Manuscript No.: acp-2022-550

Responses to Reviewer #1

This manuscript investigated future climate change impacts on near-surface O_3 concentrations over Asia from 2020-2100 using a machine learning model along with multisource data. The random forest model was trained based on results from global atmospheric chemical transport model simulations, real-time O_3 observations, and other datasets. Future climate-driven changes in O_3 concentrations were predicted by the trained model together with 18 CMIP6 multi-model simulations under four future scenarios. The paper found that future climate change would aggravate O_3 pollution in Asia and expanded the pollution from North China to South China and extended it into the cold season in a warming future. Overall, this is a good example of machine learning and big data analysis in atmospheric science. The results are of good significance and novelty. The manuscript was well-written and properly organized. Therefore, I recommend the acceptance of the manuscript.

We thank the reviewer for all the insightful comments. Below, please see our point-by-point response (in blue) to the specific comments and suggestions and the changes that have been made to the manuscript, in an effort to take into account all the comments raised here.

General:

1, In this study, the authors trained the machine learning (ML) model using O_3 precursor emissions, but did not consider the role of methane. So, I suggest the authors discuss more about how the results will be affected if the role of methane is included in the ML model prediction.

Response:

The treatment of methane in GEOS-Chem model are different from other precursor emissions. Its concentrations are directly prescribed as a surface boundary condition. Therefore, we did not consider the role of methane in machine learning model. However, the climate influence of methane is considered in the CMIP6 multi-model simulations. Therefore, the impact of future methane concentrations on O_3 via climate change is included in machine learning projections. We have now added a discussion in the revised manuscript. “Although the methane concentrations in GEOS-Chem model are prescribed and its role in the O_3 production is not considered in the ML model, the climate influence of methane is included in the CMIP6 multi-model simulations. Consequently, the impact of future changes in methane on O_3 concentrations via climate change are considered in the future projections.”

Minor:

1. Line 42: Change “primary” to “secondary”.

Response:

Revised as suggested.

2. Lines 67-74: How is performance of model in predicting ozone pollution?

Response:

Xue et al. (2020) and Wei et al. (2022) estimated maximum daily 8 h averaged (MDA8) near-surface O_3 concentrations across China using the random forest model together with multisource data, and the coefficient of determination (R^2) between the observed and predicted O_3 concentrations reached 0.70 and 0.87, respectively. Di et al. (2017) applied the neural network to simulate MDA8 O_3 concentrations in the continental United States with a mean cross-validation R^2 value of 0.76. Su et al. (2020) adopted a support vector machine model to predict summer hourly O_3 concentrations in Nanjing, and the R^2 reached 0.78. Liu et al. (2020) established a nationwide MDA8 O_3 prediction model based on the extreme gradient boosting approach, and reported a R^2 value of 0.78. Liu et al. (2022) used cluster-enhanced ensemble machine learning method to construct a global monthly near-surface O_3 dataset from 2003 to 2019, which showed an excellent overall R^2 of 0.91. We have now added a note at the end of the paragraph as follows: “The abovementioned previous studies utilizing the ML methods showed high computational efficiency and accuracy, with an overall R^2 between the observed and predicted O_3 concentrations in the range of 0.7–0.9.”

3. Line 172: Has this dataset been quality controlled or has been used in scientific studies?

Response:

The Chine Ministry of Ecology and Environment (MEE) categorized a total of 360 key cities by 2020. Each city covers several O_3 monitoring sites, and we average them hourly at city level. Since the observed O_3 concentrations reported by MEE were under the standard conditions of temperature of 273 K and air pressure of 1013 hPa until 31 August 2018, we adjusted the post-August 2018 O_3 concentrations under the ambient state of 298 K and 1013 hPa to that under the reference state.

The dataset has been widely used to examine O_3 pollution over China in previous studies (Li et al., 2020, 2021; Qian et al., 2022). We have now revised the descriptions in the manuscript.

