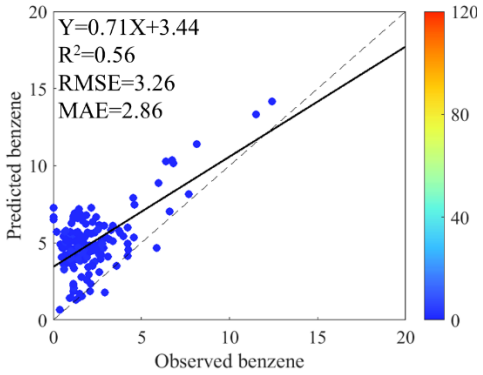
 # Reviewer 1

The authors used CTMs and ensemble machine-learning models to assess the impact of COVID-19 lockdown on ambient benzene. Overall, the manuscript is well-written and many useful information has been obtained. I think the manuscript falls into the scope of ACP. However, the manuscript still shows some minor flaws. I recommend the manuscript for publication on ACP when these issues have been adequately addressed.

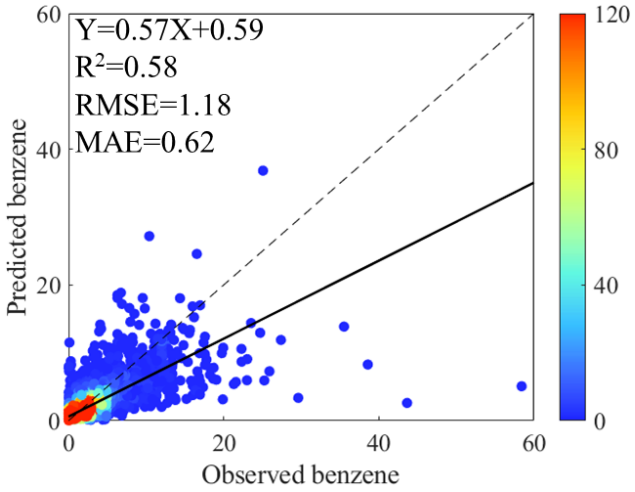
Response: Thank for reviewer's suggestions. We have revised the manuscript carefully based on reviewer's suggestions. Especially, we have added more discussion about the impact of COVID-19 lockdown on benzene variation. Besides, we also have rewritten the conclusion. The detailed revisions are shown in the revised version.

Comment 1: The sampling sites of ambient benzene focused on the United Sates, India, and Europe, while other regions lack of monitoring sites. How could you ensure the reliability of simulation results?

Response: Thank for reviewer's suggestions. Indeed, regular monitoring sites for ambient benzene observation only focused on Europe, India, and the United States. The long-term ambient benzene dataset in China and many other countries were not open access. We only obtained the short-term ambient benzene observations (during COVID-19 period) from few monitoring sites (e.g., Tangshan and Shanghai) and references to validate the performance of the ensemble model in China. The result suggested that R^2 value reached 0.56 based on the observed and simulated benzene concentrations in China, which was in good agreement with the CV R^2 value (0.60). Therefore, we believed that the modelling performance of this ensemble model was satisfied in some countries lack of monitoring sites.



Besides, we have added the spatial transferability test in the revised version. In order to examination the spatial transferability of the ensemble model, the site-based validation was performed. In each round, two-thirds of the dataset in India, Europe, and the United States were applied to train the model and the remained one was utilized to validate the model (e.g., India+Europe for training and the United States for testing). After three rounds, all of the simulated benzene concentrations were compared with the corresponding observed values. As shown in Figure S4, the out-of-bag R^2 value reached 0.58, which was slightly lower than the R^2 value (0.60) of training model. In addition, RMSE and MAE of the fitting equation for the out-of-bag data were 1.18 and 0.62, respectively. The result was in good agreement with those based on CV database, indicating the ensemble model showed satisfied spatial generalization. Although we cannot collect the benzene observations in some countries of Africa and South America, good spatial transferability confirmed that the simulation results were still reliable in some regions lack of ground observations.



