0.1*0.1-degree
GEOS-Chem chemical transport model	Near-surface NO_2 and O_3 (for disentangling meteorology impacts)	1 hr	0.5*0.625-degree

2.2.2 Vegetation index

Normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI) data were obtained from MODIS (Moderate Resolution Imaging Spectroradiometer) measurements aboard the Terra and Aqua satellites. We used the “MOD13A2 (16-day 1-km) VI” data set, which contains NDVI and EVI data at 1 km spatial resolution and 16 day temporal resolution. To generate daily intervals, the NDVI and EVI data were linearly interpolated. We considered these vegetation indexes in the O_3 ML model because vegetation contributes a considerable amount of VOCs. We also considered these vegetation indexes in the NO_2 ML model as a supplementary information to check whether changes in vegetation cover has any implications on NO_2 concentration changes.

2.2.3 Meteorology

Meteorology has both direct and indirect effects (e.g., dispersion, photochemical reactions) on pollutant concentrations. Meteorological variables such as temperature (T), relative humidity (RH), wind speed (WS), and wind direction (WD) were obtained from the ERA-5 reanalysis product. These variables were derived from the lowest model level (1000 hPa) of the “ERA-5 hourly data on pressure levels” data set. Downward UV solar radiation at the surface (DUV), boundary layer height (BLH), surface pressure (SP) and temperature of the air at 2 m above the surface (T2m) were derived from the “ERA-5 hourly data on single levels” data set. These meteorological data have a spatial resolution of 0.25 degree and a temporal resolution of one hour. In both the NO_2 and O_3 ML models, we took all meteorology variables into account.

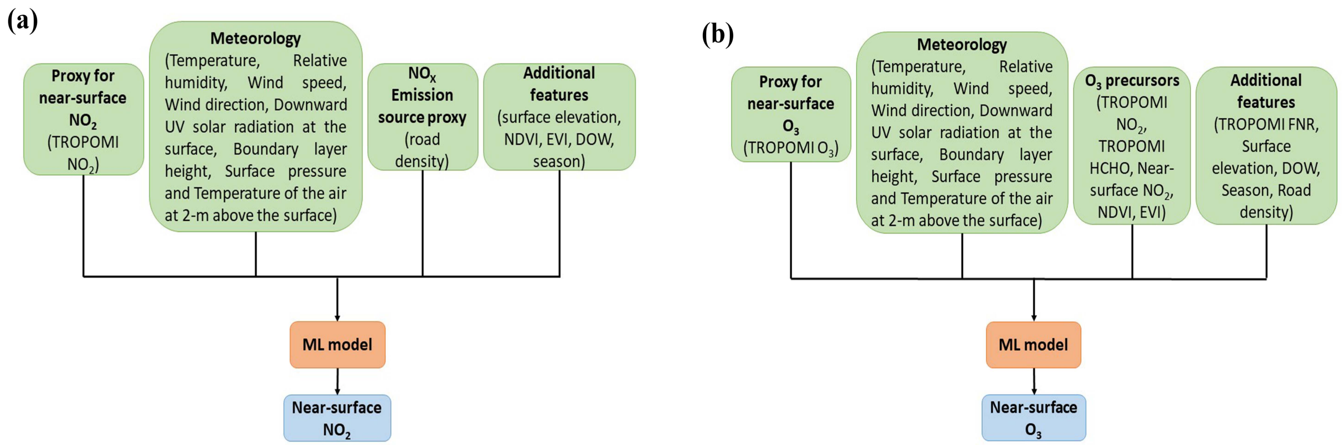

Figure 2. Predictor variables and data flow for the NO_2 (a) and O_3 (b) ML model.

2.2.4 Proxy for NO_X emission source

Because vehicle (transport sector) emissions are a significant source of NO_X emissions, considering a proxy for vehicle emissions is crucial. Therefore, we used road density as a proxy for the source of NO_X emissions. We are aware that traffic volume or density would be the ideal proxy, but data on traffic volume or density on a national/regional span is not available. The road density (RD) data was obtained from the GRIP global roads database, with a spatial resolution of 8 km.

2.2.5 Additional features

Additional supplementary data such as surface elevation (E) was obtained from the U.S. Geological Survey (USGS), with a spatial resolution of 1 km. Surface elevation was taken into account because it influences the tropospheric/total column value of measurements. We also considered "DOW" (day of the week), and "season" (season of the year) information in both the NO_2 and O_3 models since both NO_2 and O_3 have distinct weekly and seasonal cycles. Because NO_2 is an important precursor to O_3, in addition to "TROPOMI NO_2", we also included "Near-surface NO_2" modeled from NO_2 ML model as a feature variable in the O_3 ML model.

2.3 Study period and data pre-processing

The study period was chosen to be between 2018-04-30 and 2021-07-01, which corresponds to the availability of TROPOMI data retrievals with the same processing version. Despite the fact that satellites pass over the study area between 13:00 and 14:00 local standard time, we found that the satellite data represents the daily mean of air pollutants well. Therefore, we considered the daily 24-hr mean for near-surface NO_2 and the daily maximum 8-hour mean (i.e. the mean of the 8 highest hourly values during a day) for near-surface O_3 as our variables of interest (dependent variables to model), as these are commonly used metrics in air quality research (Hoffmann et al., 2021).

Table 2. Evaluation metrics of our GBT model in different testing strategies.

		Random (1-fold)	Random (5-fold)	Time-leave-out (5-fold)	Location-leave-out (5-fold)
NO_2 GBT model	**R^2**	0.88	0.89±0.002	0.74±0.07	0.68±0.12
	RMSE (μg m^{-3})	4.77	4.65±0.034	6.77±0.7	8.67±1
O_3 GBT model	**R^2**	0.92	0.92±0.001	0.74±0.09	0.8±0.06
	RMSE (μg m^{-3})	8.53	9.36±0.068	13.2±1.1	12.45±1.3

Because each data set has a different spatial and temporal resolution, we re-sampled all of the data to the same spatial (0.1*0.1 degree) and temporal (daily) resolution. The 0.1 degree ($\approx$ 10 km) resolution was chosen because it corresponds to the resolution of the main features such as road density (spatial resolution of 8 km), TROPOMI satellite measurements (spatial resolution of 7*3.5 km), and concurrent high-resolution (0.1 degree) air quality forecasts from CAMS (Copernicus Atmosphere Monitoring Service). We computed the daily 24-hr mean for near-surface NO_2 and the daily maximum 8-hr mean for near-surface O_3 for each in-situ measurement station and then calculated the mean of all stations that fell within 0.1 degree grid. The mean of surface elevation, NDVI, EVI, TROPOMI (NO_2, HCHO, O_3), and road density for each day were then calculated for the corresponding 0.1 degree grids. The surface elevation and road density were assumed to be constant during the study period. The ERA-5 meteorology product was resampled to 0.1 degree resolution using the nearest-neighbor method and the 24-hr mean was computed.

2.4 Machine learning model and evaluation strategies

We primarily used the gradient boosted tree (GBT) machine learning algorithm, XGBoost (Chen and Guestrin, 2016), to model near-surface NO_2 and O_3 concentrations. The GBT algorithm is a gradient-boosted decision tree-based algorithm that is expected to outperform deep neural network-based algorithms for structured data (Lundberg et al., 2020). Furthermore, tree-based models are more interpretable and require less time to train than deep neural network algorithms. However, for comparison, we also used the multi-layer perceptron (MLP; neural network) algorithm (Gardner and Dorling, 1998). The GBT and MLP algorithms were implemented using "scikit-learn", a Python module (https://scikit-learn.org/stable/). When training the MLP model, we normalized the discrete feature variables between 0 and 1. The corresponding predictor variables and data flow for the NO_2 and O_3 ML model is shown in Fig. 2.

To evaluate the ML model, we used the R^2 (coefficient of determination) and RMSE (root-mean-square error) metrics. We split the available data into training (70% of the data) and testing (the remaining 30%). The training data set was used to iteratively vary the hyper-parameters (combinations) and select the best set of hyper-parameters using a 5-fold CV (cross-validation). The hyper parameters used in this study are shown in Table A2 and Table A3. We also evaluated the ML model using three different 5-fold CV testing strategies (random 5-fold CV, time-leave-out 5-fold CV, and location-leave-out 5-fold CV) with 100% of the data (Meyer et al., 2018). In the random 5-fold CV testing strategy, the data was randomly split into