4. Line 217: Please describe more about the four scenarios.

Response:

Thanks for suggestion. We have now added more description about future scenarios that "Monthly meteorological parameters under four different future climate scenarios, including SSP1-2.6, SSP2-4.5, SSP3-7.0 and SSP5-8.5 (a representation of low, intermediate, medium to high, and high forcing levels, respectively), are fed to a ML model to predict near-surface O_3 concentrations."

5. Line 231: How the meteorological fields be adjusted?

Response:

To minimize the impact of inconsistency in meteorology between CMIP6 models and the GEOS-Chem model, the CMIP6 meteorological variables in 2020–2100 are adjusted by their potential bias, characterized as the difference in their historical climatological mean (2014–2019) and MERRA-2. This adjustment process also removes inconsistency in the initial conditions of meteorological fields used by different CMIP6 models. We have added the descriptions in the manuscript.

6. Usually ground-level O_3 concentrations (225, 353 lines), Surface O_3 concentrations (271 lines), near-Surface O_3 concentrations (24 lines and so on) were included in the article, how do these concentrations differ? If all of them refer to the same meaning, please use the same name.

Response:

Thanks for suggestion. We have now updated all the O_3 concentrations as near-surface O_3 concentrations.

7. As a lot of methods and data are used, I suggest to draw a structural block diagram to more clearly express the structure and ideas of this article, including how to assimilate.

Response:

We thank the reviewer for this suggestion. We have now added a structural block diagram in the revised manuscript.

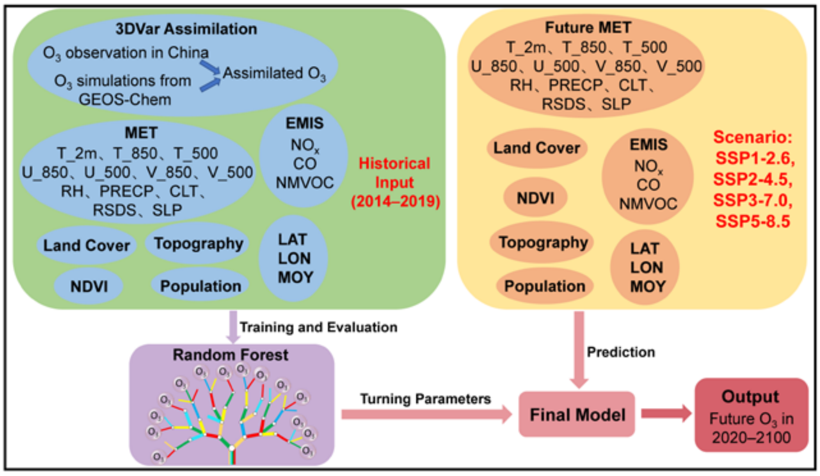

Figure 1. The structure and specific schematics for predicting future near-surface O_3 concentrations under four scenarios based on the machine learning (ML) method.

8. How are Importance scores of independent variables (meteorological parameters, emissions, land use, topography, and population density) used in the ML model for predicting future near-surface O_3 concentrations over Asia obtained?

Response:

In this study, the importance score of each variable is obtained from the training of machine learning model. To retrieve the relative importance scores of each input feature from the Random Forest model for predicting near-surface O_3 over Asia, we use the "feature_importances" attribute in Scikit-learn collection, which called Gini importance. A greater Gini importance implies a greater influence of input feature on the target variable. We have now elaborated the importance scores more in the revised manuscript.

The measurement is based on the calculation of Gini impurity when a variable is chosen to split a node. For each variable, the sum of the Gini decrease across every tree of the forest is accumulated every time that variable is chosen to split a node, and then averaged over all trees of the ensemble, expressed as:

$$\mathrm{GI}(\omega) = \sum_{n=1}^{N} \omega_n (1-\omega_n) = 1 - \sum_{n=1}^{N} \omega_n^2$$

where *n* represents the number of the categories (*N* = 1, ..., *n*) and ω_n represents the sample weight of each category. The importance of one feature (X_j) on node *m* is that the GI changes before and after node *m* branching as follows:

$$\Delta\mathrm{GI}_{jm} = \mathrm{GI}_m - \mathrm{GI}_l - \mathrm{GI}_r$$

where GI_l and GI_r represent the GI of two new nodes after branching. The importance score for one feature (IS_j) in the then extra trees with *k* trees (*i* = 1, ..., *k*) is calculated as:

$$\mathrm{IS}_j = \sum_{i=1}^{k} \Delta\mathrm{GI}_{ij} = \sum_{i=1}^{k} \sum\nolimits_{m \in M} \Delta\mathrm{GI}_{jm}$$

Where ΔGI_{ij} represents the importance of X_i in the ith tree when the node of feature X_i in decision tree *j* belongs to set *M*. Finally, an additional normalization approach is performed to all obtained importance scores for each feature.

