Dear editor,

Here we submit our revised manuscript for consideration to be published on **Atmospheric Chemistry and Physics**

The further information about our manuscript is as follows:

Topic: Developing a novel hybrid model for the estimation of surface 8-h ozone (O_3) across the remote Tibetan Plateau during 2005-2018

Type of Manuscript: article

Authors: Rui Li[a], Yilong Zhao[a], Ya Meng[a], Wenhui Zhou[a], Ziyu Zhang[a], Hongbo Fu[a-c] *

***Corresponding author:**

Hongbo Fu; Address: Department of Environmental Science and Engineering, Fudan University, Shanghai 200433, China; Tel.: (+86)21-5566-5189; Fax: (+86)21-6564-2080; Email: fuhb@fudan.edu.cn

Firstly, we acknowledge the suggestions of editor and two reviewers, and are also grateful to your efficient serving. We have updated the manuscript on the basis of these valuable comments. Our responses were listed as following:

Reviewer #2: This paper attempts to create a machine-learned statistical model to estimate 8-h surface ozone concentrations where few measurements exist (Tibetan plateau). The model formulation (RF-GAM) is interesting and the subject fits within the scope of ACP. There are several flaws with the presentation of this work, namely the lack of significance tests and alternate performance metrics in addition to English grammatical errors throughout. However, I am optimistic the authors can correct these issues and I believe this work will be a very important addition to the scientific literature of air pollution in rural regions.

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

Comment 1: Does the paper address relevant scientific questions within the scope of ACP? This paper attempts to create a machine-learned statistical model to estimate 8-h surface ozone concentrations for the Tibetan plateau. The RF-GAM model is interesting and the subject fits within the scope of ACP.

Response: Thank for reviewer's suggestions. The journal (ACP) focuses on the Earth's atmosphere and the underlying chemical and physical processes. The modelling of atmospheric components in the atmosphere is a hot topic of ACP, and many relevant references have been published in this journal. Therefore, we believe that the paper addresses relevant scientific questions within the scope of ACP.

Comment 2: Does the paper present novel concepts, ideas, tools, or data? The machine learning model formulation is interesting and the application of the GAM to remove autocorrelation in the time series data is novel.

Response: Thank for reviewer's suggestions. The paper presents a novel machine learning model named RF-GAM, which shows excellent performance in predicting the 8-h O_3 concentration over Tibetan Plateau. The development of this model is the major novelty of this present study. Besides, the present study fills the gap of statistical estimation 8-h O_3 level in a remote region, and provides useful datasets for epidemiological studies and air quality management. The implication is very important. Moreover, we also

Comment 3: Are substantial conclusions reached? Not particularly, the performance metrics presented in the work are not convincing. The authors present RMSE as an absolute measure of performance yet the best performing model had an RMSE of 14.41 μg/m^3, which is greater than 10% error relative to the WHO 8-h ozone critical value (100 μg/m^3) they are using to determine nonattainment. There is very little discussion about significant differences between performance due to model architecture and between seasons and years. Greater discussion of model uncertainty needs to be had. This study addresses an important topic that lacks much attention in scientific literature, but the performance of the model is still considerably lacking.

Response: Thank for reviewer's suggestions. The differences of performance between years has been added in the Table 1. The predictive performances in different seasons were also shown in Table 2. The contents have been expanded to discuss the temporal variation of predictive accuracy of O_3 concentration. The uncertainty of the present study is also added in the revised version: "The determination of nonattainment days showed some uncertainties owing to the predictive error of modelled O_3 concentration. First of all, meteorological data used in RF-GAM model were collected from reanalysis data and these gridded data often showed some uncertainties, which could increase the uncertainty of O_3 estimation. Second, the O_3 column amount used in the present study reflected vertical O_3 column amount rather than surface O_3 concentration. Thus, it could decrease the predictive performance of surface O_3 level". Besides, some important references were also added in the revised version.

Comment 4: Are the scientific methods and assumptions valid and clearly outlined? Scientific methods and data preprocessing are outlined well. Figure 2 presents a workflow method for the machine learning model, though including an actual architectural schematic of the final model would be better (e.g. what does the random forest look like in terms of the number of trees, nodes, etc.). Furthermore, none of the other machine learning models were given an architectural model description, i.e., how many nodes in the neural network, activation functions, etc. Machine learning is a burgeoning field with many nascent applications. Therefore, this paper would benefit going into greater detail about what other machine learning methods they tried in detail so that others could learn from their attempts.

Response: Thank for reviewer's suggestions. We have added some detailed information especially key parameters about various machine learning models in the revised version. Based on the iteration result, the optimal n_{tree} and m_{try} reached 500 and 5, respectively. For XGBoost method, the optimal tree depth and minimum child weight was set as 8 and 6, respectively. The activation functions of GRNN, BPNN, and ElmanNN were sigmoid function. The number of nodes in BPNN and ElmanNN were 5 and 4, respectively. The number of nodes in GRNN was equal to the sample size.

