Point-by-point Response to Reviewer's Comments

Report #1 by Anonymous Referee #3

The revised manuscript has improved its scientific integrity and responded to questions/comments provided during the interactive discussion. Here are a few additional comments:

We appreciate the reviewer for taking time to carefully review the manuscript and give detailed and constructive comments, which has greatly helped to improve this paper. Below is our point-by-point response to each comment.

• Page 1 (Abstract): Abstract should include major results or conclusions derived from this study.
Response: We have revised the abstract by including major results of this study, which has been highlighted in the revised manuscript and also shown below.

Revised abstract:
Numerical air quality models (AQMs) are being applied more frequently over the past decade to address diverse scientific and regulatory issues associated with deteriorated air quality in China. Thorough evaluation of a model's ability to replicate monitored conditions (i.e. a model performance evaluation or MPE) helps to illuminate the robustness and reliability of the baseline modelling results and subsequent analyses. However, with numerous input data requirements, diverse model configurations, and the scientific evolution of the models themselves, no two AQM applications are the same and their performance results should be expected to differ. MPE procedures have been developed for Europe and North America but there is currently no uniform set of MPE procedures and associated benchmarks for China. Here we present an extensive review of model performance for fine particulate matter ($PM_{2.5}$) AQM applications to China and, from this context, propose a set of statistical benchmarks that can be used to objectively evaluate model performance for $PM_{2.5}$ AQM applications in China. We compiled MPE results from 307 peer-reviewed articles published between 2006 and 2019, which applied five of the most frequently used AQMs in China. We analyse influences on the range of reported statistics from different model configurations, including modelling regions and seasons, spatial resolution of modelling grids, temporal resolution of the MPE, etc. Analysis using a Random Forest method shows that the choices of emission inventory, grid resolution, and aerosol and gas-phase chemistry are the top three factors affecting model performance for $PM_{2.5}$. We propose benchmarks for six frequently used evaluation metrics for AQM applications in China, including two tiers – "goals" and "criteria" – where "goals" represent the best model performance that a model is currently expected to achieve and "criteria" represent the model performance that the majority of studies can meet. Our results formed a benchmark framework for the modelling performance of $PM_{2.5}$ and its chemical species in China. For instance, in order to meet the goal and criteria, the normalized mean bias (NMB) for total $PM_{2.5}$ should be within 10% and 20% while the normalized mean error (NME) should be within 35% and 45%, respectively. The goal and criteria values of correlation coefficients for evaluating hourly and daily $PM_{2.5}$ are 0.70 and 0.60, respectively; corresponding values are higher when the index of agreement (IOA) is used (0.80 for goal and 0.70 for criteria). Results from this study will support the ever-growing modelling community in China by

providing a more objective assessment and context for how well their results compare with previous studies, and to better demonstrate the credibility and robustness of their AQM applications prior to subsequent regulatory assessments.

• Page 3 (2.1 Data compilation): The number of studies considered in this revision increased significantly from 128 to 307, which is astounding. Since this is the first of a series of PGM evaluation studies, it is critical that a set of selection criteria is well defined and strictly applied. The five selection criteria used in this study are scientifically sounds but excluding non-English journals or journals with <10 publications is not. In particular, the latter seems to have over 300 publications available and relevant to the PGM evaluation but excluded in this study.

Response: We agree with the reviewer that a clear description of the set of selection criteria is important before demonstrating the results. A detailed description of the selection criteria was provided in the manuscript. With these criteria strictly applied, we finally included 307 articles in this study, which is much larger than previous similar studies for U.S. (e.g. only 69 studies in Simon et al. (2012) and 76 studies in Emery et al. (2017)). We believe that this compilation could give a general picture of the model performances.

We excluded the non-English journals because: (1) compared to English journals, they have narrower audiences; (2) the majority of the evaluation results covered by the non-English journals are also covered by the English journals written by the same group of researchers; and (3) we believe that most of the evaluation results reported by the Chinese journals are comparable with those published in English.