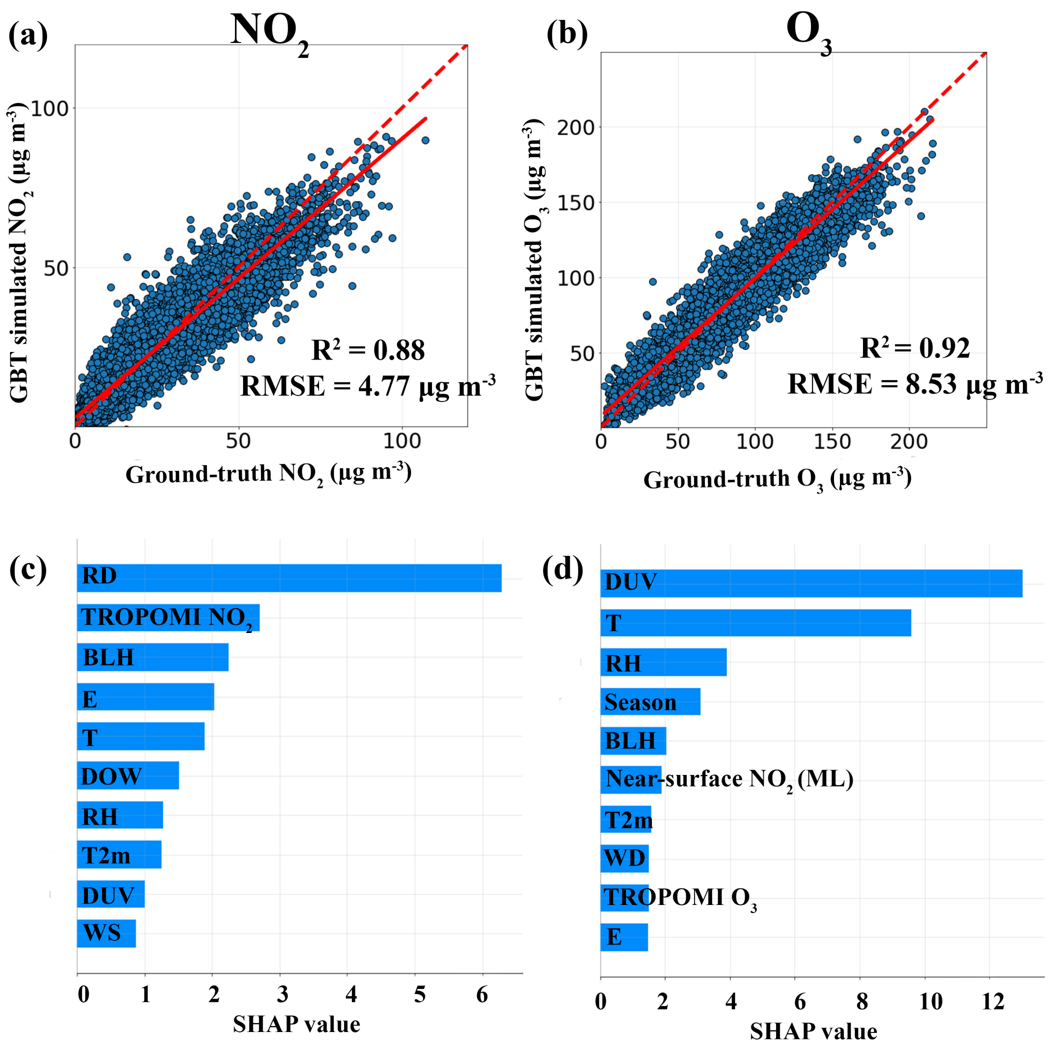

Figure 3. Comparison between ground-truth and GBT-simulated near-surface NO_2 (a) and O_3 (b). Feature importance (top 10) calculated based on SHAP (SHapley Additive exPlanations) values for NO_2 (c) and O_3 (d) GBT model. RD: Road Density, BLH: Boundary Layer Height, E: Surface Elevation, T-Temperature, DOW- Day of the week, RH-Relative Humidity, T2m: Temperature at 2 meter height, DUV: Downwind UV radiation, WS: Wind speed, WD: Wind Direction.

five parts, four of which were used for training and one for testing. This procedure was repeated until all five parts had been used as test. The mean (and standard deviation) of R^2 and RMSE from the 5-fold CV were then computed. In the time-leave-out 5-fold CV testing strategy, the 5-fold CV procedure was the same, but the data was split based on time period (by date; from the start of study period to the end of study period). Similarly, in the location-leave-out 5-fold CV testing strategy, the data was split based on location (by latitude). Figure A1 shows the first one-fold step in a 5-fold CV for time-leave-out and location-leave-out testing strategies. To interpret the importance of feature variables in the fitted model, we use SHAP (SHapley Additive exPlanations) values. The SHAP method (https://christophm.github.io/interpretable-ml-book/shap.html) is the most commonly used method for interpreting ML model output, which calculates the contribution of each feature variable to the final prediction. Thus, higher SHAP values indicate greater feature importance.

2.5 CAMS model data

We obtained near-surface NO_2 and O_3 air quality forecasts from CAMS in order to compare the performance of our ML model to that of the chemical transport model. This data set is based on a data-assimilation technique that combines real-time measurements with an ensemble of eleven air quality models to provide air quality data with high spatial resolution (0.1 degree) and 1 hr temporal resolution over Europe; however, it is only available for three years in the rolling archive. We used data from 2019-07-17 to 2020-01-31. We did not use data after 2020-01-31 due to COVID-19 lockdown restrictions, which limited many anthropogenic emission activities, and CAMS had not adjusted the emission inventory for changes in emissions. Furthermore, because NO_2 has a shorter lifetime, the effect of assimilated observations is minimal, and the CAMS forecasts NO_2 product mostly reflects emissions prescribed in the inventory (Inness et al., 2015).

2.6 GEOS-Chem model data

In this study, GEOS-Chem (GC) chemical transport model simulations were used to disentangle the meteorology contribution when estimating the influence of COVID-19 lockdown restrictions on air pollutant concentration changes. The GC simulations over the study area were obtained with a spatial resolution of 0.5×0.625 degree and 1-hr temporal resolution for the 2020 strict COVID-19 lockdown period (March 21 to May 31) and the same period in 2019. Identical anthropogenic emissions from the 2014 CEDS inventory were used for both 2020 and 2019, but with the corresponding meteorology, natural, and fire emissions in the respective years. Therefore, the difference in GC-simulated species (X) concentrations between 2020 and 2019 results from changes in meteorology, natural, and fire emissions between 2020 and 2019 ($GC\ X_{\ 2020-2019}$); here, X refers to either NO_2 or O_3. Then, we subtracted the $GC\ X_{\ 2020-2019}$ from the observed near-surface $X_{\ 2020-2019}$ to estimate the changes in concentrations of species X due to changes in anthropogenic emissions in the 2020 lockdown period (refer to studies Balamurugan et al. (2021); Qu et al. (2021) for the detailed description of the method).

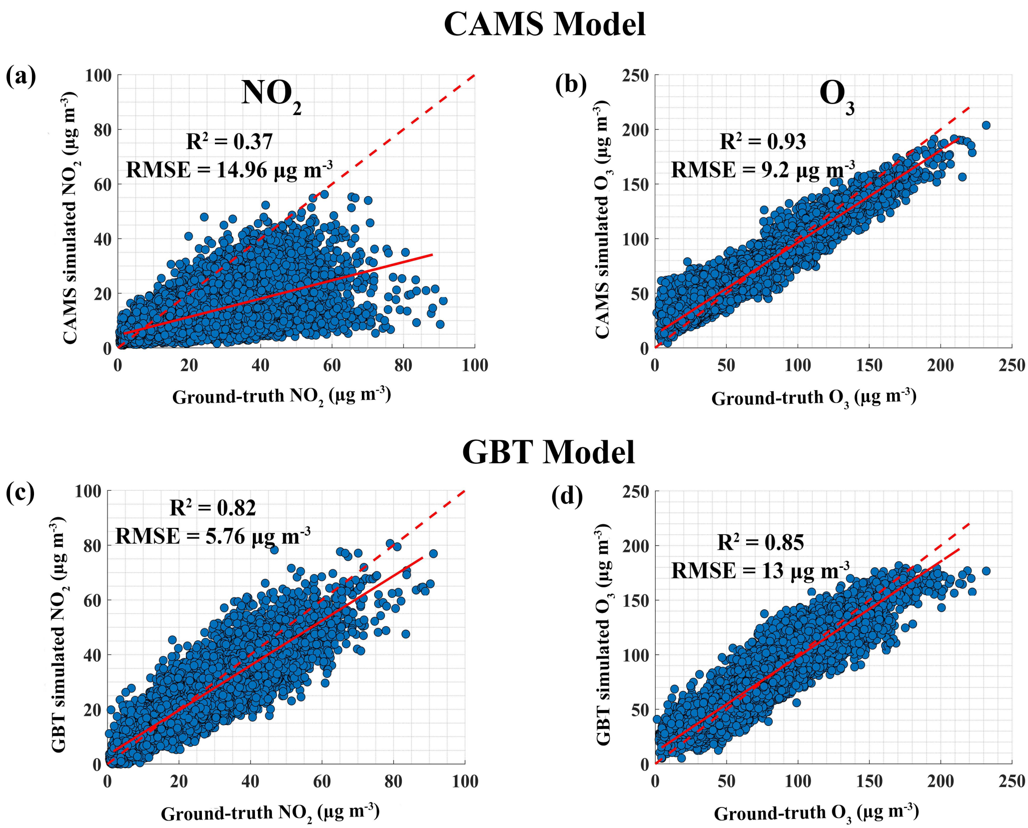

Figure 4. Top: Comparison between ground-truth near-surface NO_2 and CAMS forecasts near-surface NO_2 (a) and O_3 (b) for the period between 17-07-2019 and 31-01-2020. Bottom: Comparison between ground-truth near-surface NO_2 and GBT-simulated near-surface NO_2 (c) and O_3 (d) for the period between 17-07-2019 and 31-01-2020. The dotted line represents a 1:1 line, while the solid line represents a linear fit.

3 Results

3.1 ML model evaluation and feature importance

The trained GBT model with 70% of the data (78433) for NO_2 reproduced the observed NO_2 concentration well in the test case (33615), with an R^2 of 0.88 and RMSE of 4.77 μg m^{-3} (Fig. 3(a) and Table 2). The random 5-fold CV results were in the same range (R^2=0.89±0.002 and RMSE= 4.65±0.034 μg m^{-3}). The other two testing strategies (time-leave-out 5-fold CV and location-leave-out 5-fold CV) showed slightly worse agreement (Table 2), indicating that different validation strategies should be performed to interpret the ML model capability. Otherwise, it may result in an overoptimistic view of ML models (Meyer et al., 2018). Furthermore, the worse agreement in the location-leave-out 5-fold CV testing strategy suggests that there is less confidence in modeling the near-surface NO_2 over new locations that the GBT model has not been trained on before.

However, these results outperformed the MLP model trained by another study (Chan et al. (2021); R = 0.8 and RMSE = 6.32 μg m^{-3} obtained for the testing strategy of random split of 90% of data used for training and 10% of data used for testing) for near-surface NO_2 over Germany. Feature importance, based on the SHAP values, indicates that road density is the most important feature in the fitted model for NO_2 (Fig. 3(c)), because traffic is the main source of near-surface NO_X in urban areas. The next most important features were TROPOMI NO_2, boundary layer height, and elevation. Because the majority of NO_X sources are present at the surface, tropospheric column NO_2 data plays an important role in explaining near-surface NO_2. Near-surface NO_2 typically has a negative correlation with boundary layer height, as increasing BLH disperses more and vice versa (Balamurugan et al., 2021). Therefore, BLH is one of the most important features. It is unexpected that elevation was an important feature. The cause could be that the surface elevation varies greatly across Germany, influencing the total tropospheric column of NO_2 and thus serving as a link between the tropospheric column of NO_2 and near-surface NO_2. A previous study (Chan et al., 2021) also found that elevation was an important feature in the fitted MLP model for near-surface NO_2 over Germany.

