



OPEN

Exploring the potential of machine learning for simulations of urban ozone variability

Narendra Ojha¹✉, Imran Girach²✉, Kiran Sharma³, Amit Sharma⁴, Narendra Singh⁵ & Sachin S. Gunthe^{6,7}

Machine learning (ML) has emerged as a powerful technique in the Earth system science, nevertheless, its potential to model complex atmospheric chemistry remains largely unexplored. Here, we applied ML to simulate the variability in urban ozone (O₃) over Doon valley of the Himalaya. The ML model, trained with past variations in O₃ and meteorological conditions, successfully reproduced the independent O₃ data ($r^2 \sim 0.7$). Model performance is found to be similar when the variation in major precursors (CO and NO_x) were included in the model, instead of the meteorology. Further the inclusion of both precursors and meteorology improved the performance significantly ($r^2 = 0.86$) and the model could also capture the outliers, which are crucial for air quality assessments. We suggest that in absence of high-resolution measurements, ML modeling has profound implications for unraveling the feedback between pollution and meteorology in the fragile Himalayan ecosystem.

The chemical processes in the urban atmospheres of Himalayan foothills have strong potential to impact the regional air quality, agriculture, and therefore the economy^{1–3}. In addition, the build-up of climate-forcing pollution in the Himalayan region can have irreversible effects on the hydrological cycle and global climate^{1,2,4–7}. The atmospheric dynamics above the Himalaya also form the crossroad of so called “Atmospheric Brown Clouds” to the Tibetan Plateau⁸. Recent increase in extreme weather events triggering the calamities also indicate an intensifying interplay between the increasing pollution and meteorology over fragile ecosystem of the Himalaya^{9–12}.

The enhanced concentrations of ozone (O₃) and other climate-forcing pollutants in the Himalayan foothills are attributed to unprecedented growth in population and urbanization^{13–16}. The intense forest-fires, diverse natural factors, and the topography also play vital roles in the build-up of trace gases and aerosols here^{5,11,15,17–19}. The Himalayan atmosphere is particularly influenced by the most densely populated region of the world—the Indo-Gangetic Plain (IGP)^{20,21}. The IGP is a global hotspot of elevated O₃ and aerosol loading due to strong anthropogenic emissions and intense crop-residue burning in prevalence of favorable meteorological conditions^{22–27}. The emissions and photochemistry in the IGP affect the Himalayan atmosphere in particular through the mountain meteorology and boundary layer dynamics^{20,21,28}. A potential climate warming combined with future increase in the emissions can further intensify the atmospheric chemistry over this part of the world^{29–31}.

Considering the discussed scenario, measurements and modeling studies have been conducted to assess the effects of diverse emissions, photochemistry, and dynamics on atmospheric composition over the Himalayan region^{8,15,17,32,33}. The concentrations of O₃ and precursors were found to be enhanced during pre-monsoon (spring) and post-monsoon (autumn) seasons due to regional pollution supplemented with biomass-burning, intense solar radiation, and less precipitation^{15,20,21,34,35}. The long-term measurements of atmospheric composition and meteorological parameters however remain lacking over the Himalayan foothills in India, which are experiencing severe air quality and extreme weather events. Studies to fill this gap are of paramount significance since the chemistry-climate models also have greater biases in reproducing already sparse measurements over the Himalayan region^{20,34,36,37}. The stronger biases are suggested to be mainly due to the limitation of models in resolving the highly complex topography of Himalaya and foothills^{5,19,20,37,38}. The uncertainties in the emission inventories and parameterizations of physical and chemical processes also increase the biases in the

¹Physical Research Laboratory, Ahmedabad, India. ²Space Physics Laboratory, Vikram Sarabhai Space Centre, Thiruvananthapuram, India. ³Graphic Era (Deemed to be University), Dehradun, India. ⁴Department of Civil and Infrastructure Engineering, Indian Institute of Technology Jodhpur, Jodhpur, India. ⁵Aryabhata Research Institute of Observational Sciences, Nainital, India. ⁶EWRE Division, Department of Civil Engineering, Indian Institute of Technology Madras, Chennai, India. ⁷Laboratory for Atmospheric and Climate Sciences, Indian Institute of Technology Madras, Chennai, India. ✉email: ojha@prl.res.in; imran.girach@gmail.com

models^{19,37,39–41}. Besides higher biases, the conventional models also need intensive computing resources which poses further limitation in conducting high-resolution simulation.