9. The authors could add some discussion and some references. such as:

Li, M., et al. (2021). Rising surface ozone in China from 2013 to 2017: A

response to the recent atmospheric warming or pollutant controls? AE, 246. doi:10.1029/2021JD036393, https://doi.org/10.1016/j.scitotenv.2021.150338.
Lu, X., et al. (2020). Rapid Increases in Warm-Season Surface Ozone and Resulting Health Impact in China Since 2013. EST&L, 7, 240-247.

Response:

Thanks for the good suggestion. We have added the references and discussion as follows. "Last but not least, the near-surface O_3 have increased rapidly in China since 2013 owing to both precursor emission changes and atmospheric warming (Li M. et al., 2021), which significantly affect human health (Lu et al., 2020) and also requires further studies."

References:

Li, K., Jacob, D. J., Shen, L., Lu, X., De Smedt, I., and Liao, H.: Increases in surface ozone pollution in China from 2013 to 2019: anthropogenic and meteorological influences, Atmos. Chem. Phys., 20, 11423–11433, https://doi.org/10.5194/acp-20-11423-2020, 2020.

Li, K., Jacob, D. J., Liao, H., Qiu, Y., Shen, L., Zhai, S., Bates, K. H., Sulprizio, M. P., Song, S., Lu, X., Zhang, Q., Zheng, B., Zhang, Y., Zhang, J., Lee, H. C., and Kuk, S. K.: Ozone pollution in the North China Plain spreading into the late-winter haze season, P. Natl. Acad. Sci. USA, 118, 1–7, https://doi.org/10.1073/pnas.2015797118, 2021.

Li, M., Wang, T., Shu, L., Qu, Y., Xie, M., Liu, J., Wu, H., and Kalsoom, U.: Rising surface ozone in China from 2013 to 2017: A response to the recent atmospheric warming or pollutant controls?, Atmos. Environ., 246, 118130, https://doi.org/10.1016/j.atmosenv.2020.118130, 2021.

Lu, X., Zhang, L., Wang, X., Gao, M., Li, K., Zhang, Y., Yue, X., and Zhang, Y.: Rapid Increases in Warm-Season Surface Ozone and Resulting Health Impact in China since 2013, Environ. Sci. Technol. Lett., 7, 240–247, https://doi.org/10.1021/acs.estlett.0c00171, 2020.

Qian, J., Liao, H., Yang, Y., Li, K., Chen, L., and Zhu, J.: Meteorological influences on daily variation and trend of summertime surface ozone over years of 2015–2020: Quantification for cities in the Yangtze River Delta, Sci. Total Environ., 834, 155107, https://doi.org/10.1016/j.scitotenv.2022.155107, 2022.

Manuscript No.: acp-2022-550

Responses to Referee #2

The authors demonstrate a framework of using machine learning (ML) to project long-term (2020–2100) surface ozone levels over Asia. The machine learning algorithm (random forest, RF) is trained with ozone data from 2014 to 2018, along with data of meteorology, emissions and other auxiliary data. The trained RF is then used to make ozone projections based on meteorological fields from the four climate scenarios (i.e., SSP1-2.6, SSP2-4.5, SSP3-7.0 and SSP5-8.5) of CMIP6.

This study adopts a data assimilation approach that combines simulations from chemical transport model and observations to better represent real-world ozone levels. This manuscript is within the scope of ACP and has a good scientific quality. I suggest that this manuscript is accepted after the authors address my comments below.