Comment 2: To the best of my knowledge, many other decision tree models and deep learning models except RF, XGBoost, and LightGBM have been developed in recent years. Why do not you use other state-of-art models?

Response: Thank for reviewer's suggestions. Indeed, decision tree models did not only include RF, XGBoost, and LightGBM models, some other models such as CatBoost and GBDT were also applied to simulate the concentrations of air pollutants. Overall, these decision tree models were not sensitive to hyperparameters and the performances of these models remained relatively stable. Moreover, these models did not show marked difference in predictive accuracy. Therefore, we preferred to use some classical decision tree models such as RF, XGBoost, and LightGBM, which has been widely used by previous studies (Wei et al., 2019 RSE; Li et al., 2021 AE; Zhong et al., 2021 NSR). Compared with RF, XGBoost, and LightGBM, CatBoost and GBDT showed relatively low calculation efficiency and consumed large calculation resource. Furthermore, GBDT model slightly overestimated the higher values, while it underestimated the lower values. Therefore, we used the ensemble model of RF, XGBoost, and LightGBM to train the model.

Comment 3: Line 188-189: Why do you use some date variables such as month of year (MOY), and day of year (DOY) to remove the impact of meteorology?

Response: Thank for reviewer's suggestions. (Line 201-216) It is well known that the air pollutants were affected by emission and meteorology simultaneously. In fact, we hope to use emission inventory alone to reflect the emission contribution during COVID-19 period. However, daily global emission inventory dataset still shows large uncertainties. Meanwhile, the date variables including MOY and DOY could also reveal the impact of emission on air quality during COVID-19 period because the lockdown intensity was closely associated with the date. Therefore, we used the date variables coupled with emission inventory reflect the impact of lockdown on emission source. The final results based on XGBoost model also suggested that the use of date variables to decouple emission and meteorology was still reliable (R^2 = 0.65).

Comment 4: Line 200-218: The health risk assessment method suffers from many disadvantages. The ambient benzene derived from different sources generally showed distinct toxicity weights. I recommend the authors consider the difference in the model, which might be more valuable.

Response: Thank for reviewer's suggestions. We have reviewed many previous references and tried to quantify benzene-related health risk derived from different sources. Unfortunately, the toxicity weight of ambient benzene derived from various sources were not quantified and thus we cannot estimate the potential health risk derived from different sources. In the future work, we must perform

relevant experiment to determine the toxicity weights of different sources, and then quantified the health effects derived from emission sources.

Comment 5: Line 305-307: What is the difference of P and P*?

Response: Thank for reviewer's suggestions. P represents the ambient benzene different before and after COVID-19 lockdown; while P* denotes the change ratio (before and after COVID-19 lockdown) of ambient benzene in 2020 compared with the same period in 2019. P* is a detrended indicator in our study.

Comment 6: Line 323-352: This part was too superficial and the authors should add more discussion in this paragraph.

Response: Thank for reviewer's suggestions. (Line 342-376) We have added more discussion in the revised version and analyzed the benzene changes in some typical cities around the world.

Comment 7: The conclusion seems to be the repeat of abstract and the authors should rewrite this part.

Response: Thank for reviewer's suggestions. (Line 424-452) We have rewritten the conclusion and added some limitations of this study in this part.

"*The drastic lockdown measures largely reduced the air pollutant emissions. The meteorology-normalized ambient benzene concentrations in China (-15.6%), India (-23.6%), Europe (-21.9%), and the United States (-16.2%) experienced dramatic decreases after COVID-19 outbreak. Furthermore, the decreasing ratios in these major regions during COVID-19 lockdown period were much higher than the same period in 2019, indicating the aggressive emission control measures efficiently decreased ambient benzene concentrations. Emission reductions from industrial activities and transportation were major drivers for the decreasing of ambient benzene level during lockdown period, while the relatively stable solvent use emission could restrict the further decrease of benzene pollution. Besides, the slight increase of domestic emission during this period might be an important reason for the benzene increase in some regions (e.g., Yunnan province). There is also an urgent need to control the household combustion and solvent use emissions apart from the emissions from industry and transportation sectors.*