In the current era, the artificial intelligence (AI) and machine learning (ML) have emerged as powerful alternative tools for modeling in various fields including the Earth system science^{42–45}. Recent studies utilized AI/ML modeling in the analyses of extreme weather events and prediction of oceanic phenomenon as well as atmospheric composition^{46–48}. These studies have shown that ML models trained with data from observations or physical models can produce reliable simulations without intensive high-end computing. Nevertheless, the applications of AI/ML to simulate complex atmospheric chemistry remain still limited. Considering the scientific and societal implications, lack of measurements, and limitations of conventional models over Himalayan region, the objectives of this study are as follows:

- (1) To explore the potential of ML modeling for simulating urban O₃ variability.
- (2) To study the effects of meteorological and chemical variables on model performance.
- (3) To assess the effect of the data fraction used in the training on model performance.

The study region, datasets, and modeling are described in the “[Methodology](#)” section. Model simulations and results are presented in the “[Model simulations and results](#)” section, followed by “[Discussion](#)” section.

Methodology

Study region and datasets. The study is focussed on the urban O₃ chemistry over the Doon valley of the Himalaya. We initiated in situ O₃ measurements using an online O₃ analyser manufactured by Environnement S.A, France (model O₃ 42) at a representative station—Graphic Era deemed to be University campus (77.99° E, 30.27° N, 600 m above mean sea level). The observations are based on the UV light absorption by O₃ and instrumental uncertainty is about 5%⁴⁹. The continuous measurements are being conducted since April 2018 except during February–August 2020 when the laboratory was not open due to severe impact of the COVID-19 pandemic. Further details of these O₃ measurements are presented in the earlier study¹⁵.

Auxiliary datasets used in training the ML model include the meteorological and chemical reanalysis from the ECMWF (European Center for Medium range Weather Forecasting). The meteorological parameters: temperature, humidity, horizontal winds, and boundary layer height (BLH) are included from the ERA-Interim⁵⁰. Whereas, the chemical species: O₃ and key precursors (CO, NO, NO₂) have been included from the CAMS (Copernicus Atmosphere Monitoring Service) reanalysis⁵¹. ERA-Interim and CAMS products have been analyzed for diverse studies including over the Indian region^{15,52–54}. The CAMS data has been shown to reproduce the day-to-day variability in the noontime O₃ over the study region¹⁵. Here, we have utilized the reanalysis data of 2003–2019 period focusing on the noontime variations (6 GMT; 11:30 local time), when the urban O₃ photochemistry is most intense.

Machine learning model. This study utilizes the XGBoost (Extreme Gradient Boosting) algorithm of the ML modeling⁵⁵ to simulate the O₃ variations. Considering the dependence of O₃ on meteorological parameters and precursor gases, this modeling is under the supervised learning of AI. In the gradient boosting algorithm, a prediction model is developed in form of an ensemble of weak prediction systems i.e., decision trees. The model is built in a stage-wise manner and generalizations are made by allowing optimization of an arbitrary differentiable loss function (e.g., squared error). Further details of the XGBoost can be found elsewhere⁵⁵. The method adopted to build and evaluate the model is shown as a flow chart in the Fig. 1.

Hyper parameters have been varied iteratively following the trial and error method to achieve better prediction. The parameters were fine-tuned using the grid search function (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html). The values of hyper parameters set in the model are given in the supplementary material—Table S1. Other hyper parameters were kept to their default values (<https://xgboost.readthedocs.io/en/latest/parameter.html>). To avoid overfitting, the iterations are aborted once they cease to improve the fit parameters further, i.e., no reduction in RMSE (root mean square error) over 100 iterations. The model performance in simulating O₃ variations has been evaluated by estimating correlation (r²), slope of linear fit, and RMSE (root mean square error).

Model simulations and results

A series of simulations have been performed under this study, as summarized in the Table 1. These simulations and the evaluation of model performance are discussed in the following subsections.

Simulation utilizing in-situ O₃ measurements. In the first simulation ML_obs_O₃_met_prec, the ML model has been trained using the observational data of O₃ and reanalysis data of meteorological parameters (met) and precursors (prec). Analysis is focussed on the variations in noontime (11:30 h local time) O₃. The data of April 2018 to April 2019 (number of days N=222) has been used for training the ML model, which is 50% of total available data. Model simulation is evaluated against remaining independent observations for April–December 2019 period (N=223 days). Figure 2 shows the correlation between the ML model simulation and in-situ measurements of noontime O₃ over Doon valley for April–December 2019 period. ML model is found to successfully reproduce the temporal variability in the noontime O₃ with r² value of 0.75 (p < 0.01) and RMSE value of 10 ppbv. The estimated bias in ML model result is seen to be significantly lower as compared to the bias values reported in global and regional atmospheric models over this region^{15,34,37}. The result suggests that in absence of high-resolution measurements, the ML modeling can be combined with reanalysis and limited