The GBT model trained with 70% of the data (65705) for O_3 also well represented the observed O_3 concentrations in the test case (28160), with an R^2 of 0.92 and RMSE of 8.53 μg m^{-3} (Fig. 3(b)). Similar to the NO_2 GBT model findings, time-leave-out 5-fold CV and location-leave-out 5-fold CV testing strategies showed less agreement than the random 5-fold CV testing strategy (Table 2). In comparison to our NO_2 GBT model, our O_3 GBT model demonstrated greater confidence in modeling near-surface O_3 over locations the model was not trained on. According to SHAP values, the five most important features were DUV, T, RH, BLH, and season, with DUV having the greatest influence (Fig. 3(d)). Because ozone is formed in the atmosphere from the photolysis of NO_2, DUV plays a significant role in the fitted model that explains near-surface O_3. Temperature is the second most important feature, which is also not surprising as it drives biogenic VOC emissions (an important precursor to O_3). Previous studies also show similar findings (Diao et al., 2021; Hu et al., 2021). GBT-modeled near-surface NO_2 was the sixth most important feature in the fitted model, according to the SHAP values, and it was also more important than TROPOMI NO_2.

Figure A2 shows the results obtained from the MLP model. Both the NO_2 and O_3 MLP models performed worse than the NO_2 and O_3 GBT models, respectively (Table A4 vs. Table 2). In particular, MLP model findings showed low agreement in time-leave-out 5-fold CV and location-leave-out 5-fold CV testing strategies. This supports previous studies (Heaton, 2020; Lundberg et al., 2020) showing MLP model is unlikely to outperform tree-based models for tabular data. Because the GBT model outperforms the MLP model, we only considered the GBT model results in the following.

It is important to note that deep learning models are data-intensive, and their performance and generalization capabilities tend to improve with larger amounts of data. In our study, we utilized the simplest deep learning algorithm known as MLP. However, it is essential to explore the capabilities of other deep learning algorithms, such as CNN and LSTM, in future studies to gain further insights. Additionally, employing multiple ML models through bagging techniques could potentially lead to improved performance, despite the computational expense involved.

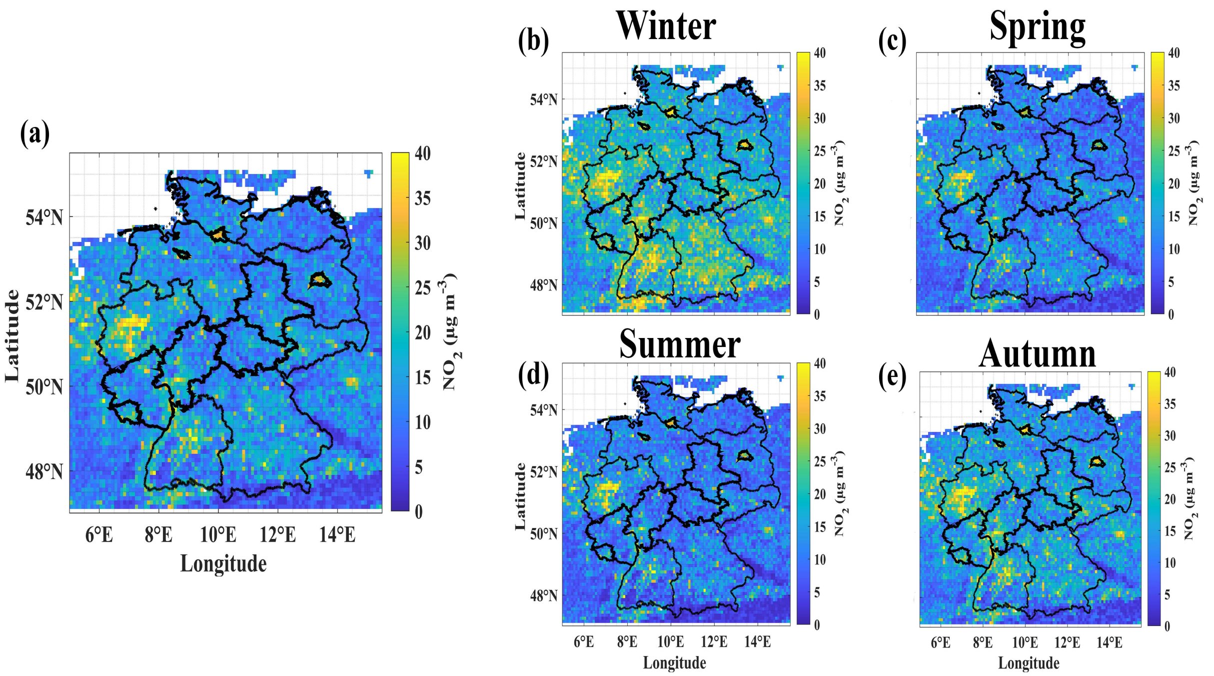

Figure 5. (a) Averaged GBT-simulated daily near-surface NO_2 concentrations over the study domain ~~during~~ for the study period between 2018-04-30 and 2021-07-01. (b-e) Averaged GBT-simulated daily near-surface NO_2 concentrations for each season during the study period. Winter: December, January and February. Spring: March, April and May. Summer: June, July and August. Autumn: September, October and November.

3.2 GBT model performance compared to CAMS

To evaluate how well our GBT model performs compared to CAMS, we compared the high-resolution near-surface NO_2 and O_3 forecasts from CAMS with observations, and GBT-simulated near-surface NO_2 and O_3 with observations, for the period between 2019-07-17 and 2020-01-31, i.e., CAMS comparison period, (Fig. 4). Please note this time period was not used for training the GBT model for this comparison. Our NO_2 GBT model reproduced the observed near-surface NO_2 concentrations well during this comparison period, with an R^2 of 0.82 and RMSE of 5.76 μg m^{-3}, while CAMS NO_2 forecasts showed poor representation (R^2 = 0.37 and RMSE = 14.96 μg m^{-3}). However, CAMS O_3 forecasts agreed slightly better with observed concentrations (R^2 = 0.93 and RMSE of 9.2 μg m^{-3}) compared to our O_3 GBT model (R^2 = 0.85 and RMSE = 13 μg m^{-3}). It should be noted that CAMS model forecasts were based on data assimilation techniques. Therefore, the CAMS models are expected to outperform our GBT models. However, our NO_2 GBT model outperforms CAMS, possibly because the effect of data assimilation is minimal in the CAMS forecasts product due to the short NO_2 lifetime.

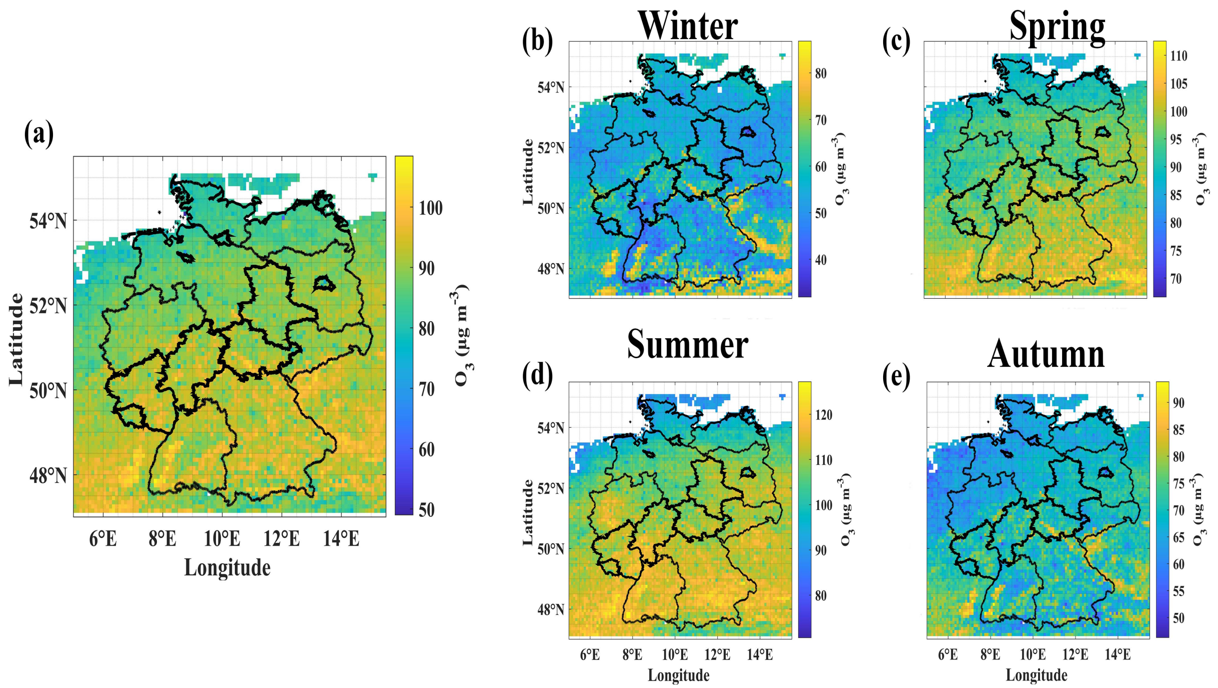

Figure 6. (a) Averaged GBT-simulated daily near-surface O_3 concentrations over the study domain ~~during~~ for the study period between 2018-04-30 and 2021-07-01. (b-e) Averaged GBT-simulated daily near-surface O_3 concentrations for each season during the study period. Winter: December, January and February. Spring: March, April and May. Summer: June, July and August. Autumn: September, October and November.