Besides, substantial decreases of atmospheric benzene levels could save sufficient health benefits. Dramatic decreases of benzene emissions in Europe and the United States cannot save effective health benefits because the ambient benzene levels in both of these regions during business-as-usual scenario were significantly lower than the risk threshold. However, the benzene decreases in North China Plain (NCP), China and Bihar, India could save abundant health benefits because these regions often suffered from severe atmospheric benzene pollution during business-as-usual scenario. Thus, more targeted abatement measures are needed to reduce the benzene emission in these areas. For instance, the stricter industrial and vehicle emission standards for VOC control should be implemented in China and India. Moreover, some measures including limiting the amount of coal-fired power plants, adding environmentally friendly cars and clean fuels for vehicles and vessels, and strengthening the labeling system for vehicles in use should be strengthened.

It should be noted that our study still suffered from some limitations. First of all, the monitoring sites were not evenly distributed around the world, and thus the simulation result might show the higher uncertainty in the regions lack of monitoring sites. Besides, the GEOS-Chem model still suffered from some uncertainties due to imperfect chemical mechanism and inaccurate emission inventory. In the future work, the model should be further improved." has been added in the revised version.

Comment 8: The reference format is not standard and the authors should revise carefully.

Response: Thank for reviewer's suggestions. We have significantly revised the reference format based on reviewer's suggestions.

 # Reviewer 2

Ling et al. used GEOS-Chem coupled with machine-learning models to predict the ambient benzene level before and after COVID-19 lockdown. Many studies have analyzed the impacts of COVID-related anthropogenic emission on regional air quality. It is a really interesting topic since there are few studies looking at the responses of global atmospheric benzene to COVID-19 lockdown. However, the manuscript still showed some major flaws especially in the model test and discussion, which should be addressed first.

Response: Thank for reviewer's suggestions. We have significantly revised the manuscript based on reviewer's suggestions carefully. The detailed comments are as follows:

Comment 1: The abstract includes too many results rather than the important findings. Thus, the important implications should be condensed in the abstract. I suggest the authors should reorganize the abstract.

Response: Thank for reviewer's suggestions. (Line 15-32) We have rewritten the abstract and add many important findings in the revised version. One of the most important findings is the benzene pollution mitigation in China and India might bring out more health benefits around the world.

Comment 2: There are numerous studies focusing on modelling surface air pollutants like PM and polluted gases using machine learning models (especially those adopted in the current study) globally or regionally. Thus, the authors are suggested to summarize related studies in the Introduction.

Response: I agree with reviewer's suggestions. (Line 85-87) We have citied many previous studies about the air quality estimates using machine-learning models in the revised version. Wei et al. have performed many pioneering works in this field, and thus we have cited their references in the revised version.

Comment 3: Line 58: How about the global or regional (like in China) O_3 and aerosol precursors (e.g., SO_2, CO) changes during the COVID-19? The authors are also suggested to discuss since only PM and NO_2 mentioned here.

Response: Thank for reviewer's suggestions. (Line 60-64) We have also added some studies about other air pollutant (e.g., $PM_{2.5}$, SO_2, NO_2, and O_3) changes during COVID-19.

Comment 4: Line 60-67: Some field measurement of ambient benzene in China or Europe during COVID-19 period should be introduced. There should be several studies that have analyzed the temporal variation of ambient benzene in Chinese cities before and after lockdown.

Response: Thank for reviewer's suggestions. (Line 70-74) We have added some studies about benzene variations in China and Europe during COVID-19 period.

Comment 5: Line 82-85: Why do you use the GEOS-Chem coupled with machine-learning models to decouple the emission and meteorology contributions? The GEOS-Chem model could also distinguish the emission and meteorology contributions. Are there any differences or advantages? Please clarify.

