Preliminary response to reviewer's comment on "Recommendations on benchmarks for photochemical grid model applications in China: Part I – PM2.5 and chemical species" by Ling Huang et al.

Air pollution is a major environment problem and a hot scientific topic in China. Air quality model is a crucial kit to perform mechanism study, source apportionment study, strategy study and policy consultant. The usage of different air quality models increased exponentially over the past years. This work compiles studies during 2006-2019 using air quality models over China comprehensively, and analyses the accuracy of these studies over different regions with different models. Although the performance of some model results are compiled and evaluated in this work and the language presentation is good, however, I find this evaluation failed to follow the suggestion made by authors themselves and may be not based on a thoroughly review of previous modelling works. Furthermore, I find little improvement in this new reversion, it failed address my major concerns in the quick review. I could not suggest for publishing the current version, unless the following concerns are well addressed.
Response: In this short response, we reply to address the reviewer's comments in a quick and preliminary manner. A more detailed response with corresponding revisions will be provided separately.

1) A quick search on Web of Science tells me that there are about 74 papers published 2006-2019 using Geos-Chem to study air quality in China. This figure is much more than the other 3 models analysed in this study, CAMx, CMAQ, NAQPMS. Without include GEOS-Chem, I can not agree this samples used in this can represent the air quality modelling study in China and lead to a benchmark suggestion. Furthermore, I use the key word WRF-Chem, China and air quality, Web of Science gives me a result of 174 publications during 2006-2019. This figure is 3 times higher than the number of samples used in this study, which is only 56 samples. Authors need to fully justify the criteria them used for selecting samples.
Response: As suggested by the reviewer, we will include GEOS-Chem results in our revised manuscript. We will also provide a detailed description of our processes to select samples along with the revised manuscript. One thing we need to point out is that we started this work back in July 2019 so the results from Web of Science are expected to be a little different from what the reviewer found.

2) The title does not reflect the present work. This work mainly focuses on PM, but the title highlights photochemical model. I feel more discussion about ozone pollution need to be included, given that ozone is the key secondary pollution of photochemistry and is becoming more and more important for air quality in China. Without including ozone, this study is far from any recommendation on benchmarks for photochemical models.
Response: As specified in our current manuscript (Page 3, Line 17-20), we plan to prepare three companion papers: the first one (i.e. current one) focuses on $PM_{2.5}$ and speciated chemical components, considering that significant attention has been given

to $PM_{2.5}$ pollution in China for the past decade; the second one, which is currently under preparation, will be solely focusing on ozone, given that ozone pollution is becoming a more prominent problem over $PM_{2.5}$ in recent years; the last one will be focusing on other pollutants (e.g. PM_{10}, SO_2, NO_2). The purpose of this set of work is to give a comprehensive review of air quality model applications in China and the resulting model performance. We feel that it would be too much to include all information into one single manuscript. That's why we decided to present them separately.

3) Authors need to include the evaluation of meteorology performs in this study, instead of will be discussed as a future work . As suggested by authors themselves in the conclusion part: It is always good practise to present model performance results of meteorological field. . . Performance results of meteorological model could also help explain potential causes of unsatisfactory PGM simulated results. Analyse the air quality performance in conjunction with meteorological performance will certainly improve the value of this work. Separating a nice and comprehensive work to individual pieces is not a good practise and also not good for a prestigious journal such as ACP.

Response: We agree with the reviewer that meteorological performance is a critical part of a comprehensive and complete evaluation of air quality model application. However, for three reasons we decided to present meteorological evaluation results as a separate work. Firstly, the evaluation of meteorological modeling is a standalone scientific question by itself that requires a separate discussion for that. There are many more applications of meteorological simulations than providing inputs for air quality simulations. Secondly, as mentioned in our current manuscript (Page 5, Line 12), not all studies that performed evaluation of air quality simulations also evaluated meteorological simulations, given that good performance of air pollutants implicitly suggest accurate meteorological simulations. Lastly, including discussions on meteorological simulations would considerably increase the length of the current manuscript. Again, the current manuscript is aiming to focus on $PM_{2.5}$ and its chemical components. In summary, we acknowledge the importance of evaluating meteorological simulations and we feel it deserves a separate discussion.

4) As suggested by authors themselves in the conclusion part: In addition to providing numerical values of statistical metrics for model performance evaluation, graphs/plots are strongly recommended to further support model validation. To give a few examples visualizing data via time series plots of modelled and observed data could help illustrate periods with better or poorer performances. I believe audiences are also expecting to see a time series plots of model performance over 2006-2019. Did we improve the ability of air quality simulation over past decades? If yes, what is the critical step we have improved; if no, where is key problem we should focus on in future? These are the key questions/suggestions we are keen to know from this comprehensive review study, and will add great value to this work and large help for the modelling community. However, this information is absent. I would like to

suggest some further discuss in this direction, in addition to the summary of performance in previous works.

Response: We agree that graphical analysis is an important component of model performance evaluation. Graphical and statistical analyses are complementary. The reviewer provides relevant examples of how graphical analysis can be used to explain and illustrate important aspects of model performance. In response to the reviewer's suggestion, we can add a section that reviews best practices for using graphical analysis in model performance evaluation.

With respect to the time series plots of model performance over 2006-2019, we will look into this more and provide a detailed response in our upcoming revised manuscript.

5) As suggested by authors themselves in the conclusion part: Provide as much details as possible with respect to how observation and modelling results are used to obtain the statistical results. However, I feel very limited details are provided for some statistical analyses of this work. At lease, for me, it is difficult to understand or reproduce the Fig. 9. What does x-axis mean? Sample fraction, fraction of what? Why the sum of fractions is larger than 100%, are they integrated values? Here is just an example, more details need to be provided in captions.

Response: A short answer to questions regarding Figure 9. This is how we produce Figure 9. To give an example of IOA values reported for $PM_{2.5}$. There are in total 32 studies that reported IOA values for $PM_{2.5}$ and the total number of IOA reported is 47 (multiple IOA values could be reported in a single study). We sorted these 47 numbers from high to low and the corresponding sample fraction for individual number is calculated as the sorted rank divided by 47 (total number). Then we plot these 47 IOA numbers as y-axis and the corresponding sample fraction as x-axis (as shown in Figure 9). Based on this plot, we can directly tell one third (first dashed vertical line in Figure 9) of previously reported IOA values for $PM_{2.5}$ is greater than 0.91 and another third (second dashed vertical line) of previous reported IOA values is lower than 0.69. In this sense, the audiences could place their IOA results on Figure 9 and get a sense of where their results are located with respect to previous studies. We will provide more details in the upcoming revised manuscript.