We excluded journals with less than 10 publications for the following reasons: (1) The 307 studies (out of 464 studies found by our Web of Science search) that we included are published in main stream air quality-related journals (especially in the field of air quality modeling). In contrast, many of the excluded studies were not in air quality-related journals and they appeared in the Web of Science search simply because of a key word. (2) 307 studies is a large body of data from which to draw conclusions suggesting that adding more studies is statistically unlikely to change our findings. In summary, we believe that the 307 included studies are representative of current results for air quality model applications and including more studies would be unlikely to change our major conclusions. We revised the manuscript and inserted the explanations.

Revised manuscript (Page 3, Line 16-20):

Our investigation started by searching for combinations of three key words on the Web of Science: model name, “air quality”, and “China”, and limited the timespan between 2006 and 2019. This initial search gave 446 (CMAQ), 84 (CAMx), 256 (WRF-Chem), 117 (NAQPMS), and 58 (GEOS-Chem) records (a total of 961). Duplicated records were excluded. We then excluded records that were listed as conference papers or not published in English-language journals (for example, Chinese and Korean-language journals) due to narrower audiences. This resulted in 826 records published in 61 journals. We further reduced the number of journals considered by excluding those that had less than ten publications during 2006-2019, since most of the excluded journals are not air quality-related journals, which results in 464 studies. Table S1 shows the list of journals that were included in this study, which is believed to cover the mainstream journals in atmospheric research, especially in applications of air quality models.

• Page 8 (3.3. Recommended metrics and benchmarks): Figure 10 shows 6 graphs in order of NMB, NME, FB, FE, R, and IOA while the text discusses the Figure in order of R, IOA, NMB, NME, FB and FE. Can the order be consistent? It seems that some numbers have not been updated from the previous version so please check. Most of the values in the text reflect what is shown in Figure 10 but some numbers rounded.

Response: Thanks to the reviewer for pointing out this inconsistent issue. The order of Figure 10 and Table 2 are revised to match text discussions. We double checked the values. We rounded up some numbers to nearest 0.5 or 5% to consistently recommend a set of round values.

Revised figure and table:

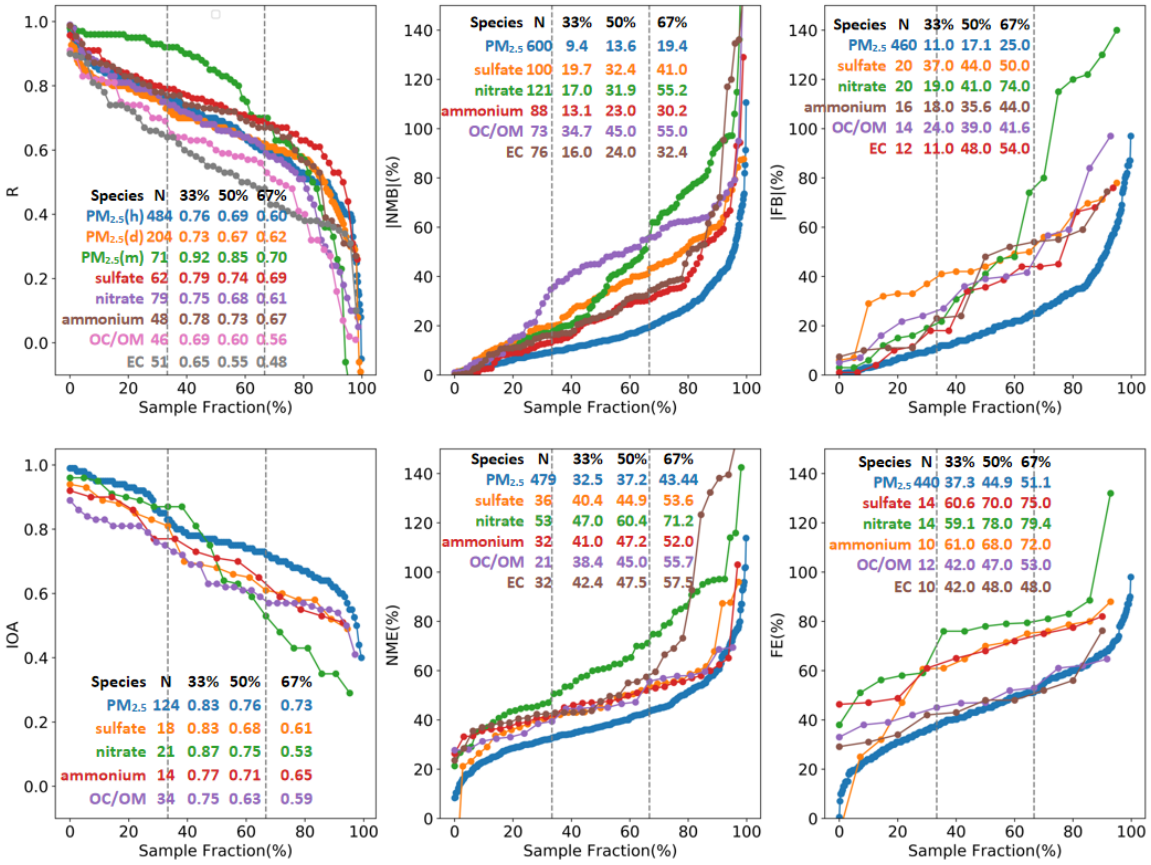

Figure 10: Rank-ordered distributions of R, IOA, NMB, NME, FB, and FE for total $PM_{2.5}$ and speciated components. The number of data points and the 33rd, 50th, and 67th percentile values are also listed. For instance, one third of reported R value for predicted hourly $PM_{2.5}$ concentration is higher than 0.76; half is higher than 0.69; and two thirds higher than 0.60.

Table 2: Recommended benchmarks for evaluating AQM applications in China for total $PM_{2.5}$ and speciated components [a, b]

Metrics	Benchmark level	$PM_{2.5}$	sulfate	nitrate	ammonium	OC/OM	EC
R	Goal	>0.70 (hourly/daily) >0.90 (monthly)	>0.75*	>0.70	>0.75*	>0.65	>0.65
	Criteria	>0.60* (hourly/daily)	>0.65*	>0.60	>0.65*	>0.55	>0.45

		>0.70 (monthly)					
IOA	**Goal**	>0.80	>0.80	>0.85	>0.75	>0.75	None
	Criteria	>0.70	>0.60	>0.50	>0.60	>0.55	None
NMB	**Goal**	<±10%	<±20%	<±20%	<±15%	<±35%	<±20%
	Criteria	<±20%*	<±45%	<±60%	<±35%	<±55%	<±35%*
NME	**Goal**	<35%	<45%	<50%*	<45%	<40%*	<45%*
	Criteria	<45%*	<55%	<75%*	<55%	<60%*	<60%*
FB	**Goal**	<±15%	<±40%	<±20%	<±20%	<±25%	<±15%
	Criteria	<±25%	<±50%	<±75%	<±45%	<±45%	<55%
FE	**Goal**	<40%	<65%	<60%	<65%	<45%	<45%
	Criteria	<55%	<75%	<80%	<75%	<55%	<50%

[a] Values with an asterisk in Table 2 indicate that our benchmarks are stricter than corresponding values in Emery et al. (2017)

[b] Shaded values indicate that less than 20 data points were available to develop the benchmarks.

Report 2 by Referee #1

Thanks for your revision, which addressed my concerns in part, but there are still some important issues have not been totally addressed. And I think, these issues could be critical for making this work valuable for other studies in future and therefore shaping this work be suitable for publishing in ACP.

We are grateful to the reviewer for taking time to carefully review the manuscript and give detailed and constructive comments, which has greatly helped to improve this paper. Below is our point-by-point response to each respective comment.

1) Authors include GEOS-Chem and conclude that GEOS-Chem's performance is less satisfied due to its relative coarse resolution leading to insufficient resolve details in interactions between emission and chemistry in a city-scale. Then, could authors provide more discussion in the manuscript to stress the necessary/advance of fine-res. models could be a promising trend of future air quality modelling studies to help improve our understanding.

Response: Thanks for the comment. We have added discussions regarding the application of GEOS-Chem in regional scale in the revised manuscript. GEOS-Chem is a global 3-D atmospheric chemistry model with state-of-science developments and large international user base. However, due to its coarse resolution, limitations exist when applying GEOS-Chem to simulations at regional or local scale. To tackle this issue, Lin et al. (2020) developed a new online regional atmospheric chemistry model - WRF-GC (v1.0), that integrates the WRF meteorology model and GEOS-Chem chemistry model. This new WRF-GC model has been successfully configured at finer resolution (27 km x 27 km) and applied to quantify the changes of NOx emissions due to COVID-19 for Eastern China (Zhang et al., 2020), illustrating the potential applications of GEOS-Chem at finer spatial scale. These discussions have been added to the revised manuscript.