Response: Thank for reviewer's suggestions. Indeed, GEOS-Chem model could distinguish the emission and meteorology contributions. Unfortunately, the simulation results derived from GEOS-Chem model alone still showed the higher uncertainty due to the uncertainty of emission inventory and the incomplete chemical mechanisms. In order to improve the modelling performance, we

integrated the GEOS-Chem output into the ensemble machine-learning model. Then, the XGBoost model was applied to isolate the emission and meteorology contributions to ambient benzene.

Comment 6: The specific lockdown time in different regions (e.g., China, India, and United States) should be introduced in the methods.

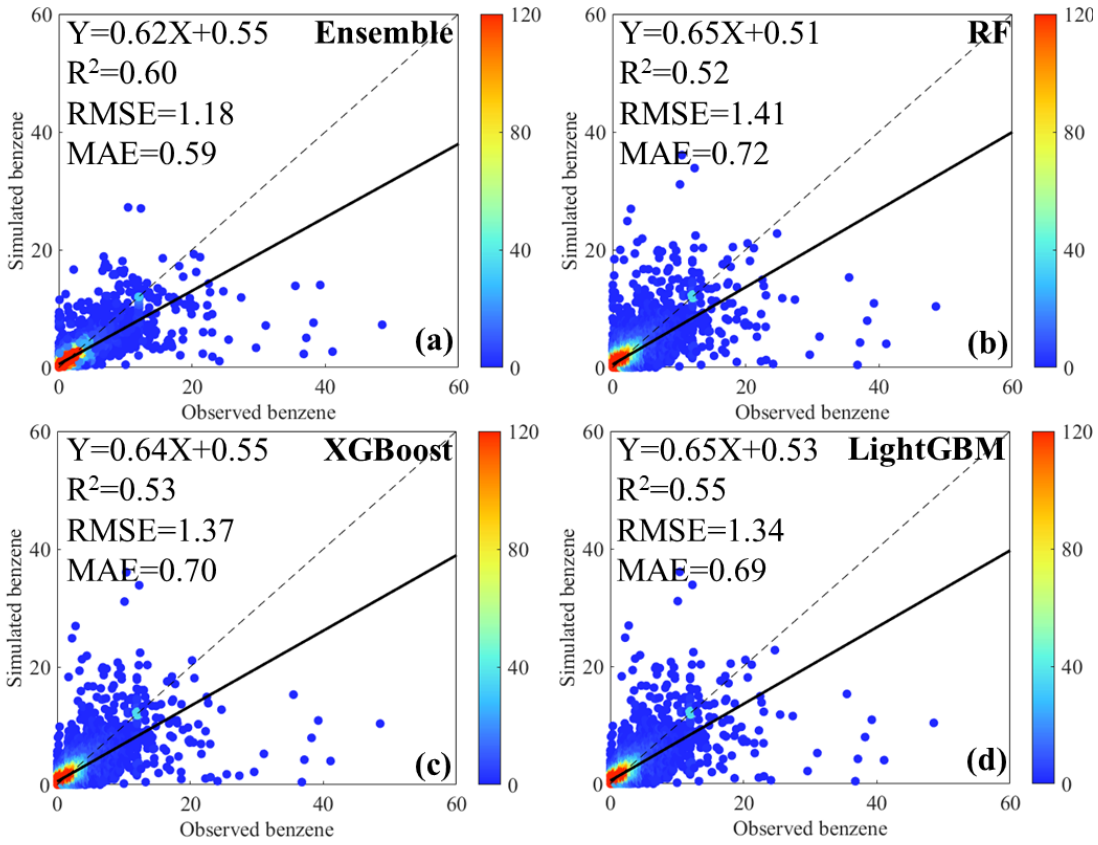
Response: I agree with reviewer's suggestions. (Line 114-118) The start date of COVID-19 lockdown in China was January 23th and the national lockdown lasted for about one month. However, the deblocking date in Wuhan was April 8th. The start and end dates of lockdown in India were March 25th and April 25th, respectively. The lockdown in the United States occurred firstly in California in March 19th, and then the lockdown lasted for 1-2 months. The lockdown dates in most European countries lasted from March to May.