3.3 Spatio-temporal changes in near-surface NO_2 and O_3 over the study domain

After the discussed model evaluation, we trained the GBT model using 100% of the data and modeled the near-surface NO_2 and O_3 concentrations over the study domain at 0.1 degree resolution and daily (24-hr mean for NO_2 and 8-hr maximum mean for O_3) intervals. The averaged GBT-modeled near-surface NO_2 concentrations over the study domain during the study period are shown in Fig. 5(a). The spatial variability of near-surface NO_2 correlates with Germany's population density, and the main hotspots correspond to Germany's major metropolitan areas (Figure A3). The study domain's main hotspot is western Germany (North Rhine-Westphalia; a federal state of Germany), Germany's industrial heartland. The number of days (%) that exceeded the 2021 WHO NO_2 limit (24-hr mean > 25 $\mu g\ m^{-3}$) over major metropolitan areas in Germany was more than 50%, with western Germany exceeding the WHO NO_2 limit on more than 80% of the days during the study period (Fig. 7). Around 36% of people live in locations where more than 25% of days exceed the WHO NO_2 limit during the study period (Fig. 8). The GBT-simulated near-surface O_3 showed distinct spatial variability compared to NO_2, with high O_3 concentrations over southern Germany and low O_3 concentrations over northern Germany (Fig. 6). This could be due to the fact that O_3 is a secondary pollutant that is primarily driven by photochemical reactions influenced by meteorology; DUV and temperature values, which were the most influencing factors for photochemical reactions and accordingly the most important features fitted in the O_3 GBT model, were higher in southern Germany than northern Germany (Figure A4). During the study period, more than 50% of days in southern Germany exceeded the 2021 WHO O_3 limit (maximum 8-hr mean > 100 μg m^{-3}). Nearly 90% of people live in locations where more than 25% of days exceed the WHO O_3 limit (Fig. 8). Another interesting fact is that southern metropolitan areas and high NO_X regions have less days that exceeded the WHO O_3 limit than southern rural regions (Fig. 7). It is a well-known fact that rural regions have higher ozone levels than urban regions (Malashock et al., 2022). It could be because NO is a significant O_3 scavenger in higher NO_X (NO_2 is a proxy for NO_X) regions or due to being in a NO_X saturated regime. Furthermore, it is due to the fact that rural regions being the downwind locations of emission plume and are the primary source of biogenic VOC emissions (Zong et al., 2018).

We also evaluated the model capability in capturing the exceedance events (above WHO limit) using time-leave-out evaluation strategy. The exceedances of NO_2 and O_3 events simulated by GBT model compared with Ground-truth events in each iteration. This allows us to assess the model's ability to reproduce the exceedance events that have not been used in the training process. The 82% of the WHO NO_2 and O_3 exceedance events in the whole dataset (Ground-truth) were correctly identified as WHO NO_2 and O_3 exceedance events (True Positives) in both the NO_2 and O_3 GBT models (Table A5). However, we also noted that 6.6% and 7.3% of the data were incorrectly identified as exceedance events by our NO_2 and O_3 GBT models, respectively (False Positives). This indicates that our GBT model might slightly underestimate the exceedance events for both NO_2 and O_3. This could be due to unknown drivers that are not included in the model.

The GBT-simulated near-surface NO_2 showed seasonal variations, as expected, with higher values in the winter season (Fig. 5). This is because of high-residential heating demand and favorable meteorology (e.g., a low boundary layer height) for pollutant accumulation and less NO_2 photolysis due to low solar radiation in the winter. The near-surface NO_2 hotspots were the same in all seasons, as seen in the overall study period average. In contrast, near-surface O_3 showed strong seasonal variations, with high values in the spring and summer due to high solar radiation (Fig. 6). It is worth noting that, as seen in the overall study period average, O_3 values in southern Germany were significantly higher in spring and summer than in northern Germany. Because near-surface O_3 is mainly driven by meteorology (DUV and temperature, which drive photochemical reactions and precursor emissions), the spatial and temporal variability is attributed to changes in meteorology. We also compared the spatial variability of GBT-simulated near-surface NO_2 and O_3 to the CAMS forecasts product for the period between 2019-07-17 and 2020-01-31 (Figure A5 and A6). The spatial variability of GBT-simulated near-surface NO_2 and O_3 agreed well with CAMS model. This implies that the ML model can supplement or replace the computationally expensive chemical transport models.

3.4 Influence of COVID-19 lockdown restrictions on near-surface NO_2 and O_3 changes

Due to the COVID-19 out-break, many nations, including Germany, announced a lockdown in the spring of 2020. During that time period, various anthropogenic emission activities were restricted, affecting particularly traffic-related emissions. To estimate the influence the lockdown restrictions on air pollutant concentration changes, we compared the GBT-simulated 2020

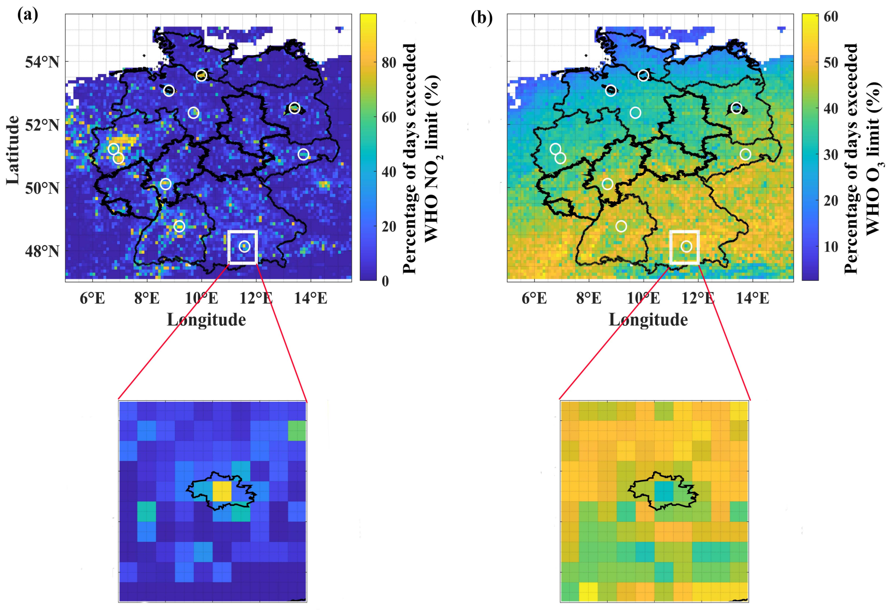

Figure 7. Number of days (%) that exceeded the WHO 24-hr mean NO_2 (a) and maximum 8-hr mean O_3 (b) limits over the study domain during the study period based on GBT-model simulations. White circles represent major metropolitan areas. The metropolitan area of Munich and its surroundings (rectangular box) are enlarged to illustrate the urban vs. rural gradient. The administrative boundaries of Munich are marked in black in the inset panel.

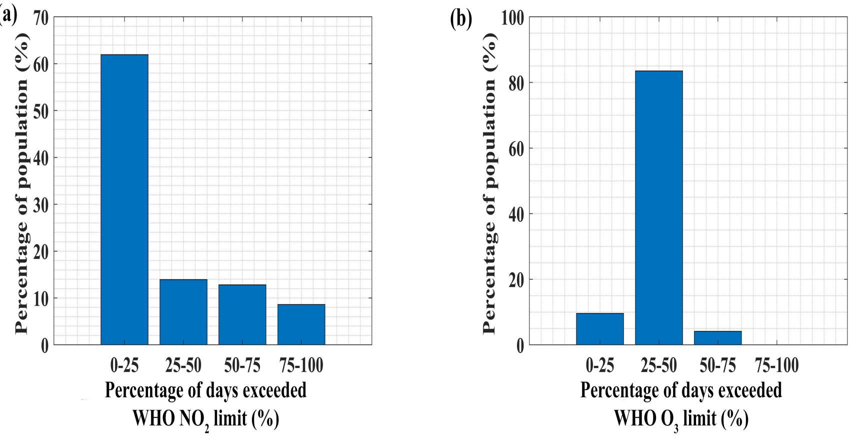

Figure 8. The population distribution in terms of the number of days (%) that exceeded the WHO 24-hr mean NO_2 (a) and maximum 8-hr mean O_3 (b) limits over the study domain during the study period based on GBT-model simulations.

lockdown concentration with the same period in 2019. The 2020 lockdown period measurements were not used for GBT model training in this comparison. This can also be regarded as the critical performance evaluation of the GBT model.

When comparing different time periods, it is crucial to account for meteorological effects when estimating the impact of anthropogenic emission reductions (i.e., lockdown effects) on changes in air pollutant concentrations. Therefore, as described in the method section, we used GC simulations to exclude the meteorology contribution from GBT-simulated concentrations. After disentangling the meteorology contribution, it is noticeable that high near-surface NO_2 levels decreased primarily over the previously observed hotspots (Fig. 9). The near-surface O_3 increased over western Germany while decreasing elsewhere, particularly over low NO_X regions. We already observed that western Germany was a NO_X hotspot, possibly a NO_X saturated regime, so a reduction in NO_X increases ozone. Also, we could see that changes in near-surface O_3 were either negligible or slightly increased over metropolitan areas. The meteorology-accounted for mean lockdown near-surface NO_2 decreased by about 23 ($\pm$5.3)%, while meteorology-accounted for mean lockdown near-surface O_3 increased by 1 ($\pm$4.6)%, over ten major metropolitan areas (Berlin, Bremen, Cologne, Dresden, Düsseldorf, Frankfurt, Hamburg, Hanover, Munich, and Stuttgart), compared to 2019. It increased by about 9% in the Cologne and Düsseldorf metropolitan areas (located in western Germany) and slightly increased or decreased (between -3 and +2%) in other metropolitan areas, compared to 2019. This finding is consistent with other studies that found a decrease in meteorology-accounted for lockdown near-surface NO_2 and the small increase in lockdown near-surface O_3 over German metropolitan areas compared to 2019 using in-situ measurements (Balamurugan et al., 2021, 2022b). We also evaluated our GBT model's ability to represent different emission scenarios by comparing weekends and weekdays; typically, anthropogenic NO_X emissions on weekends are lower than on weekdays due to reduced vehicle transportation. Our GBT model was also able to distinguish between the weekend and weekday emission scenarios; weekend near-surface NO_2 was lower than weekday near-surface NO_2, and, as expected, there were no or only slight changes in weekend near-surface O_3 compared to weekdays, with slight increases particularly over metropolitan areas (Figure A7).