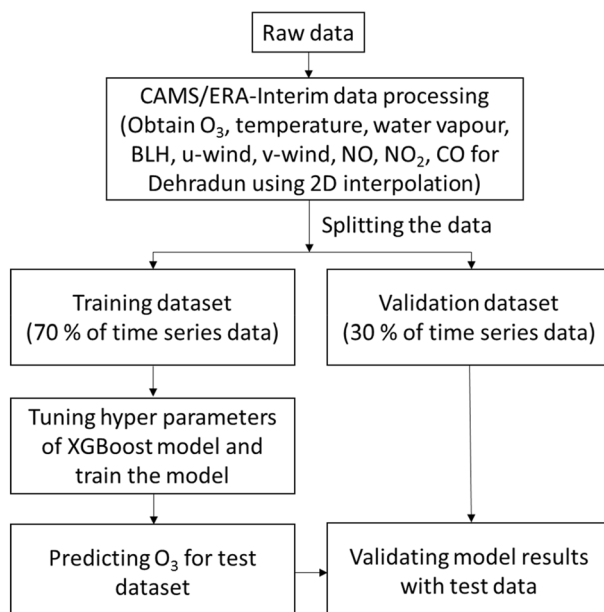


Figure 1. Flow chart of the steps in building the ML model for simulation of urban O₃ variations and evaluation.

Serial no	Simulation	Training data		
		O ₃	Meteorology	Precursors
1	ML_obs_O ₃ _met_prec	Measurements	ERA-interim	CAMS
2	ML_cams_O ₃	CAMS	-	-
3	ML_cams_O ₃ _met	CAMS	ERA-interim	-
4	ML_cams_O ₃ _prec	CAMS	-	CAMS
5	ML_cams_O ₃ _met_prec	CAMS	ERA-interim	CAMS

Table 1. Different ML model simulations performed in the study.

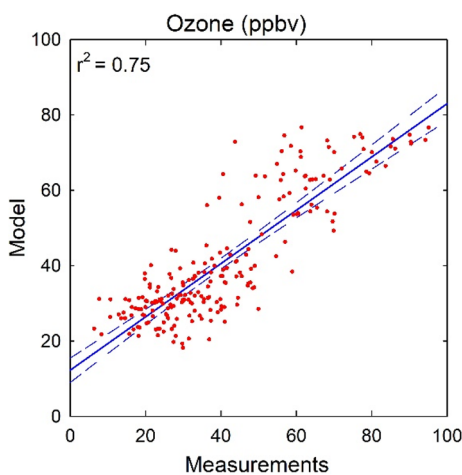


Figure 2. Correlation between measurements and model (ML_obs_O₃_met_prec) simulated variations in noontime O₃ over Doon valley during April–December 2019. Solid blue line shows the linear regression fit and dashed lines show the 99% confidence intervals.

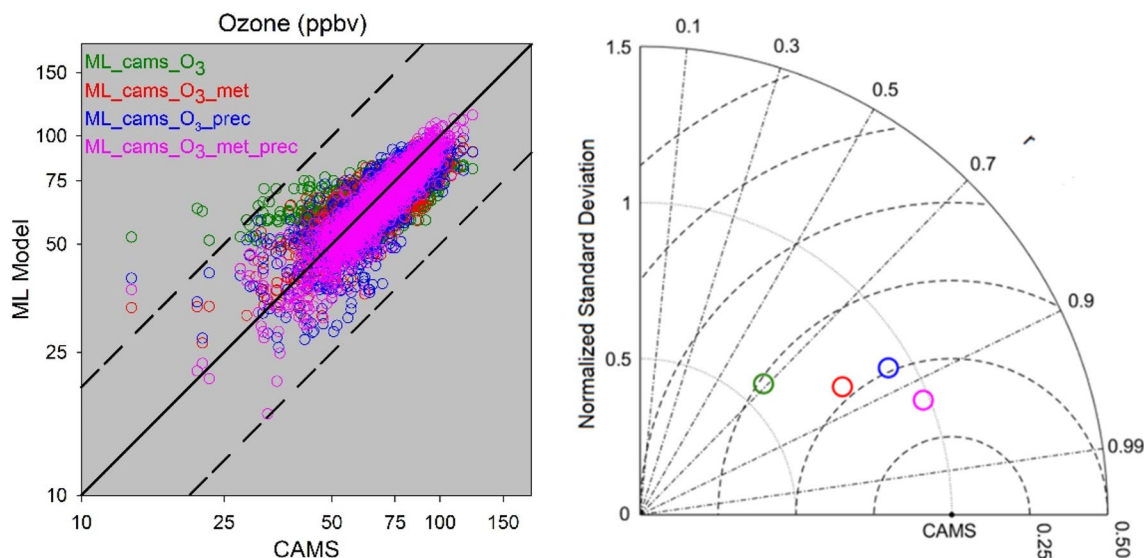


Figure 3. Scatter plot and Taylor's diagram evaluating the ML model simulations of noontime O₃ variations as compared to the CAMS reanalysis.

in-situ data over the Himalayan region. The strong correlation between the ML model and in-situ measurements further opens up the possibility to utilize such simulations for assessing the impacts of O₃ on agriculture and health in this region.

