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.