3.5 Transferability of our GBT model

Our study domain also covered parts of other European countries. However, we trained our GBT model using data from German measurement stations only. Therefore, comparing our trained GBT model simulations with measurements in other countries demonstrates how well our GBT model models near-surface NO_2 and O_3 concentrations in neighboring parts of the world; similar to the location-leave-out testing strategy. We chose five major cities (Salzburg, Prague, Strasbourg, Liège, and Groningen) in different European countries covered by our study domain and compared their measured NO_2 and O_3 concentrations with GBT modeled NO_2 and O_3 concentrations (Fig. 10 & Table ~~A5~~A6).

Our trained NO_2 GBT model based on German measurement stations explained 32-64% (R^2 ranges between 0.32 and 0.64, and RMSE ranges between 9.76 and 13 μg m^{-3}) of near-surface NO_2 measured in five metropolitan areas located outside of Germany, while O_3 GBT model simulations agreed well with observations (R^2 ranges between 0.87 and 0.94, and RMSE ranges between 9.55 and 14.32 μg m^{-3}). Since near-surface O_3 is mainly driven by meteorology, the O_3 GBT model trained using German measurement stations explains a large portion of near-surface O_3 in other locations. The worse agreement between NO_2 GBT model predictions and NO_2 observations in other European countries suggests that information is lacking

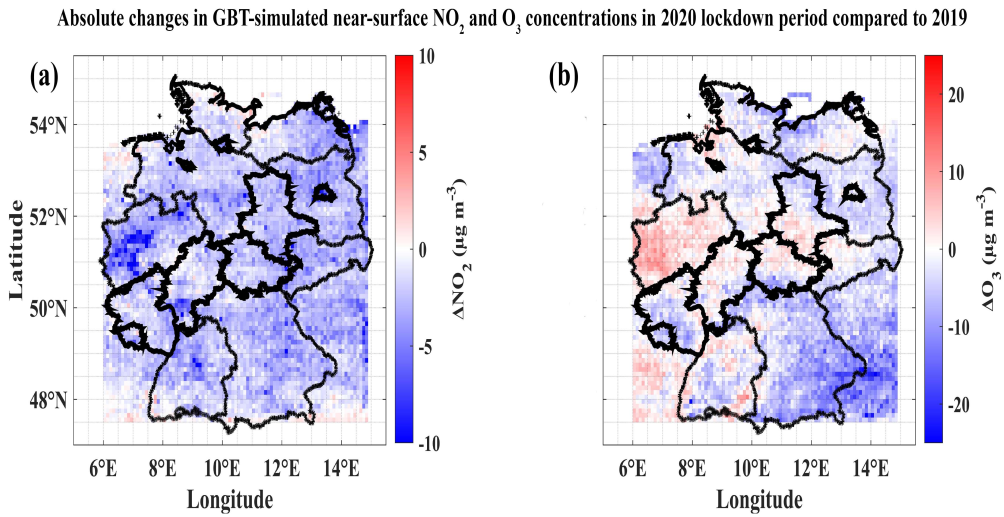

Figure 9. Absolute changes in GBT-simulated near-surface NO_2 and O_3 concentrations in 2020 lockdown period compared to the same period in 2019 after accounting for meteorology.

in the NO_2 GBT model for better representation of other locations, similar to location-leave-out 5-fold CV, which also showed low agreement for the NO_2 GBT model when modeling new locations (Table 2). Differences in vehicle fleet composition and emission standards across different countries/locations would have an impact on our NO_2 GBT model predictions when applied to other countries/locations. In future work, other features/proxies besides road density could be considered to represent traffic emission.

4 Conclusion

This study simulated near-surface NO_2 and O_3 concentrations using an ML model over Germany at 0.1 degree resolution and daily intervals. The ML model was used to link satellite column measurements (proxies for near-surface air pollutants), meteorology and proxies of emission source information to near-surface NO_2 and O_3 concentrations. The ML models are extremely effective at learning the complex non-linear relationships between variables. Therefore, in this study, we explored the capabilities of ML models in the spatio-temporal prediction of air pollutants. In addition, we investigated three aspects of the ML model: 1. how well our ML model performs compared to the chemical transport model, 2. how well our ML model can be used to assess the effectiveness of mitigation initiatives; and 3. how well our ML model can be transferred to locations where measurements are unavailable.

Four different testing strategies were performed to evaluate the ML model's spatio-temporal prediction: 1. Random split of data (70% for training and 30% for testing), 2. Random 5-fold CV, 3. Time-leave-out 5-fold CV, and 4. Location-leave-out 5-fold CV. The gradient boosted tree (GBT) model trained for NO_2 explained about 68-88% of observed NO_2 concentrations

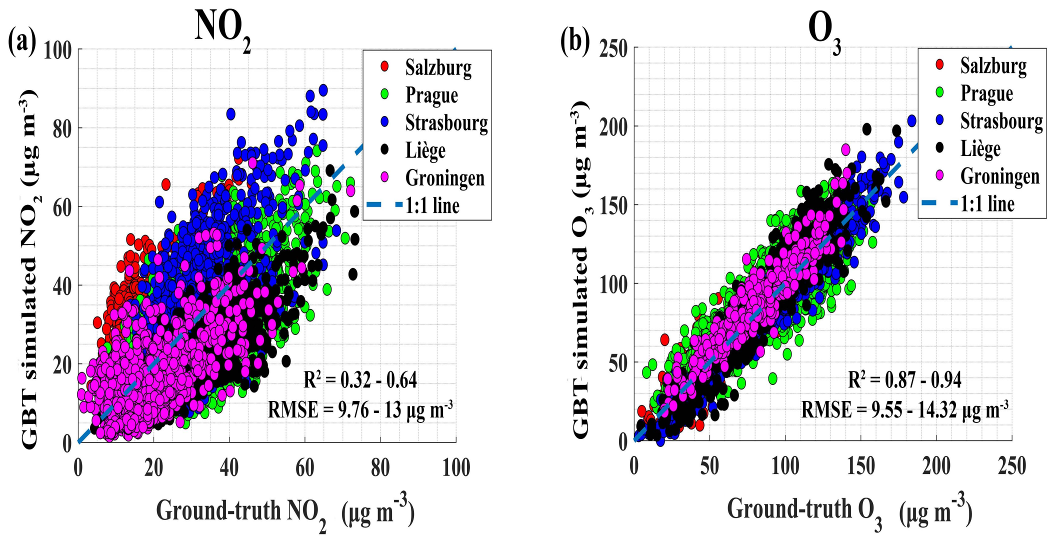

Figure 10. Comparison between ground-truth and GBT-simulated near-surface NO_2 (a) and O_3 (b) for five different European metropolitan areas.

in Germany, with RMSE of 4.77-8.67 μg m^{-3}, whereas the GBT model trained for O_3 performed even better, with an R^2 of 0.74-0.92 and RMSE of 8.53-13.2 μg m^{-3}. The evaluation metrics of the GBT model for different testing strategies differed significantly. The location-leave-out 5-fold CV testing strategy showed poor agreement for the NO_2 GBT model, whereas the time-leave-out 5-fold CV testing strategy showed poor agreement for the O_3 GBT model. This points out the importance of performing different testing strategies to interpret the true capability of the ML model. The road NO_X emission source proxy (road density) and TROPOMI tropospheric column NO_2 were the most important features in the fitted NO_2 GBT model. However, for O_3, the most important features were downward UV radiation at the surface and temperature. The multi-layer perceptron (MLP) model trained for both NO_2 and O_3 performed worse than the GBT model.

We also showed that our NO_2 GBT model outperforms the CAMS model, while slightly under-performing for near-surface O_3. The CAMS model forecasts data set uses real-time observations with an ensemble of eleven air-quality models through data assimilation techniques, which are expected to be more computationally expensive than our GBT model. Therefore, the spatio-temporal variability of near-surface NO_2 and O_3 concentrations and human exposure at a locations where no measurements are available can be studied with lower computational effort when using our GBT model. Near-surface NO_2 hotspots were found over German metropolitan areas, particularly western Germany. The near-surface NO_2 hotspots locations did not change with the seasons but had high values in the winter. However, near-surface O_3 showed high seasonal variability, with high values in the spring and summer and no definite hotspots. Overall, southern Germany experiences higher ozone levels than northern Germany due to higher downward UV radiation and temperatures in southern Germany compared to northern Germany. Even though metropolitan areas were the NO_2 hotspots, rural regions, particularly in southern Germany, had higher O_3 concentrations than metropolitan areas. It is because rural areas are dominated by meteorology-driven biogenic VOC emissions and are generally situated downwind of the emission plume. About 36% of people live in locations where WHO NO_2 limit exceeds more than 25% of days during the study period. Meanwhile, 90% of the people lives in areas where the WHO O_3 limit is exceeded for more than 25% of days.

Our study also demonstrated the GBT model's capability to assess the efficacy of mitigation strategies. For example, our GBT model reproduced the observations that, during the 2020 COVID-19 lockdown period, meteorology-accounted for near-surface NO_2 was significantly reduced, while meteorology-accounted for near-surface O_3 was slightly increased or decreased over metropolitan and industrial areas over Germany, compared to 2019. These findings agreed with those of other studies that used in-situ measurements.

Our GBT ML model's transferability is assessed by comparing simulations from our GBT model trained with measurements in Germany to measurements in other European countries. Our NO_2 GBT model showed moderate agreement with observations from other countries (R^2 ranges between 0.32 and 0.64, and RMSE ranges between 9.76 and 13 μg m^{-3}), implying a lack of information in the GBT model when modeling near-surface NO_2 over other countries, which may have different vehicle fleet composition and emissions standards. However, our O_3 GBT model performed well (R^2 ranges between 0.87 and 0.94, and RMSE ranges between 9.55 and 14.32 μg m^{-3}), indicating that our O_3 GBT model can be used to model the O_3 concentrations in other countries, at least in neighboring European countries.

Code and data availability. The various data sets and code used to conduct this study will be made available on GitHub following publication.

Appendix A

Table A1. Different type of stations (%) considered in this study (based on locations specified by the European Environment Agency).

	Traffic	**Industrial**	**Background**
Near-surface NO_2	37.1%	5.3%	57.6%
Near-surface O_3	2.7%	5.8%	91.4%

Table A2. The hyperparameters of the GBT model for each pollutant used in the study.