Comment 5: Are the results sufficient to support the interpretations and conclusions? (cite sections) There must be a greater statistical argument presented to support the interpretations of the authors. Oftentimes the authors present only an R^2 metric or an RMSE value to evaluate the goodness of fit among conditions. These performance measures are relatively obfuscating as even though ozone might increase or decrease between temporal/spatial extents, these changes, I suspect, are statistically insignificant. There are practically no significance tests conducted. These performance metrics must be put into greater context. Perhaps including confidence intervals (rather than only standard deviations) would also help. For the variable importance section 3.2, the authors should present an error covariance matrix for the inputs. For random forest algorithms, the variable importance determined by the model is meaningless if the variables are co-linear. Furthermore, oftentimes the explanations for the concentration of ozone are too speculative (e.g. lines 334-351). There are numerous mentions of NO_x and VOCs contributing to ozone loss or formation, yet this analysis cannot address such concerns directly as the model does not take into account these variables. These explanations are well thought out but exceed the scope of what this model can answer. Perhaps place them in the discussion rather than with the results. In addition to RMSE, an absolute metric, perhaps the authors should consider reporting a relative metric such as percent error in tandem. R^2 values of 0.60 and RMSE values of 14 μg/m^3 are not readily obvious/interpretable if this is adequate performance. Also, conducting significance tests between different models might be beneficial. I am not convinced that RF and XGBoost are markedly different. The addition of the GAM seems like a different model architecture than typical machine learning models. That is, why not attach the GAM to the XGBoost and see how that performs? It is fine if the authors wanted to create an RF-GAM model from the outset, but the comparison between the other models does not seem very robust.

Response: Thank for reviewer's suggestions. We employed one-way ANOVA method to compare the predictive values of eight models. The results are as follows:

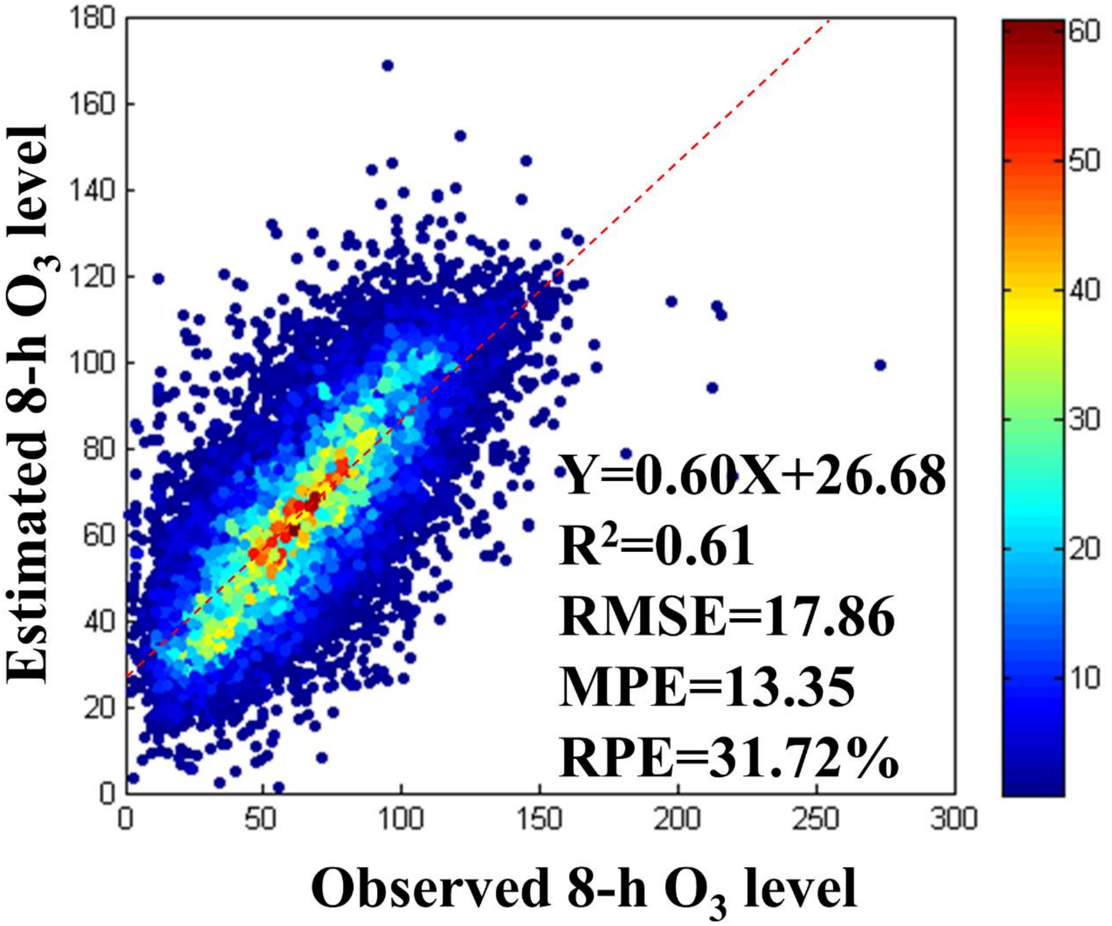
Source	SS	df	MS	F	p value
Groups	24542.6	7	3506.09	9.01	<0.01
Error	59209467.7	152154	389.14		
Total	59234010.3	152161			