Revised manuscript (Page 8, Line 21-33):

Fine resolution simulations have been conducted with the intention of improving model performance. With finer grid resolution, the spatial allocation of certain features in emission patterns is significantly improved, which is especially important for air quality simulations at local scale (Tan et al., 2015; Liu et al., 2020). Additionally, meteorological simulations could also be improved at finer resolution given more detailed land cover and structures in topography (Tao et al., 2020), which in turn improves the subsequent air quality simulations. Estimation of PM2.5 related health impacts are reported to be biased high/low at coarse spatial resolution (Li et al., 2017; Thompson and Selin, 2012). Lin et al. (2020) developed a new online regional atmospheric chemistry model - WRF-GC (v1.0), that integrates the WRF meteorology model and GEOS-Chem chemistry model. This new WRF-GC model has been configured with a spatial resolution of 27km and successfully applied to quantify the changes of NOx emissions due to COVID-19 for East China (Zhang et al., 2020), illustrating the potential applications of GEOS-Chem at finer spatial scale. However, not all fine resolution simulations lead to improved model performance, especially when the input data are not available with the same high resolution (Jiang and Yoo, 2018; Tao et al., 2020). Therefore, grid resolution should be determined depending on the purpose of the study and the availability of input data.

2) I do not object to your reply-02, however, reply-02 still did not address my question. Ozone is the central pollutant of photochemistry. And you do not talk a work about ozone in this manuscript, whose title highlight the "photochemical". Here is an example. A manuscript entitled "How deadly is COVID-19?", but it only talks about how to develop a vaccine, do not mention a word of global number of confirmed cases and death toll. Do you think this is a good title reflect the content presented? So, my point is, either title is not suitable or ozone need to be discussed.

Response: Thanks for the kind comment, we agree that ozone is a key air pollutant if we talk about "photochemistry". We changed the title from "photochemical grid model" to "Numerical air quality model". We use "Numerical" to make clear that we do not consider Lagrangian air quality models (for example, Hysplit, Calpuff) that have much simpler chemistry or none.

3) I do not agree that evaluation of meteorology is a "standalone scientific question". First, as suggest by author in the conclusion that performance of meteorology simulation can directly influence air quality simulation. Second, as replied by authors to Reviewer-02:

"Consequently, errors in inputs and algorithms present in this group of simulations are likely to be correlated (e.g., tendency for higher emissions correlated with more dispersive meteorological simulations), which will hinder reliable diagnosis of factors contributing to model error."

Exactly, a good air quality simulation can be achieved for wrong reasons when meteorological performance is poor, this will hind the reasons of uncertainty and hamper improvement in our understanding.

So, a convincing evaluation of air pollutants should always build on the top the evaluation of meteorology. I suggest to include meteorology in the work. If authors do insist to separate them, then at least, the meteorology evaluation work should publish first, and this work can cite it and build on the top of it.

Response: We totally agree with the reviewer that meteorological performance can influence the air quality simulation results to some extent and meteorological performance is an essential part of any air quality model evaluation. However, we decided not to include discussions on meteorological evaluation in this paper for the following reasons. First, not all of the 307 studies included in this study reported model performance results for their respective meteorological simulations, so the development of MPE benchmarks for meteorology would necessarily consider a different set of studies and lead to inconsistent meteorology-air quality performance connections. There are studies that performed evaluations for air quality simulations but did not mention meteorological evaluation and vice versa. Second, we see evaluations of meteorology and air quality are rather distinct issues, and it doesn't really matter which publication comes first. Indeed, the reviewer raised an important scientific question regarding how the meteorological MPE influences the air quality MPE, which is a very interesting topic and needs much more complex analysis. However, this is beyond the scope of the current study and we will possibly consider it in our following research. We have inserted explanations in the revised manuscript.