Comment 7: Line 103-104: How about the data quality in India? The authors should add some data quality assurance about benzene dataset in India. Besides, the data assurance in other regions should be also added.

Response: Thank for reviewer's suggestions. (Line 122-128) The CPCB database provides data quality assurance (QA) or quality control (QC) programs by establishing strict procedures for sampling, analysis, and calibration (Gurjar et al., 2016). The ground-level benzene observations in Europe and the United States were collected from air quality data portal of the European Environment Agency (EEA) and United States Environmental Protection Agency (EPA), respectively. Furthermore, only days with more than 12 h of available data are included in the analysis.

Comment 8: Why do you use the ensemble model to predict benzene level? Please compare and show the advantage of the ensemble model compared with individual one.

Response: Thank for reviewer's suggestions. We have evaluated the performances of ensemble model, RF, XGBoost, and LightGBM, respectively. The result suggested that the CV R^2 values followed the order of ensemble model (0.60) > LightGBM (0.55) > XGBoost (0.53) > RF (0.52). Furthermore, both of the root-mean-square error (RMSE) (1.18 $\mu g/m^3$) and the mean absolute error (MAE) (0.59 $\mu g/m^3$) of the ensemble model were significantly lower than those of RF (RMSE and MAE: 1.41 and 0.72 $\mu g/m^3$), XGBoost (RMSE and MAE: 1.37 and 0.70 $\mu g/m^3$), and LightGBM (RMSE and MAE: 1.34 and 0.69 $\mu g/m^3$). Therefore, we selected to use the ensemble model to predict global benzene level.

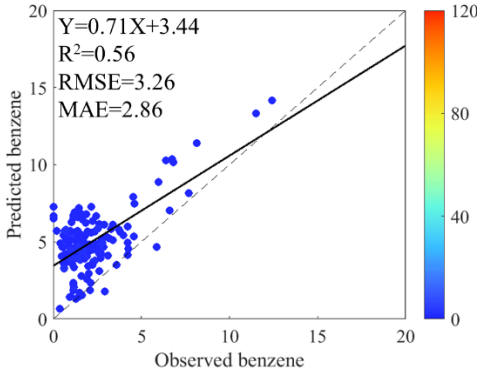


Comment 9: Line 178: Why do you use 5-fold CV test instead of 10-fold test? The later one is the most commonly used one.

Response: Thank for reviewer's suggestions. In our study, we only used the 5-fold CV test to train the submodel, and the final model still uses 10-fold CV test. For the training of submodels, we used 5-fold CV test instead 10-fold CV test to save calculation resource.

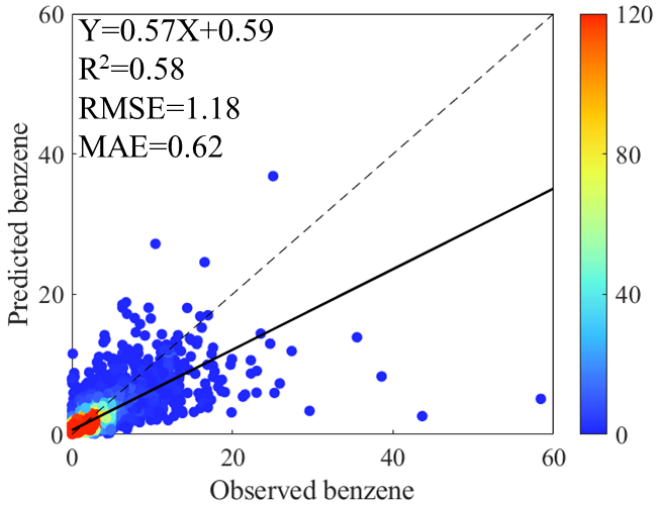
Comment 10: The monitoring sites only located in Europe, India, and the United States, but no site is available in China. This could lead to larger uncertainties in China. How did the authors resolve this issue?