We thank the reviewer for all the insightful comments. Below, please see our point-by-point response (in blue) to the specific comments and suggestions and the changes that have been made to the manuscript, in an effort to take into account all the comments raised here.

Specific comments:

In section 2.3, a more detailed description of data assimilation approach should be added to the main text for readers to follow. As a minimum, the authors should include some citations for this section.

Response:

Thanks for the suggestion. We have now detailed the assimilation system in the revised paper as below:

"The assimilation system, which is used to combine the O_3 observations across China with results from GEOS-Chem simulations, is based on a three-dimensional variational (3DVar) data assimilation (Kalnay, 2003; Evensen et al., 2022). The goal of the 3DVar is to find the maximum likelihood estimation of a state vector x, which is the O_3 concentrations here in this study, given the available observations y through minimizing the cost function:

$$J(x) = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^b)^{\mathrm{T}}\, \mathbf{B}^{-1}\, (\boldsymbol{x} - \boldsymbol{x}^b) \; + \frac{1}{2}(\boldsymbol{y} - \mathrm{H}(\boldsymbol{x}))^{\mathrm{T}}\, \mathbf{O}^{-1}\, (\boldsymbol{y} - \mathrm{H}(\boldsymbol{x})\,)$$

Here $\boldsymbol{x}^b$ represents the priori simulation. **B** is the empirical background covariance matrix formulated as a product of the uncertainty in the simulated value and a distance-based correlation matrix **C**, and the individual element is calculated as:

$$\mathbf{B}_{i,j} = 0.2 * \boldsymbol{x}_i^b * 0.2 * \boldsymbol{x}_j^b * \mathbf{C}_{i,j}$$

Here we have used 20% choice to characterize uncertainty of the O_3 simulation, the correlation matrix is empirically set as:

$$\mathbf{C}_{i,j} = e^{-(\frac{d_{i,j}}{200km})^2/2}$$

Here $d_{i,j}$ represents the spatial distance between the grid cell i and j.

H denotes the linear observation operator that converts the simulation results into the observational space. Here all observations are assumed to be independent, and therefore **O** is a diagonal covariance matrix storing the square of the observation uncertainty, which is also set as 20% similarly."

In section 4, the authors mention that one of the limitations to this study is in only using observations across China for the data assimilation. I recommend that the authors also highlight this limitation in section 2.3. For instance, in lines 191 to 193, uncertainties of GEOS-chem simulation are only minimized in China.

Response:

We thank the reviewer for this suggestion. We have now made modifications in the revised manuscript accordingly to emphasize the limitation, as "…, suggesting that the assimilated data have an excellent representation of O_3 observations and minimize the uncertainties of GEOS-Chem simulations in China.".

The sentence from lines 196 to 198 appears to suggest that all of the ozone concentrations from the study domain have been assimilated. I don't think this is the case for regions outside of China. I would suggest the authors to be more specific. For example, how have the ozone concentrations from outside of China been processed? Are these directly from the simulation of GEOS-chem?

Response:

Thank you for the comment and suggestion. We have made the statement clear in the manuscript as follows: "In this study, a random forest (RF) model is used to predict O_3 concentrations, similar to our previous studies (Li H. et al., 2021, 2022), with input data of assimilated O_3 concentrations in China that combine observations and results from GEOS-Chem model simulations, GEOS-Chem simulated O_3 concentrations outside of China, MERRA-2 meteorological variables, O_3 precursor emissions, land cover (LC), normalized difference vegetation index (NDVI), topography (TOPO), population density (POP), and the month of the year (MOY) and geographic location of each model grid as spatiotemporal information."

In section 2.2, the observational network of CNEMC has an inconsistent number of observational sites through 2014–2019 as the number of sites has grown. Does the inconsistency of sites affect data assimilation? How the authors handle this potential issue?

Response:

The China Ministry of Ecology and Environment (MEE) categorized a total of 360 key cities by 2020. Each city covers several O_3 monitoring sites, and we average them hourly at city level. In this study, we assessed the changes of observed O_3 concentrations from MEE in 360 cities across China during 2014–2019, which would not affect the data assimilation. We have now clearly stated it in section 2.2 as follows: "In this study, the quality controlled hourly O_3 observations in 360 cities are averaged within each 0.5° latitude × 0.625° longitude gride of the GEOS-Chem model."