Simulations utilizing long-term CAMS O₃. The in-situ measurements are largely unavailable in the Indian Himalayan region and the temporal coverage is also very limited. In view of this, we include the long-term CAMS data to assess the potential and performance of ML modeling more deeply. With availability of long-term data, here, we train the ML Model with noontime (11:30 local time) CAMS O₃ and reanalysis meteorology for 2003–2015 (70% of total data). This makes a significant fraction (30%) of total data during 2015–2019 period available for the evaluation. The simplest simulation is ML_cams_O₃ in which model is trained only with the O₃ time series without including any additional parameter. This simulation is found to predict the independent O₃ variations with r^2 value of 0.47 and RMSE of 11.6 ppbv (Fig. 3). This result is a manifestation of a periodicity in O₃ data embedded by the seasonal cycle in India.

The relative effect of including variations in the meteorological parameters versus major precursors (CO, NO, NO₂) has been evaluated by performing additional simulations (Table 1, Fig. 3). Model trained with O₃ and meteorology (ML_cams_O₃_met) reproduces independent O₃ variations with r^2 value of 0.71 and slope value of 0.65. Another simulation in which the ML model is trained with O₃ and precursors but not with the meteorology shows similar or slightly improved performance ($r^2 = 0.74$, slope = 0.79, $p < 0.01$). The inter-comparison of these two simulations suggests that reasonable predictions of urban O₃ variability can be made with ML models trained with either of the meteorological or precursor dataset. This is important as this region lacks comprehensive datasets especially of the precursors, and in such cases the meteorological datasets can be used to predict O₃. Further, to explore the potential of ML approach, we performed another simulation ML_cams_O₃_met_prec in which both meteorology as well as precursors have been included in the model. This led to significant improvement in the model performance with r^2 value as high as 0.86 and slope value of 0.91. For this simulation, the RMSE value also drops drastically to 6 ppbv and the mean bias is also smaller (~3 ppbv). An important finding is that when the potentials of both meteorological as well as chemical datasets are combined, the model's ability to predict outliers improves drastically, which is of major significance in air quality assessments.

A comparison of r^2 values among all these simulations (numbered 2–5 in the Table 1) suggests that ~47% of O₃ variations can be explained ($r^2 = 0.47$ in ML_cams_O₃) by the periodicities embedded in the data originated from the seasonal cycle. As precursors and meteorology act in tandem, higher r^2 values (~0.7) in simulations trained with either meteorology or precursors suggest that this additional ~25% of O₃ variability can be attributed to the changes in meteorology or precursor levels. Meteorology plus major precursors could explain ~86% ($r^2 = 0.86$ in ML_cams_O₃_met_prec) of the variations in the urban O₃. The remaining variability could be due to diverse unaccounted factors such as deposition, vertical transport, and volatile organic compounds, etc. The analysis suggests that ML simulations can provide deep insights into the relative importance of the physical and chemical processes affecting the air quality.

The performance of different simulations has been compiled in form of a Taylor's diagram (Fig. 3). The figure includes statistics like r , normalized RMSE, and normalized standard deviation (SD) where normalization is done with respect to the SD in the reference (CAMS). The relative performance of different simulations is assessed by comparing how close a simulation is to the reference point (CAMS). For an ideal agreement, ML simulation should coincide with the reference point ($r = 1$, normalized SD = 1, and normalized RMSE = 0). It is evident that the ML simulation exploiting the potentials of both meteorology and precursors (ML_cams_O₃_met_prec)