Hyper paramertes	**NO_2 model**	**O_3 model**
Max_depth	10	10
Learning_rate	0.3	0.3
reg_lambda	12	4
reg_alpha	18	26
gamma	20	8
min_child_weight	16	8
n_estimators	2500	2500

Table A3. The hyperparameters of the MLP model for each pollutant used in the study.

Hyper paramertes	**NO_2 model**	**O_3 model**
Hiddern_layers (neurons in each layer)	3 (200,100,50)	4 (350,150,75,37)
activation	tanh	tanh
alpha	0.04	0.1
learning rate	adaptive	adaptive
solver	sgd	lbfgs
Max_iter	2000	1500

Table A4. Evaluation metrics of our MLP model in different testing strategies.

		Random (70%/30%)	Random (5-fold)	Time-leave-out (5-fold)	Location-leave-out (5-fold)
NO_2 MLP model	R^2	0.79	0.82±0.006	0.54±0.29	0.46±0.25
	RMSE (μg m^{-3})	4.77	5.9±0.11	8.6±1.76	13.2±1.07
O_3 MLP model	R^2	0.83	0.9±0.001	0.42±0.37	0.71±0.13
	RMSE (μg m^{-3})	12.15	9.6±0.027	20.1±7.3	14.9±3.2

Table A5. Comparison between WHO NO_2 and O_3 exceedance events in the ground-truth dataset and GBT simulated WHO NO_2 and O_3 exceedance events using time-leave-out testing strategy

	Ground-truth exceedance	Correct detection as exceedance by NO_2 GBT model (True Positives)	Correct detection as exceedance by O_3 GBT model (False Positives)
Near-surface NO_2	36772	30125	7439
Near-surface O_3	35860	29396	6924

Table A6. Metropolitan areas in other European cities considered for the evaluation of GBT model. The evaluation metrics (comparison between GBT simulations and in-situ measurements) for NO_2 and O_3 shown in last two columns for each city.

Metropolitan area (country)	Coordinates	R^2 and RMSE (μg m^{-3}) for NO_2	R^2 and RMSE (μg m^{-3}) for O_3
Salzburg (Austria)	47.80° N, 13.05° E	0.32 and 12.52	0.87 and 12.43
Prague (Czech Republic)	50.07° N, 14.43° E	0.43 and 10.05	0.79 and 14.32
Strasbourg (France)	48.57° N, 7.75° E	0.47 and 13	0.94 and 9.55
Liège (Belgium)	50.63° N, 5.56° E	0.64 and 11.9	0.88 and 12.04
Groningen (Netherlands)	53.21° N, 6.56° E	0.34 and 9.76	0.87 and 11.33

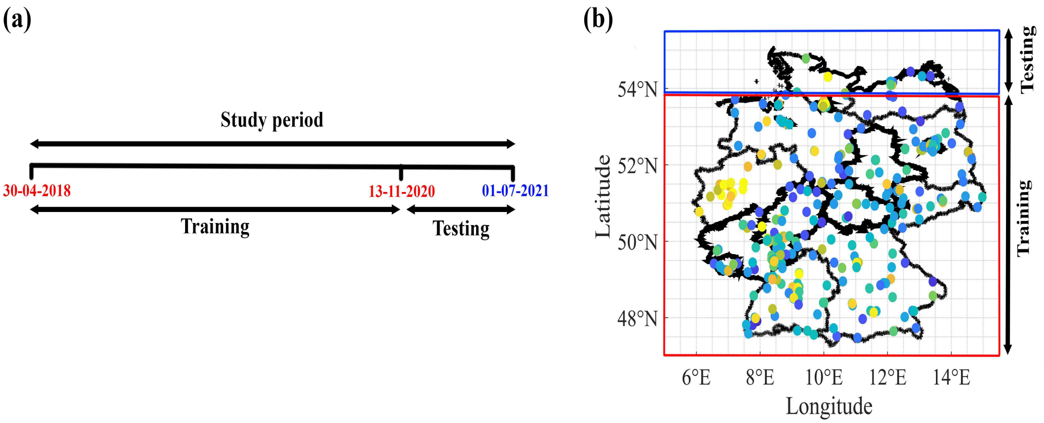

Figure A1. A first one-fold step in 5-fold CV is illustrated for time-leave-out (a) and location-leave-out (b) testing strategies. In time-leave-out 5-fold CV, the data was divided into 5 parts based on time period (date-wise), with four parts used for training and one part tested. This process is repeated until each part (a total of 5) has been tested. Similarly, in location-leave-out 5-fold CV, the data was divided into 5 parts based on location (latitude), with four parts used for training and one part tested. This process is repeated until each part (a total of 5) has been tested.

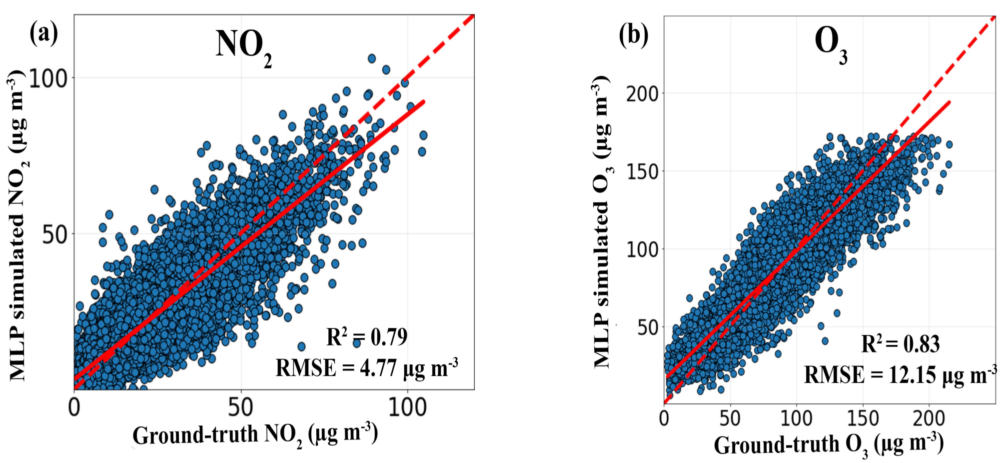

Figure A2. Comparison between ground-truth and MLP-simulated near-surface NO_2 (a) and O_3 (b). The dotted line represents a 1:1 line, while the solid line represents a linear fit.

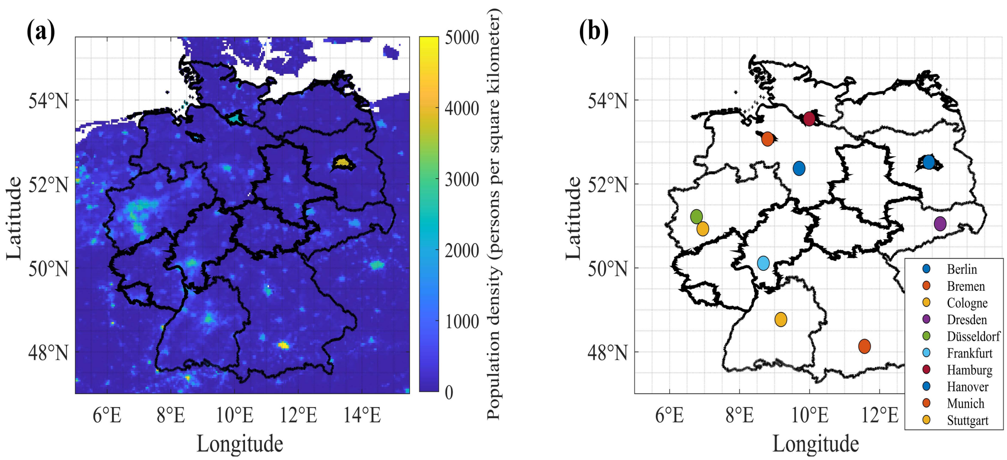

Figure A3. Population density for the year 2020 (a) and the locations of major German metropolitan areas (b).

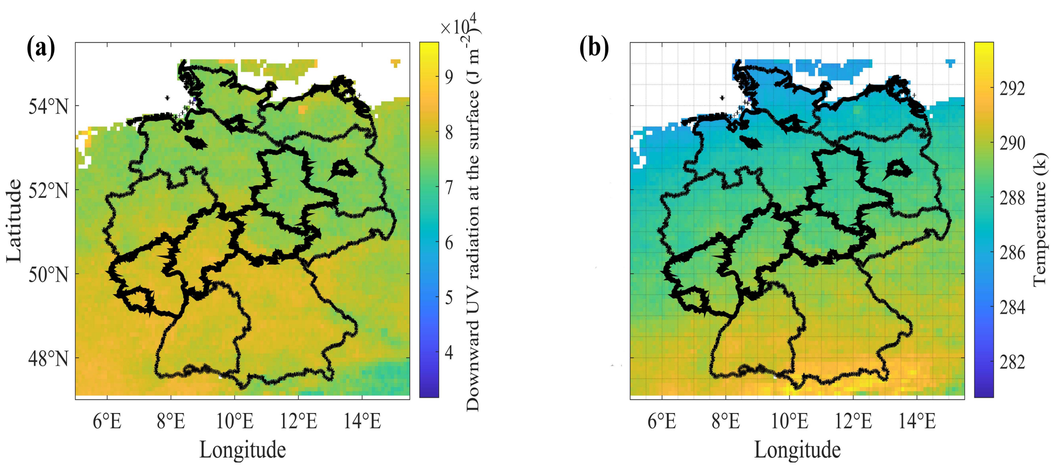

Figure A4. Averaged “Downward UV radiation at the surface” (a) and “Temperature” (b) over the study domain during the study period.