Based on the result of one-way ANOVA, we found that the RF-GAM model showed significantly better predictive performance compared with other seven models.

Indeed, the traditional statistical model required that all of the independent variables should be deviated from the autocorrelation. The multicollinearity of variables might decrease the reliability of these models. However, the random forest method did not consider the autocorrelation of all the variables because the algorithm was nonlinear.

Indeed, some explanations for the concentration of ozone were speculative, and thus we revised many redundant or speculative explanations especially for the contribution of NO_x and VOCs to ozone formation. However, some explanations were well-founded because it could be supported by the Fig. S4. Besides, the effects of meteorological factors on ozone formation were confirmed by some previous studies and Fig. S5-S11, which was reliable.

In addition to RMSE, we also added a new indicator named relative prediction error (RPE) to assess the predictive accuracy (Fig. 3 and Fig. 4). We have employed the XGBoost method to combine the GAM to estimate the surface O_3 concentration (see the following figure), and found that the hybrid model showed the worse performance compared with RF-GAM model. The R^2 value of XGBoost-GAM model was significantly lower than that of RF-GAM model. Besides, both of RMSE and MPE for XGBoost-GAM model were also significantly higher than those of RF-GAM. Thus, we believed that RF-GAM was more appropriate to estimate the O_3 concentration in Tibetan Plateau.



Comment 6: Is the overall presentation well structured and clear? The overall structure and flow of the paper are coherent, though oftentimes the English is not clear.

Response: Thank for reviewer's suggestion. We have significantly revised the language throughout the paper.

Comment 7: Is the language fluent and precise? This study needs considerable corrections to numerous grammatical and language errors. There are far too often awkward phrasings and mixed tenses.

Response: Thank for reviewer's suggestion. We have significantly revised the language throughout the paper and corrected many grammar errors.

Comment 8: Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated? Many of the figures are too small to view. Needs to be increased in size.

Response: I agree with reviewer's suggestions. We have redrawn the Fig. 7 and Fig. 10 to increase the figure size.

Comment 9: Are the number and quality of references appropriate? Great overview and command of the literature.

Response: Thank for reviewer's suggestions. We have added some latest references in the revised version though many relevant references have been added.

Following are some line comments, though not all-inclusive:

Comment 10: a few times you use the word 'beneficial' to describe conditions that create pollution, perhaps consider different word choices as this is a bit awkward.

Response: I agree with reviewer's suggestions. "beneficial" was changed into "caused", "triggered", or "promoted".

Comment 11: Also found on line 127. 102-103: "decision tree models such as random forest (RF) and extreme gradient boosting (XGBoost) strike a perfect balance between prediction performance and computing cost". This seems to be a wide sweeping claim without a reference. Decision trees may be faster to train than neural networks, but they are also slower to evaluate. That is, decision trees are slow to train and slow to test, whereas neural networks are very slow to train and very fast to test. Consider revising your claim.

Response: I agree with reviewer's suggestions. (Line 102-103) The sentence has been changed into "Among these machine learning algorithms, decision tree models such as random forest (RF) and extreme gradient boosting (XGBoost) generally showed fast training speed and excellent prediction accuracy".

Comment 12: 135-138: awkward, confusing phrasing

Response: Thank for reviewer's suggestions. (Line 136-139) The sentence has been changed into "Unfortunately, these scarce monitoring sites in Tibetan Plateau cannot capture real O_3 pollution status especially in the remote areas (e.g., Northern part of Tibetan Plateau) because each site only possessed limited spatial representativeness".

Comment 13: Line 190: Still not sure why you included GDP as an input variable. You write that you integrated it because “these data were available each five years”. Please provide a scientific explanation for the inclusion of this variable.

Response: Thank for reviewer’s suggestion. GDP was included in the original model because many previous studies confirmed that GDP might be linked with the O_3 pollution (Li et al., 2020 JCP). It was well known that the hotspot of O_3 pollution focused on the area with the higher VOCs and NO_x (e.g., industrial points and residential areas). These regions often displayed the relatively higher GDP compared with other regions.