Response: Thank for reviewer's suggestions. Indeed, regular monitoring sites for ambient benzene observation only focused on Europe, India, and the United States. The long-term ambient benzene dataset in China was not open access. We only obtained the short-term ambient benzene observations (during COVID-19 period) from few monitoring sites (e.g., Tangshan and Shanghai) and references to validate the performance of the ensemble model in China. The result suggested that R^2 value reached 0.56 based on the observed and simulated benzene concentrations in China, which was in good agreement with the CV R^2 value (0.60). Therefore, we believed that the modelling performance of this ensemble model was satisfied in China though no substantial observation data in China was integrated into the model.



Comment 11: Section 3.1: The authors must add the spatial transferability test in this part to confirm the robustness of the ensemble model.

Response: I agree with reviewer's suggestions. (Line 262-271) In fact, we have performed the spatial transferability test in the original version. In order to examination the spatial transferability of the ensemble model, the site-based validation was performed. In each round, two-thirds of the dataset in India, Europe, and the United States were applied to train the model and the remained one was utilized to validate the model (e.g., India+Europe for training and the United States for testing). After three rounds, all of the simulated benzene concentrations were compared with the corresponding observed values. As shown in Figure S4, the out-of-bag R^2 value reached 0.58, which was slightly lower than the R^2 value (0.60) of training model. In addition, RMSE and MAE of the fitting equation for the out-of-bag data were 1.18 and 0.62, respectively. The result was in good agreement with those based on CV database, indicating the ensemble model showed satisfied spatial generalization.



Comment 12: Line 250: What does out-of-bag R^2 mean? In fact, out-of-bag refers to out-of-sample. Do you mean out-of-station/site?

Response: Thank for reviewer's suggestions. The out-of-bag R^2 value denotes the modelling performance of samples not used for training in 10-fold cross validation. For 10-fold cross validation, 10% of dataset was not used to train the model for each round. After 10 rounds, the dataset could be compared with the corresponding observation dataset and then to obtain the R^2 value. This part means the out-of-bag dataset. In our study, the out-of-bag refers to out-of-sample.

Comment 13: Line 356-357: Too many decimal places are meaningless.

Response: I agree with reviewer's suggestions. (Line 381) We have corrected the error in the revised version.

Comment 14: Section 3.2: The discussion is too general and more detailed reasons for benzene change in different cities should be introduced.

Response: Thank for reviewer's suggestions. We have added more discussion in the revised version and analyzed the benzene changes in some typical cities around the world (section 3.2).

Comment 15: Line 353-360: the paragraph is too simple. The impact of each meteorological parameter should be discussed in this paragraph.

Response: Thank for reviewer's suggestions. (Line 385-396) We have assessed the impact of many meteorological parameters on ambient benzene. Especially, the air temperature plays the most important role on the ambient benzene. The importance of other meteorological parameters is relatively low compared with air temperature. We have rewritten this part in the revised version.

Comment 16: The environmental implications of this study should be condensed in the conclusions. How can we control the ambient benzene pollution around the world?

Response: Thank for reviewer's suggestions. We have added some concrete control measures for benzene pollution around the world. (Line 448-456) "*For instance, the stricter industrial and vehicle emission standards for VOC control should be implemented in China and India. Moreover, some measures including limiting the amount of coal-fired power plants, adding environmentally friendly cars and clean fuels for vehicles and vessels, and strengthening the labeling system for vehicles in use should be strengthened.*" has been added in the revised version.

Comment 17: There are many grammar errors and thus the English throughout the manuscript should be further edited carefully.

Response: Thank for reviewer's suggestions. We have corrected these grammar errors throughout the manuscript.