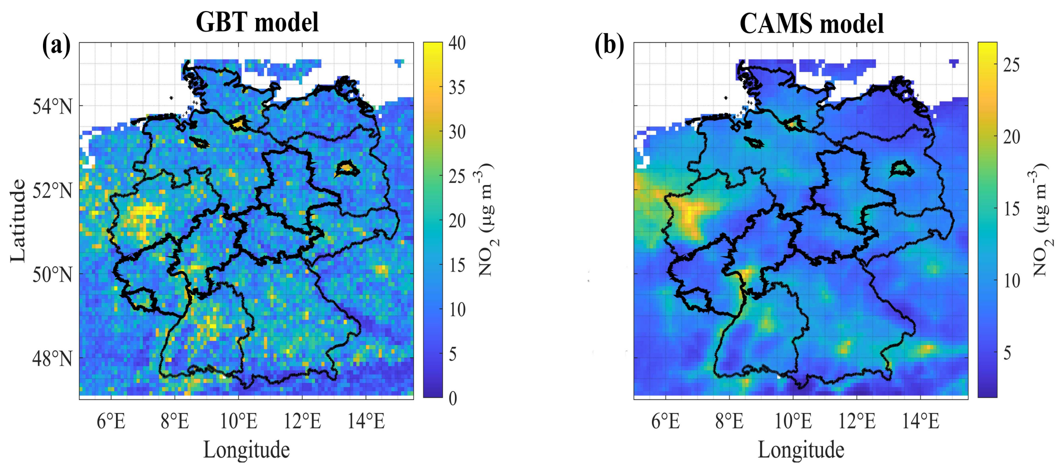

Figure A5. Averaged GBT-simulated near-surface NO_2 concentrations (a) and CAMS forecasts near-surface NO_2 concentrations (b) over the study domain for the period between 2019-07-17 and 2020-31-01.

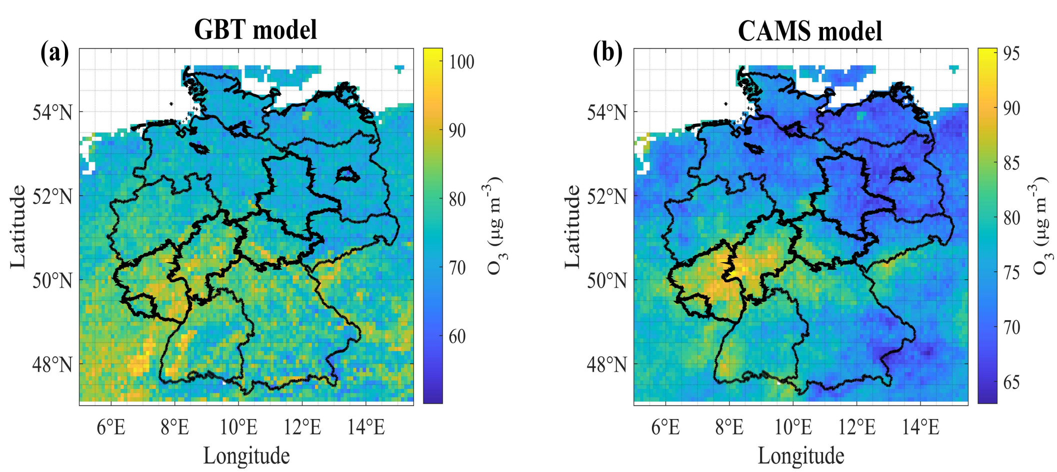

Figure A6. Averaged GBT-simulated near-surface O_3 concentrations (a) and CAMS forecasts near-surface O_3 concentrations (b) over the study domain for the period between 2019-07-17 and 2020-31-01.

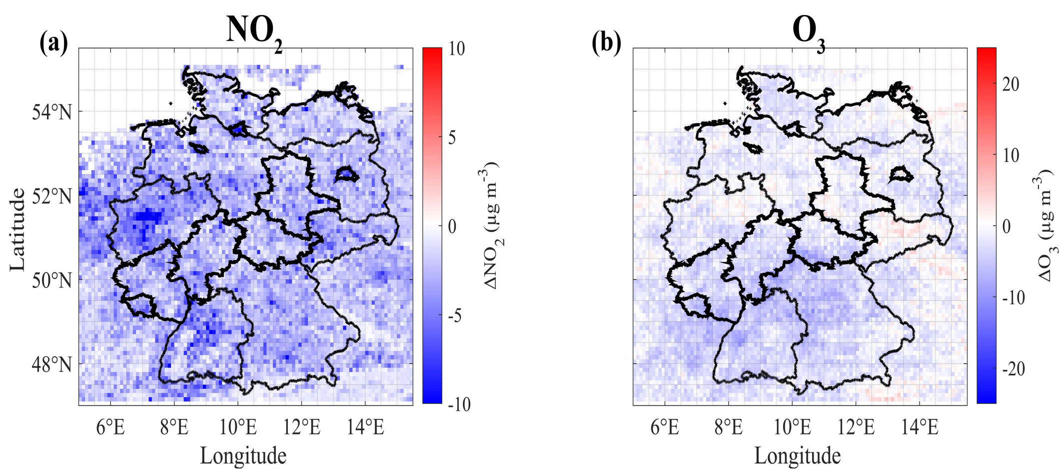

Figure A7. The difference in GBT-simulated near-surface NO_2 (a) and O_3 (b) concentrations between weekend and weekday during the study period.

Author contributions. VB, JC and FNK conceived the study and designed the concept. VB obtained all of the data, performed the modelling work and analysed the results. VB and AW developed the methodology. JC and FNK acquired the funding and supervised the work. VB wrote the manuscript. JC, AW and FNK reviewed and edited the manuscript

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This research has been funded by the Institute for Advanced Study, Technical University of Munich (grant no. 291763).

The authors thank the European Environment Agency, the Copernicus Services, the GES DISC data archive and the United States Geological Survey for providing free access to the various data sets used in this study.

References

S5P HCHO Readme, S5P Mission Performance Centre Formaldehyde [L2 HCHO] Readme https://sentinels.copernicus.eu/documents/247904/3541451/Sentinel-5P-Formaldehyde-Readme.pdf, 2022a.

S5P NO2 Readme, S5P Mission Performance Centre Nitrogen Dioxide [L2 NO2] Readme. https://sentinel.esa.int/documents/247904/3541451/Sentinel-5P-Nitrogen-Dioxide-Level-2-Product-Readme-File, 2022b.

S5P O3 Readme, S5P Mission Performance Centre Readme OFFL Total Ozone. https://sentinels.copernicus.eu/documents/247904/3541451/Sentinel-5P-Readme-OFFL-Total-Ozone.pdf, 2022c.

Balamurugan, V., Chen, J., Qu, Z., Bi, X., Gensheimer, J., Shekhar, A., Bhattacharjee, S., and Keutsch, F. N.: Tropospheric NO2 and O3 response to COVID-19 lockdown restrictions at the national and urban scales in Germany, Journal of Geophysical Research: Atmospheres, 126, e2021JD035 440, 2021.

Balamurugan, V., Balamurugan, V., and Chen, J.: Importance of ozone precursors information in modelling urban surface ozone variability using machine learning algorithm, Scientific reports, 12, 1–8, 2022a.

Balamurugan, V., Chen, J., Qu, Z., Bi, X., and Keutsch, F. N.: Secondary PM 2.5 decreases significantly less than NO 2 emission reductions during COVID lockdown in Germany, Atmospheric Chemistry and Physics, 22, 7105–7129, 2022b.

Bell, J., Power, S. A., Jarraud, N., Agrawal, M., and Davies, C.: The effects of air pollution on urban ecosystems and agriculture, International Journal of Sustainable Development & World Ecology, 18, 226–235, 2011.

Chan, K. L., Khorsandi, E., Liu, S., Baier, F., and Valks, P.: Estimation of surface NO2 concentrations over Germany from TROPOMI satellite observations using a machine learning method, Remote Sensing, 13, 969, 2021.

Chen, J., Dietrich, F., Maazallahi, H., Forstmaier, A., Winkler, D., Hofmann, M. E., Denier van der Gon, H., and Röckmann, T.: Methane emissions from the Munich Oktoberfest, Atmospheric Chemistry and Physics, 20, 3683–3696, 2020.

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

Cheng, X., Zhang, W., Wenzel, A., and Chen, J.: Stacked ResNet-LSTM and CORAL model for multi-site air quality prediction, Neural Computing and Applications, 34, 13 849–13 866, 2022.

Council, N. R. et al.: Rethinking the ozone problem in urban and regional air pollution, National Academies Press, 1992.

Crippa, M., Janssens-Maenhout, G., Guizzardi, D., Van Dingenen, R., and Dentener, F.: Contribution and uncertainty of sectorial and regional emissions to regional and global PM 2.5 health impacts, Atmospheric Chemistry and Physics, 19, 5165–5186, 2019.

Crutzen, P. J.: Tropospheric ozone: An overview, Tropospheric ozone, pp. 3–32, 1988.

De Hoogh, K., Saucy, A., Shtein, A., Schwartz, J., West, E. A., Strassmann, A., Puhan, M., Ro¨o¨sli, M., Stafoggia, M., and Kloog, I.: Predicting fine-scale daily NO2 for 2005–2016 incorporating OMI satellite data across Switzerland, Environmental science & technology, 53, 10 279–10 287, 2019.

Diao, L., Bi, X., Zhang, W., Liu, B., Wang, X., Li, L., Dai, Q., Zhang, Y., Wu, J., and Feng, Y.: The Characteristics of Heavy Ozone Pollution Episodes and Identification of the Primary Driving Factors Using a Generalized Additive Model (GAM) in an Industrial Megacity of Northern China, Atmosphere, 12, 1517, 2021.

Gardner, M. W. and Dorling, S.: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences, Atmospheric environment, 32, 2627–2636, 1998.

Gensheimer, J., Chen, J., Turner, A. J., Shekhar, A., Wenzel, A., and Keutsch, F. N.: What Are the Different Measures of Mobility Telling Us About Surface Transportation CO2 Emissions During the COVID-19 Pandemic?, Journal of Geophysical Research: Atmospheres, 126, e2021JD034 664, 2021.

Ghahremanloo, M., Lops, Y., Choi, Y., and Yeganeh, B.: Deep Learning Estimation of Daily Ground-Level NO2 Concentrations From Remote Sensing Data, Journal of Geophysical Research: Atmospheres, 126, e2021JD034 925, 2021.

Guenther, A. B., Zimmerman, P. R., Harley, P. C., Monson, R. K., and Fall, R.: Isoprene and monoterpene emission rate variability: model evaluations and sensitivity analyses, Journal of Geophysical Research: Atmospheres, 98, 12 609–12 617, 1993.

