Interactive comment on "Recommendations on benchmarks for photochemical grid model applications in China: Part I – $PM_{2.5}$ and chemical species" by Ling Huang et al.

Response to Reviewer's Comments

Anonymous Referee #1

Received and published: 28 July 2020

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: We thank the reviewer for these valuable comments. We have made extensive efforts to improve the existing manuscript and aim to address the reviewer's concerns. Our major improvements include:

(1) re-write the Abstract/Introduction part to emphasize the importance and applications of results obtained from this work;

(2) adding GEOS-Chem results thus leading to a more complete picture of PGM applications in China;

(3) a detailed description of literature selection process is provided and the number of peer-reviewed publications increased from 128 to 307 (Section 2.1);

(4) more discussions were added to provide more in-depth insights into our findings, including trends of model performance results over the past decade (Section 3.2), application of Random Forest Model to rank feature importance of key model inputs

(Section 3.3 and Supplemental information).

Our responses to the reviewer's comment are given below in blue. Revised manuscript with revisions highlighted in yellow is attached after the response.

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 included GEOS-Chem studies in our revised manuscript, which now leads to a total of five models included: CMAQ, CAMx, WRF-Chem, NAQPMS, and GEOS-Chem. The reason that we did not include GEOS-Chem at first is because GEOS-Chem is a global model with relatively coarse resolution (~200 km as default grid spacing), while the other four models are considered regional air quality models with relatively finer resolution (for example, 36 km and lower). Of the 20 GEOS-Chem studies that we compiled in this study, the finest grid resolution is $1/4° \times 5/16°$ (~30 km) and the coarsest resolution is $2° \times 2.5°$ (~200 km). As a result, GEOS-Chem cannot resolve details in the interactions between emissions and chemistry at city-scales and dispersion patterns will be rather different from the regional models. In the revised manuscript (Section 2.1), we also provide a detailed description of how we selected the samples from an initial 900+ Web of Science records down to a final set of 307 papers that were ultimately compiled in this study. This addresses the reviewer's concern regarding the criteria that we used for sample selection.

2.1 Data compilation

A total of five photochemical models – the Community Multiscale Air Quality (CMAQ, Foley et al., 2010), the Comprehensive Air Quality Model with Extensions (CAMx, Ramboll Environment and Health, 2018), the Goddard Earth Observing System (GEOS)-Chem (http://geos-chem.org), the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem, Grell et al., 2005), and the Nested Air Quality Prediction Modelling System (NAQPMS, Z. Wang et al. 2006) – are included in this compilation. While the former four models are developed by institutes and/or companies outside China, the NAQPMS is developed by the Institute of Atmospheric Physics of Chinese Academy of Sciences and has mostly been utilized for applications in China.

GEOS-Chem is a global chemical transport model with coarser resolution (only 20% of complied GEOS-Chem studies has a grid resolution less than 50 km), as opposed to the other four regional models that are applied with finer spatial resolution at local scale (for example, less than 10 km). 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 gives 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). 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, 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. The next filtering stage needed substantial manual effort. The 464 records were downloaded and manually checked to exclude (1) studies that were accidentally included in the search but did not apply any of the models in their study; (2) studies that were intended for other purposes (for example, evaluating meteorological simulations); (3) studies that were not focused on China (for example, the target region was Korea, Japan, etc.); (4) studies that did not provide any air quality model performance evaluation or the evaluation results were referred to previous studies; (5) studies that did conduct model performance evaluation but no numerical values were given (for example, only graphical plots were given). The final selection included a total of 307 papers (see a complete list in Table S2). We defined ten regions of China as shown in Figure 1, namely Beijing-Tianjin-Hebei (BTH) region, Yangtze River Delta (YRD) region, Pearl River Delta (PRD) region, Sichuan Basin (SCB), North China Plain (NCP), Central, Northwest, Northeast, Southeast, and Southwest (see Table S3 for provinces covered in this region).