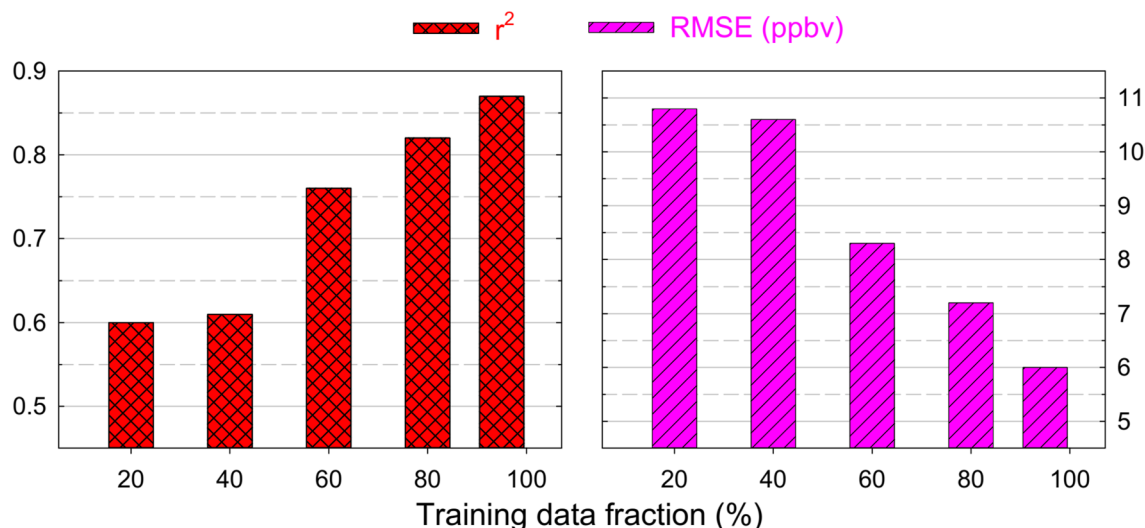


Figure 4. Variation in r^2 and RMSE with change in the percentage data used in training of ML model.

performed the best. Besides stronger r value, a normalized SD value close to 1 suggests that the simulation produces similar extent of the variability as in the CAMS. On the other hand, ML simulations using either meteorology or precursors had similar performance. Also, ML_cams_O₃_prec produced more variability likely due to non-linearities in chemistry as compared with the simulation using meteorological variations (ML_cams_O₃_met).

Effect of training data length. We further investigate the sensitivity of model performance to the fraction of available data being used for the training. In this regard, a series of simulations have been performed using the best performing model set up (ML_cams_O₃_met_prec) by using 20–95% data for model training. Figure 4 shows the variations in r^2 and RMSE values due to variation in the training data fraction. The analysis shows that the model performance is highly sensitive to the length of total data being used in its training. The r^2 value is found to increase significantly from about 0.6–0.87 and RMSE shows reduction from ~11 to 6 ppbv with increase in the training data fraction. The analysis suggests that longer time-dependent datasets are highly desirable for optimizing performance of ML models in predicting air quality variation. This underlines that long-term in situ measurements and validated chemistry-climate simulations can help in further exploiting the potential offered by the ML approach.

Discussion

Our study unravels the strong potential of ML modeling for computationally inexpensive simulations of urban O₃ variability in the Himalayan foothills region. The periodicity in O₃ and meteorological parameters due to systematic seasonal cycle of India tends to allow ML model to reproduce data fairly well. In lack of high-resolution measurements, ML simulations can be used to assess the impacts of O₃ on health and agriculture in this region. Additionally, the series of simulations conducted here would serve as a reference for further applications of AI/ML based modeling to complement conventional Earth system models. It is however pointed out that here the environment is urban and the O₃ variations are greatly governed by the regional photochemistry. The scenario could be very different for cleaner remote regions where O₃ variability is dominated by transport from upwind polluted regions or from the higher altitudes. In this regard, we recommend establishing baseline stations to continuously monitor the atmospheric composition as well as the meteorology to exploit the full potential of ML modeling. Model performance is already promising with inclusion of only meteorology, nevertheless, the inclusion of precursors enhances the model's ability to capture outliers, which are critical in air quality assessments. Future studies may extend the scope to additional climate-forcing pollutants and to unravel feedback between pollution and meteorology causing calamities in the fragile ecosystem of the Himalaya experiencing strong anthropogenic pressure.

Received: 24 August 2021; Accepted: 1 November 2021

Published online: 18 November 2021

References

1. Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D. & Pozzer, A. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature* **525**(7569), 367–371. <https://doi.org/10.1038/nature15371> (2015).
2. Sharma, A., Ojha, N., Pozzer, A., Beig, G. & Gunthe, S. S. Revisiting the crop yield loss in India attributable to ozone. *Atmos. Environ.* **X 1**, 100008. <https://doi.org/10.1016/j.aeaoa.2019.100008> (2019).
3. Ghude, S. D. *et al.* Premature mortality in India due to PM_{2.5} and ozone exposure. *Geophys. Res. Lett.* **43**(9), 4650–4658. <https://doi.org/10.1002/2016GL068949> (2016).