Revised text (Page 5, Line 18-25):

Meteorological data are needed to drive air quality simulations and the performance of meteorological modelling is a key source of uncertainty for air quality modelling performance. Meteorological data were mostly simulated by the Weather Research Forecasting (WRF) model (Skamarock et al., 2005) in our compiled studies; the Fifth Generation Penn State/NCAR

Mesoscale Model (MM5) (Grell et al., 1994) and the Regional Atmospheric Modelling system (RAMS) were used in a few studies. Model performance of meteorological results should be evaluated in addition to air quality simulation results. However, several studies did not report any results with respect to their meteorological simulations. The performance of meteorological results used to drive air quality simulations and how it could affect the air quality simulations is beyond the scope of the current work and will need to be discussed as a future work.

4) in reply-04, authors state "Our results indicate that the top three factors involve emission inventory, grid resolution, and boundary conditions, while the choice of model and source of meteorology are least important. This is a very preliminary analysis and thus we have decided to include these discussions in the supporting materials.".
I think this is the most important part of the present work, it could be the unique contribution to the modelling community and hence a unique value of this work. We do want to have detailed discussions and promote the discussions to main text. Even though there are limitations, we could still have discussions of these limitations (all studies have limitations) and the uncertainties, and provide valuable advices for future studies to overcome these limitations. Put this part in SI really under value the present work.
Response: Thanks for the positive comment. We agree that moving this part to the main text could improve the depth of this study. As suggested by the reviewer, we moved this part to the main work.

Revised manuscript:
2.4 Feature importance based on Random Forest

Random Forest is a machine learning method suitable for classification and regression (Liu et al., 2012). It is a collection of a series of decision trees and each tree is generated from a bootstrap sample. Both continuous and categorical input variables are allowed. It can provide the order of feature importance (FI) so that we can determine and rank which parameter choices most influence the simulation results.
We reviewed the model configurations for studies that reported correlation coefficient, IOA, MB, NMB, mean error (ME), normalized mean error (NME), fractional bias (FB), and fraction error (FE) for $PM_{2.5}$ (a total of 176 studies). Model configurations include the meteorological data that are used to drive air quality simulations (e.g. from WRF, MM5, or GEOS), the emission inventory (e.g. public available dataset vs. locally developed), gas-phase chemistry (for example, carbon bond vs. Statewide Air Pollution Research Center (SAPRC)), aerosol chemistry (including inorganic aqueous chemistry, inorganic gas-particle partitioning, organic gas-particle partitioning and oxidation), boundary conditions (e.g. model default values vs. results generated from global model), grid resolution and the temporal resolution (Table S7). We ignored the study region and period for FI selection because these two options are more restricted by the user's specific needs and focus (i.e., more subjective/uncontrollable and less objective/controllable). We ranked each statistical metric from good to poor performance. For example, values of R and IOA that are close to 1 represent good performance and values close to 0 represent poor performance. For MB and NMB, we used absolute values so that deviations from zero represent the performance level. These results were classified into three tiers with breaks at 33% and 67% of the ranked values so that each tier includes the top one third, the middle one third, and the bottom one third of the reported performance results. The random forest model was performed using the 'sklearn' module in Python to obtain the FI metric.

3.3 Recommended metrics and benchmarks

As mentioned earlier, AQM applications involve numerous driving inputs as well as diverse model configurations, which lead to an abundant database from which to assess their relative influences on model performance. The similarities between the benchmarks derived in this study and Emery's study suggest that important model input data (e.g. emission inventories) have comparable accuracy for China and North America and model formulations (e.g. algorithms such as chemistry, deposition, transport) seem to be equally applicable to China and North America. In additional to the need for model performance benchmarks, there also is a need for more studies that quantify contributions to model uncertainty, such as the recent study by Dunker et al. (2020), which quantifies contributions of chemistry, boundary concentrations, deposition and emissions to uncertainty in simulated ozone results. In this study, we applied the Random Forest method for pattern recognition to identify and rank model attributes (inputs, grid resolutions, etc.) that have important influences on $PM_{2.5}$ model performance. The choice of emission inventory is shown to affect the model performances most, followed by grid resolution, aerosol and gas chemistry (Figure 11). Meteorological input and the choice of model itself is of least importance.