In section 2.4, could the authors give the ranges of the hyperparameters used in the tuning during cross validation and the final selected hyperparameters? Besides, Is the whole set of training data (i.e., all data from 2014-2018 over Asia) randomly split into 10-folds for the cross validation? If in this case, why does the caption of Fig. 2 indicate that the 10-fold cross-validation results are from the year 2019? I suggest that the authors clarify this and give more information regarding the cross-validation process. Moreover, I am concerned that spatial autocorrelation may exist in the cross-validation because of the random split of the training data. For instance, a grid kept for training while the adjacent grid that shares high similarity with this grid is used for validation. This may violate the assumption of data independence. See Ploton et al. (2020) (https://doi.org/10.1038/s41467-020-18321-y) that is relevant to the spatial autocorrelation issue.

Response:

Thanks for the comment. We tuned the n_estimators (the number of decision trees in the forest) from 50 to 250 with interval of 50 and min_samples_split (the minimum of samples required to split a node) from 2 to 8 with interval of 2. The grid search and 10-fold cross-validation were applied to tune the hyperparameters. We found that changes of hyperparameters have a little impact on the performance of RF model. In this study, the best hyperparameters (n_estimators=200, min_samples_split=2, max_features= "sqrt", bootstrap= "True") of the RF model are utilized. We have now added a note in the revised manuscript.

The training set from 2014 to 2018 was split into 10 folds for cross-validation, and the performance of the machine learning model is only determined by the testing data, which were not used at the training/validation stage. We have now revised the caption of Fig. 3 accordingly: "**Figure 3**. Density scatterplots of predicted vs assimilated monthly near-surface O_3 concentrations (ppb) in 2019 over Asia. The gray and red lines are the 1:1 line and linear regression line, respectively. Statistical metrics including the number of samples (N), correlation of determination (R^2, unitless), root mean square error (RMSE, ppb), mean absolute error (MAE, ppb), and mean relative error (MRE, %) are shown at the top left."

We agree with the reviewer about the issue of spatial autocorrelation in the raw data. We have now added the following comment to the discussion section associated with the uncertainties and limitations in the machine learning method. "Moreover, the spatial autocorrelation in random split of training data for cross-validation would lead to the overly optimistic statistics of ML model predictive power (Ploton et al., 2020)."

Same in section 2.4, variables such as month of the year (MOY) and geographical locations of model grids may not have actual physical meaning. I'm not sure why variables such of these are necessary. Could the authors provide some explanations?

Response:

Wei et al. (2019) applied the spatial-time random forest model to account for the spatiotemporal heterogeneity of $PM_{2.5}$ concentrations over China, with considering both the spatial heterogeneity and temporal variations of variables. The results showed the newly developed model performed better than the traditional random forest, which demonstrated that considering both geographical and temporal information would improve the model performance. Recent studies have widely used the spatiotemporal information as inputs to investigate air pollution based on machine learning method (e.g., Wei et al., 2020, 2022; Li et al., 2021, 2022; Gong et al., 2022). The O_3 concentrations also vary dramatically in space and time, thus we applied month of the year and geographical locations of the model grids as spatiotemporal information in this study.

It seems that the authors construct a single RF emulator to model ozone over the entirety of Asia. One of the advantages of using a single emulator is in the large size of the training data. However, a single emulator is not able to provide information about feature importance for any specific regions. For instance, humidity in southern China is more important, while temperature and solar radiation may be the key features in northern China (e.g., Weng et al., 2022) (https://doi.org/10.5194/acp-22-8385-2022). The importance scores in Fig. 4 can only reflect the overall importance of the features from the whole study domain, and the interpretation of these scores should be treated with caution. For instance, if the study domain covers more regions with humidity as the key feature for suppressing ozone production, it is likely that humidity is weighted to be more important than other features. I suggest the authors to address and discuss this limitation.