Heaton, J.: Applications of Deep Neural Networks, 2020.

Hoffmann, B., Boogaard, H., de Nazelle, A., Andersen, Z. J., Abramson, M., Brauer, M., Brunekreef, B., Forastiere, F., Huang, W., Kan, H., et al.: WHO Air Quality Guidelines 2021–Aiming for Healthier Air for all: A Joint Statement by Medical, Public Health, Scientific Societies and Patient Representative Organisations, International journal of public health, p. 88, 2021.

Hu, C., Kang, P., Jaffe, D. A., Li, C., Zhang, X., Wu, K., and Zhou, M.: Understanding the impact of meteorology on ozone in 334 cities of China, Atmospheric Environment, 248, 118 221, 2021.

Hu, J., Chen, J., Ying, Q., and Zhang, H.: One-year simulation of ozone and particulate matter in China using WRF/CMAQ modeling system, Atmospheric Chemistry and Physics, 16, 10 333–10 350, 2016.

Inness, A., Blechschmidt, A.-M., Bouarar, I., Chabrillat, S., Crepulja, M., Engelen, R., Eskes, H., Flemming, J., Gaudel, A., Hendrick, F., et al.: Data assimilation of satellite-retrieved ozone, carbon monoxide and nitrogen dioxide with ECMWF's Composition-IFS, Atmospheric chemistry and physics, 15, 5275–5303, 2015.

Jacob, D. J.: Introduction to atmospheric chemistry, Princeton University Press, 1999.

Jin, X., Fiore, A. M., Murray, L. T., Valin, L. C., Lamsal, L. N., Duncan, B., Folkert Boersma, K., De Smedt, I., Abad, G. G., Chance, K., et al.: Evaluating a space-based indicator of surface ozone-NOx-VOC sensitivity over midlatitude source regions and application to decadal trends, Journal of Geophysical Research: Atmospheres, 122, 10–439, 2017.

Jin, X., Fiore, A., Boersma, K. F., Smedt, I. D., and 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, Environmental science & technology, 54, 6518–6529, 2020.

Kang, Y., Choi, H., Im, J., Park, S., Shin, M., Song, C.-K., and Kim, S.: Estimation of surface-level NO2 and O3 concentrations using TROPOMI data and machine learning over East Asia, Environmental Pollution, 288, 117 711, 2021.

Kim, M., Brunner, D., and Kuhlmann, G.: Importance of satellite observations for high-resolution mapping of near-surface NO2 by machine learning, Remote Sensing of Environment, 264, 112 573, 2021.

Lee, M., Lin, L., Chen, C.-Y., Tsao, Y., Yao, T.-H., Fei, M.-H., and Fang, S.-H.: Forecasting air quality in Taiwan by using machine learning, Scientific reports, 10, 4153, 2020.

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., and Pozzer, A.: The contribution of outdoor air pollution sources to premature mortality on a global scale, Nature, 525, 367–371, 2015.

Li, H., Yang, Y., Jin, J., Wang, H., Li, K., Wang, P., and Liao, H.: Climate-driven deterioration of future ozone pollution in Asia predicted by machine learning with multisource data, Atmospheric Chemistry and Physics Discussions, pp. 1–40, 2022.

Li, T., Wang, Y., and Yuan, Q.: Remote sensing estimation of regional NO2 via space-time neural networks, Remote Sensing, 12, 2514, 2020.

Liang, Y.-C., Maimury, Y., Chen, A. H.-L., and Juarez, J. R. C.: Machine learning-based prediction of air quality, Applied Sciences, 10, 9151, 2020.

Lin, X., Trainer, M., and Liu, S.: On the nonlinearity of the tropospheric ozone production, Journal of Geophysical Research: Atmospheres, 93, 15 879–15 888, 1988.

Liu, Y., Wang, P., Li, Y., Wen, L., and Deng, X.: Air quality prediction models based on meteorological factors and real-time data of industrial waste gas, Scientific Reports, 12, 9253, 2022.

Lou, S., Liao, H., Yang, Y., and Mu, Q.: Simulation of the interannual variations of tropospheric ozone over China: Roles of variations in meteorological parameters and anthropogenic emissions, Atmospheric Environment, 122, 839–851, 2015.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I.: From local explanations to global understanding with explainable AI for trees, Nature machine intelligence, 2, 56–67, 2020.

Malashock, D. A., DeLang, M. N., Becker, J. S., Serre, M. L., West, J. J., Chang, K.-L., Cooper, O. R., and Anenberg, S. C.: Estimates of ozone concentrations and attributable mortality in urban, peri-urban and rural areas worldwide in 2019, Environmental Research Letters, 17, 054 023, 2022.

McDuffie, E. E., Smith, S. J., O'Rourke, P., Tibrewal, K., Venkataraman, C., Marais, E. A., Zheng, B., Crippa, M., Brauer, M., and Martin, R. V.: A global anthropogenic emission inventory of atmospheric pollutants from sector-and fuel-specific sources (1970–2017): an application of the Community Emissions Data System (CEDS), Earth System Science Data, 12, 3413–3442, 2020.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., and Nauss, T.: Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation, Environmental Modelling & Software, 101, 1–9, 2018.

Nussbaumer, C. M. and Cohen, R. C.: The role of temperature and NO x in ozone trends in the Los Angeles basin, Environmental Science & Technology, 54, 15 652–15 659, 2020.

Osses, M., Rojas, N., Ibarra, C., Valdebenito, V., Laengle, I., Pantoja, N., Osses, D., Basoa, K., Tolvett, S., Huneeus, N., et al.: High-resolution spatial-distribution maps of road transport exhaust emissions in Chile, 1990–2020, Earth System Science Data, 14, 1359–1376, 2022.

Pisoni, E., Albrecht, D., Mara, T. A., Rosati, R., Tarantola, S., and Thunis, P.: Application of uncertainty and sensitivity analysis to the air quality SHERPA modelling tool, Atmospheric environment, 183, 84–93, 2018.

Pusede, S. and Cohen, R.: On the observed response of ozone to NO x and VOC reactivity reductions in San Joaquin Valley California 1995–present, Atmospheric Chemistry and Physics, 12, 8323–8339, 2012.

Pusede, S., Gentner, D., Wooldridge, P., Browne, E., Rollins, A., Min, K.-E., Russell, A., Thomas, J., Zhang, L., Brune, W., 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 Chemistry and Physics, 14, 3373–3395, 2014.

Qu, Z., Jacob, D. J., Silvern, R. F., Shah, V., Campbell, P. C., Valin, L. C., and Murray, L. T.: US COVID-19 shutdown demonstrates importance of background NO2 in inferring NOx emissions from satellite NO2 observations, Geophysical research letters, 48, e2021GL092 783, 2021.

Sicard, P., Paoletti, E., Agathokleous, E., Araminienė, V., Proietti, C., Coulibaly, F., and De Marco, A.: Ozone weekend effect in cities: Deep insights for urban air pollution control, Environmental Research, 191, 110 193, 2020.

Sillman, S.: The relation between ozone, NOx and hydrocarbons in urban and polluted rural environments, Atmospheric Environment, 33, 1821–1845, 1999.

Sillman, S., Logan, J. A., and Wofsy, S. C.: The sensitivity of ozone to nitrogen oxides and hydrocarbons in regional ozone episodes, Journal of Geophysical Research: Atmospheres, 95, 1837–1851, 1990.

Singh, J., Singh, N., Ojha, N., Sharma, A., Pozzer, A., Kiran Kumar, N., Rajeev, K., Gunthe, S. S., and Kotamarthi, V. R.: Effects of spatial resolution on WRF v3. 8.1 simulated meteorology over the central Himalaya, Geoscientific Model Development, 14, 1427–1443, 2021.

Trombetti, M., Thunis, P., Bessagnet, B., Clappier, A., Couvidat, F., Guevara, M., Kuenen, J., and López-Aparicio, S.: Spatial intercomparison of Top-down emission inventories in European urban areas, Atmospheric Environment, 173, 142–156, 2018.

Vlasenko, A., Matthias, V., and Callies, U.: Simulation of chemical transport model estimates by means of a neural network using meteorological data, Atmospheric Environment, 254, 118 236, 2021.

Wang, W., van der A, R., Ding, J., van Weele, M., and Cheng, T.: Spatial and temporal changes of the ozone sensitivity in China based on satellite and ground-based observations, Atmospheric Chemistry and Physics, 21, 7253–7269, 2021.

Xie, X., Wang, T., Yue, X., Li, S., Zhuang, B., Wang, M., and Yang, X.: Numerical modeling of ozone damage to plants and its effects on atmospheric CO2 in China, Atmospheric Environment, 217, 116 970, 2019.

Zaini, N., Ean, L. W., Ahmed, A. N., Abdul Malek, M., and Chow, M. F.: PM2. 5 forecasting for an urban area based on deep learning and decomposition method, Scientific Reports, 12, 17 565, 2022.

Zhang, J., Chen, Q., Wang, Q., Ding, Z., Sun, H., and Xu, Y.: The acute health effects of ozone and PM2. 5 on daily cardiovascular disease mortality: A multi-center time series study in China, Ecotoxicology and Environmental Safety, 174, 218–223, 2019.

Zhao, Z., Wu, J., Cai, F., Zhang, S., and Wang, Y.-G.: A hybrid deep learning framework for air quality prediction with spatial autocorrelation during the COVID-19 pandemic, Scientific Reports, 13, 1015, 2023.

Zhu, Q., Bi, J., Liu, X., Li, S., Wang, W., Zhao, Y., and Liu, Y.: Satellite-Based Long-Term Spatiotemporal Patterns of Surface Ozone Concentrations in China: 2005–2019, Environmental health perspectives, 130, 027 004, 2022.

Zong, R., Yang, X., Wen, L., Xu, C., Zhu, Y., Chen, T., Yao, L., Wang, L., Zhang, J., Yang, L., et al.: Strong ozone production at a rural site in theNorth China Plain: Mixed effects of urban plumesand biogenic emissions, Journal of Environmental Sciences, 71, 261–270, 2018.