4. Ramanathan, V. *et al.* Atmospheric brown clouds: Impacts on South Asian climate and hydrological cycle. *Proc. Natl. Acad. Sci.* **102**(15), 5326–5333. <https://doi.org/10.1073/pnas.0500656102> (2005).
5. Pant, G. B., Kumar, P., Revadekar, J. V. & Singh, N. *Climate Change in the Himalayas* (Springer, Cham, 2018). <https://doi.org/10.1007/978-3-319-61654-4>.
6. Kotamarthi, V. R. Ganges Valley Aerosol Experiment: Science and Operations Plan. DOE/SC-ARM-10-019 (2010).
7. Sarangi, C. *et al.* Dust dominates high-altitude snow darkening and melt over high-mountain Asia. *Nat. Clim. Change* **10**(11), 1045–1051. <https://doi.org/10.1038/s41558-020-00909-3> (2020).
8. Lüthi, Z. L. *et al.* Atmospheric brown clouds reach the Tibetan Plateau by crossing the Himalayas. *Atmos. Chem. Phys.* **15**(11), 6007–6021. <https://doi.org/10.5194/acp-15-6007-2015> (2015).
9. Chug, D. *et al.* Observed evidence for steep rise in the extreme flow of western Himalayan rivers. *Geophys. Res. Lett.* **47**(15), e2020GL087815. <https://doi.org/10.1029/2020GL087815> (2020).
10. Choudhury, G. *et al.* Aerosol-induced high precipitation events near the Himalayan foothills. *Atmos. Chem. Phys.* **2015**, 1–17. <https://doi.org/10.5194/acp-2020-440> (2020).
11. Saikawa, E. *et al.* Air pollution in the Hindu Kush Himalaya. In *The Hindu Kush Himalaya assessment: Mountains, climate change, sustainability and people* (eds Wester, P. *et al.*) 339–387 (Springer, Berlin, 2019). https://doi.org/10.1007/978-3-319-92288-1_10.
12. Singh, O., Arya, P. & Chaudhary, B. S. On rising temperature trends at Dehradun in Doon valley of Uttarakhand, India. *J. Earth Syst. Sci.* **122**, 613–622. <https://doi.org/10.1007/s12040-013-0304-0> (2013).
13. Pandit, M. K. The Himalayas must be protected. *Nature* **501**(7467), 283. <https://doi.org/10.1038/501283a> (2013).
14. Prabhu, V. *et al.* Black carbon and biomass burning associated high pollution episodes observed at Doon valley in the foothills of the Himalayas. *Atmos. Res.* **243**, 105001. <https://doi.org/10.1016/j.atmosres.2020.105001> (2020).
15. Ojha, N. *et al.* Surface ozone in the Doon Valley of the Himalayan foothills during spring. *Environ. Sci. Pollut. Res.* **26**(19), 19155–19170. <https://doi.org/10.1007/s11356-019-05085-2> (2019).
16. Deep, A. *et al.* Evaluation of ambient air quality in Dehradun city during 2011–2014. *J. Earth Syst. Sci.* <https://doi.org/10.1007/s12040-019-1092-y> (2019).
17. Kumar, R. *et al.* Influences of the springtime northern Indian biomass burning over the central Himalayas. *J. Geophys. Res. Atmos.* **116**(19), 1–14. <https://doi.org/10.1029/2010JD015509> (2011).
18. Singh, N. *et al.* Boundary layer evolution over the central Himalayas from radio wind profiler and model simulations. *Atmos. Chem. Phys.* <https://doi.org/10.5194/acp-16-10559-2016> (2016).
19. Singh, J. *et al.* Effects of spatial resolution on WRF v3.8.1 simulated meteorology over the central Himalaya. *Geosci. Model Dev.* **14**, 1427–1443 (2021).
20. Ojha, N. *et al.* Variabilities in ozone at a semi-urban site in the Indo-Gangetic Plain region: Association with the meteorology and regional processes. *J. Geophys. Res. Atmos.* **117**(20), 1–19. <https://doi.org/10.1029/2012JD017716> (2012).
21. Sarangi, T. *et al.* First simultaneous measurements of ozone, CO, and NO_y at a high-altitude regional representative site in the central Himalayas. *J. Geophys. Res. Atmos.* **119**(3), 1592–1611. <https://doi.org/10.1002/2013JD020631> (2014).
22. Kumar, V. & Sinha, V. Season-wise analyses of VOCs, hydroxyl radicals and ozone formation chemistry over north-west India reveal isoprene and acetaldehyde as the most potent ozone precursors throughout the year. *Chemosphere* **283**, 131184. <https://doi.org/10.1016/j.chemosphere.2021.131184> (2021).
23. Ojha, N. *et al.* On the widespread enhancement in fine particulate matter across the Indo-Gangetic Plain towards winter. *Sci. Rep.* **10**(1), 5862. <https://doi.org/10.1038/s41598-020-62710-8> (2020).
24. Nelson, B. S. *et al.* In situ Ozone Production is highly sensitive to Volatile Organic Compounds in the Indian Megacity of Delhi. *Atmos. Chem. Phys. Discuss.* **2021**, 1–36. <https://doi.org/10.5194/acp-2021-278> (2021).