Response:

We thank the reviewer for this suggestion. We have added a sentence in the section 3.1 as follows. “However, it is noted that the O_3 variations in different regions are dominated by different meteorological factors (Weng et al., 2022). The importance score of each independent feature quantified in this study can only reflect the overall importance across Asia, which is less representative of any specific regions.”

We have also addressed this concern in the discussion section. “Additionally, the overall importance scores of the features in this study can only reflect that from the whole study domain. Further investigations are required to identify and quantify the importance score of each local variable contributed to the near-surface O_3 predictions in different specific regions.”

Minor and technical comments:

Line 77: Citation of Gong et al. (2019) should be replaced by Gong and Liao (2019). This should be consistent with the citation in Line 76.

Response:

Corrected.

Line 206: Mis-spelling of author name. It should be “Rodriguez”.

Response:

Corrected.

In the supplementary, I’m not sure whether Fig. S11 and Fig. S12 follow the same caption as Fig. S8. Are these still percentage differences (%) between 2020–2029 and 2091–2100?

Response:

Thanks for the note. The Fig. S11 and Fig. S12 are the spatial distributions of absolute difference (m/s) in the CMIP6 multi-model seasonal averaged wind fields at 850 hPa and 500 hPa between 2020–2029 and 2091–2100, respectively. We have now made corrections.

References:

Evensen, G., Vossepoel, F. C., and van Leeuwen, P. J.: Data Assimilation Fundamentals: A Unified Formulation of the State and Parameter Estimation Problem, Springer Nature, https://doi.org/10.1007/978-3-030-96709-3, 2022.

Gong, C., Wang, Y., Liao, H., Wang, P., Jin, J., and Han, Z.: Future co-occurrences of hot days and ozone polluted days over China under scenarios of Shared Socioeconomic Pathways predicted through a machine learning approach, Earth's Futur., 10, e2022EF002671, https://doi.org/10.1029/2022EF002671, 2022.

Li, H., Yang, Y., Wang, H., Li, B., Wang, P., Li, J., and Liao, H.: Constructing a spatiotemporally coherent long-term $PM_{2.5}$ concentration dataset over China during 1980–2019 using a machine learning approach, Sci. Total Environ., 765, 144263, https://doi.org/10.1016/j.scitotenv.2020.144263, 2021.

Li, H., Yang, Y., Wang, H., Wang, P., Yue, X., and Liao, H.: Projected Aerosol Changes Driven by Emissions and Climate Change Using a Machine Learning Method, Environ. Sci. Technol., 56, 7, 3884–3893, https://doi.org/10.1021/acs.est.1c04380, 2022.

Kalnay, E.: Atmospheric Modeling, Data Assimilation and Predictability, Cambridge University Press, Cambridge, United Kingdom, 2003.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., and Pélissier, R.: Spatial validation reveals poor predictive performance of large-scale ecological mapping models, Nat. Commun., 11, 1–11, https://doi.org/10.1038/s41467-020-18321-y, 2020.

Wei, J., Li, Z., Cribb, M., Huang, W., Xue, W., Sun, L., Guo, J., Peng, Y., Li, J., Lyapustin, A., Liu, L., Wu, H., and Song, Y.: Improved 1 km resolution $PM_{2.5}$ estimates across China using enhanced space–time extremely randomized trees, Atmos. Chem. Phys., 20, 3273–3289, https://doi.org/10.5194/acp-20-3273-2020, 2020.

Wei, J., Li, Z., Li, K., Dickerson, R. R., Pinker, R. T., Wang, J., Liu, X., Sun, L., Xue, W., and Cribb, M.: Full-coverage mapping and spatiotemporal variations of ground-level ozone (O_3) pollution from 2013 to 2020 across China. Remote Sens., Environ., 270, 112775, https://doi.org/10.1016/j.rse.2021.112775, 2022.

Weng, X., Forster, G. L., and Nowack, P.: A machine learning approach to quantify meteorological drivers of ozone pollution in China from 2015 to 2019, Atmos. Chem. Phys., 22, 8385–8402, https://doi.org/10.5194/acp-22-8385-2022, 2022.