25. Chen, Y. *et al.* Avoiding high ozone pollution in Delhi. *Faraday Discuss.* **226**, 502–514. <https://doi.org/10.1039/d0fd00079e> (2021).
26. Fishman, J., Wozniak, A. E. & Creilson, J. K. Global distribution of tropospheric ozone from satellite measurements using the empirically corrected tropospheric ozone residual technique: Identification of the regional aspects of air pollution. *Atmos. Chem. Phys.* **3**(4), 893–907. <https://doi.org/10.5194/acp-3-893-2003> (2003).
27. Pallavi, S. B. & Sinha, V. Source apportionment of volatile organic compounds in the northwest Indo-Gangetic Plain using a positive matrix factorization model. *Atmos. Chem. Phys.* **19**(24), 15467–15482. <https://doi.org/10.5194/acp-19-15467-2019> (2019).
28. Solanki, R., Singh, N., Kiran Kumar, N. V. P., Rajeev, K. & Dhaka, S. K. Time variability of surface-layer characteristics over a mountain ridge in the central Himalayas during the spring season. *Bound. Layer Meteorol.* **158**(3), 453–471. <https://doi.org/10.1007/s10546-015-0098-5> (2016).
29. Lelieveld, J. *et al.* The South Asian monsoon—Pollution pump and purifier. *Science (80-)* **361**(6399), 270–273. <https://doi.org/10.1126/science.aar2501> (2018).
30. Coates, J., Mar, K. A., Ojha, N. & Butler, T. M. The influence of temperature on ozone production under varying NO_x conditions—A modelling study. *Atmos. Chem. Phys.* <https://doi.org/10.5194/acp-16-11601-2016> (2016).
31. Kumar, R. *et al.* How will air quality change in South Asia by 2050?. *J. Geophys. Res. Atmos.* **123**(3), 1840–1864. <https://doi.org/10.1002/2017JD027357> (2018).
32. Rupakheti, D. *et al.* Pre-monsoon air quality over Lumbini, a world heritage site along the Himalayan foothills. *Atmos. Chem. Phys.* **17**(18), 11041–11063. <https://doi.org/10.5194/acp-17-11041-2017> (2017).
33. Sarkar, C. *et al.* Overview of VOC emissions and chemistry from PTR-TOF-MS measurements during the SusKat-ABC campaign: High acetaldehyde, isoprene and isocyanic acid in wintertime air of the Kathmandu Valley. *Atmos. Chem. Phys.* **16**(6), 3979–4003. <https://doi.org/10.5194/acp-16-3979-2016> (2016).
34. Kumar, R., Naja, M., Venkataramani, S. & Wild, O. Variations in surface ozone at Nainital: A high-altitude site in the central Himalayas. *J. Geophys. Res. Atmos.* **115**(August), 1–12. <https://doi.org/10.1029/2009JD013715> (2010).
35. Bhardwaj, P. *et al.* Variations in surface ozone and carbon monoxide in the Kathmandu Valley and surrounding broader regions during SusKat-ABC field campaign: Role of local and regional sources. *Atmos. Chem. Phys.* **18**(16), 11949–11971. <https://doi.org/10.5194/acp-18-11949-2018> (2018).
36. Sharma, A. *et al.* WRF-Chem simulated surface ozone over south Asia during the pre-monsoon: Effects of emission inventories and chemical mechanisms. *Atmos. Chem. Phys.* <https://doi.org/10.5194/acp-17-14393-2017> (2017).
37. Kumar, R. *et al.* Simulations over South Asia using the Weather Research and Forecasting model with Chemistry (WRF-Chem): Chemistry evaluation and initial results. *Geosci. Model Dev.* **5**(3), 619–648. <https://doi.org/10.5194/gmd-5-619-2012> (2012).
38. Sanjay, J., Krishnan, R., Shrestha, A. B., Rajbhandari, R. & Ren, G. Y. Downscaled climate change projections for the Hindu Kush Himalayan region using CORDEX South Asia regional climate models. *Adv. Clim. Change Res.* **8**(3), 185–198. <https://doi.org/10.1016/j.accre.2017.08.003> (2017).
39. Kumar, R. *et al.* What controls the seasonal cycle of black carbon aerosols in India?. *J. Geophys. Res.* <https://doi.org/10.1002/2015JD023298> (2015).
40. Girach, I. A. *et al.* Variations in O₃, CO, and CH₄ over the Bay of Bengal during the summer monsoon season: Shipborne measurements and model simulations. *Atmos. Chem. Phys.* <https://doi.org/10.5194/acp-17-257-2017> (2017).
41. Chutia, L. *et al.* Distribution of volatile organic compounds over Indian subcontinent during winter: WRF-chem simulation versus observations. *Environ. Pollut.* <https://doi.org/10.1016/j.envpol.2019.05.097> (2019).
42. Arcomano, T. *et al.* A machine learning-based global atmospheric forecast model. *Geophys. Res. Lett.* **47**(9), e2020GL087776. <https://doi.org/10.1029/2020GL087776> (2020).

43. Betancourt, C., Stomberg, T., Roscher, R., Schultz, M. G. & Stadler, S. AQ-Bench: A benchmark dataset for machine learning on global air quality metrics. *Earth Syst. Sci. Data* **13**(6), 3013–3033. <https://doi.org/10.5194/essd-13-3013-2021> (2021).
44. Amato, F., Guignard, F., Robert, S. & Kanevski, M. A novel framework for spatio-temporal prediction of environmental data using deep learning. *Sci. Rep.* **10**(1), 1–11. <https://doi.org/10.1038/s41598-020-79148-7> (2020).
45. Wang, J., Balaprakash, P. & Kotamarthi, R. Fast domain-aware neural network emulation of a planetary boundary layer parameterization in a numerical weather forecast model. *Geosci. Model Dev.* **12**(10), 4261–4274. <https://doi.org/10.5194/gmd-12-4261-2019> (2019).
46. Davenport, F. V. & Diffenbaugh, N. S. Using machine learning to analyze physical causes of climate change: A case study of U.S. midwest extreme precipitation. *Geophys. Res. Lett.* **48**(15), e2021GL093787. <https://doi.org/10.1029/2021GL093787> (2021).
47. Ratnam, J. V., Dijkstra, H. A. & Behera, S. K. A machine learning based prediction system for the Indian Ocean Dipole. *Sci. Rep.* **10**(1), 284. <https://doi.org/10.1038/s41598-019-57162-8> (2020).
48. Sayeed, A. *et al.* A novel CMAQ-CNN hybrid model to forecast hourly surface-ozone concentrations 14 days in advance. *Sci. Rep.* **11**(1), 10891. <https://doi.org/10.1038/s41598-021-90446-6> (2021).
49. Tanimoto, H. *et al.* Direct assessment of international consistency of standards for ground-level ozone: Strategy and implementation toward metrological traceability network in Asia. *J. Environ. Monit.* **9**(11), 1183–1193. <https://doi.org/10.1039/B701230F> (2007).
50. Dee, D. P. *et al.* The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137**(656), 553–597. <https://doi.org/10.1002/qj.828> (2011).
51. Inness, A. *et al.* The CAMS reanalysis of atmospheric composition. *Atmos. Chem. Phys.* **19**(6), 3515–3556. <https://doi.org/10.5194/acp-19-3515-2019> (2019).
52. Girach, I. A., Tripathi, N., Nair, P. R., Sahu, L. K. & Ojha, N. O₃ and CO in the South Asian outflow over the Bay of Bengal: Impact of monsoonal dynamics and chemistry. *Atmos. Environ.* **233**, 117610. <https://doi.org/10.1016/j.atmosenv.2020.117610> (2020).
53. Kunchala, R. K. *et al.* On the understanding of surface ozone variability, its precursors and their associations with atmospheric conditions over the Delhi region. *Atmos. Res.* **258**, 105653. <https://doi.org/10.1016/j.atmosres.2021.105653> (2021).
54. Singh, A. *et al.* Impact of increasing carbon dioxide on dinitrogen and carbon fixation rates under oligotrophic conditions and simulated upwelling. *Limnol. Oceanogr.* <https://doi.org/10.1002/lno.11795> (2021).
55. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. <https://doi.org/10.1145/2939672.2939785>. (2016).

Acknowledgements

We gratefully acknowledge the open source library, XGBoost algorithm, developed by the contributors from DMLC/XGBoost community (<https://github.com/dmlc/xgboost/blob/master/CONTRIBUTORS.md>). ERA-Interim meteorological data from ECMWF (European Center for Medium range Weather Forecasting; <https://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/>); and CAMS data from <https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-global-reanalysis-eac4?tab=form> used in the study are highly acknowledged. We also acknowledge the Atmospheric Trace gases-Chemistry, Transport and Modelling (AT-CTM) project under the Geosphere Biosphere Programme of the Indian Space Research Organisation (ISRO-GBP). We are thankful to D. Pallamraju, S. Suresh Babu, Radhika Ramachandran and Anil Bhardwaj for valuable support during the study. The constructive comments and suggestions from the two anonymous reviewers and the handling editor are greatly appreciated.

Author contributions

N.O. and I.G. conceived the idea and performed the modeling. K.S. and I.G. conducted the ozone measurements. N.O., I.G., and A.S. performed the analyses with inputs from N.S. and S.S.G. N.O. wrote the manuscript with contributions from all the co-authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-01824-z>.

Correspondence and requests for materials should be addressed to N.O. or I.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021