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 mentioned in our manuscript (Page 2, Line 34-37), we plan to develop three studies in series: the current study focuses on $PM_{2.5}$ and speciated chemical components (as stated in the title), considering that significant attention has been given to $PM_{2.5}$ pollution in China over the past decade. The second study, which is currently under preparation, will solely focus on ozone, given that ozone pollution is becoming a more prominent problem in recent years. The last study will focus 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 their resulting performance against measurements. We feel that it would be too much to include all information in a single manuscript.

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. First, the evaluation of meteorological modeling is a standalone scientific question by itself that requires a separate in-depth analysis and discussion. There are many more applications of meteorological simulations than providing inputs for air quality applications. Second, as mentioned in our current manuscript (Page 4, Line 40), not all studies that performed evaluation of their air quality simulations also evaluated or reported meteorological results, given that good air quality performance implicitly suggests acceptable meteorological simulations. Last, including discussions on meteorological evaluations 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 as a way to guide future modeling applications with context about how well their modeling results compare to well-performing historical simulations specifically in China. 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: The reviewer brought up an excellent question: “Did we improve the ability of air quality simulation over past decades and if yes, what is the critical step that has been improved”. The answer to this question is valuable for the modeling community, yet extremely difficult to answer, because model applications over the decades have been so diverse , with different and evolving models and physical/chemical treatments, different and evolving model inputs (e.g. emission inventory, boundary conditions, meteorological inputs), different model configurations (e.g. vertical layers, spatial resolutions), and different modeling periods and regions. These peer-reviewed studies were conducted independently and they are not designed as a set of controlled experiments to limit variability (i.e. varying one input/algorithm while holding others unchanged). Several studies were focused on improving model performance by developing better emission inventories. Some studies focused on model chemical mechanisms such as number of model species (SAPRC vs. carbon bond), incorporating new formation pathways (e.g. heterogeneous reactions, chlorine chemistry) or improving the modeling framework (two-product vs. VBS for SOA modeling). It would not be possible for us to analyze what factors influenced model performance in ~300 individual studies.

Nevertheless, we attempted to address the reviewer’s comments as follows:

- To answer the reviewer’s question, “Did we improve the ability of air quality simulation over past decades”, we added time series plots of commonly used statistical metrics for $PM_{2.5}$ and speciated components (depending on data availability) reported in literature during 2006-2019 and associated discussions are added in the revised manuscript (see Section 3.2).

  In revised manuscript:

  Trends over the past decade

  In an attempt to assess whether model performance results have evolved over the past decades, we present time series of selected statistical metrics for total PM2.5 in Figure 9 (plots for inorganic species are shown in Figure S3). Results published prior to 2013 were aggregated into one group because there were a limited number of studies prior to 2013. For total $PM_{2.5}$, reported R values have remained relatively consistent over the past decade with the median fluctuating within 0.6~0.8. The ranges of reported RMSE and MB become narrower in recent years even though the number of studies has increased substantially. Reported IOA and RMSE values fluctuated upward and downward over the period. On the other hand, there seems to be an improving trend in terms of FB, FE, and NME as the reported values for these three metrics shift towards zero. For instance, the median value of reported FE decreased from 56.9% prior to 2013 to around 33% in 2019. However, it is important not to over-interpret these results as the number of studies published each year could affect the results.

- In an effort to answer the reviewer's question, "what is the critical step we have improved", we did a preliminary analysis of 176 studies that reported model performance results for $PM_{2.5}$ based on the random forest method (see "Analysis of feature importance based on random forest method" in the revised supplemental information). 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.

  In supplemental information:

  Analysis of feature importance based on Random Forest Method

  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. 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. Like other machine learning methods, random forest is also a black box. 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 collected detailed model configurations for studies that reported results of correlation coefficient (R), index of agreement (IOA), mean bias (MB), normalized mean bias (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. 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 did not include the study region and period for FI selection because we feel 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 bad performances. For MB and NMB, we used absolute values so that deviations from zero represent the performance level. We then classified these results 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. We then ran the random forest model using the 'sklearn' module in python to obtain the FI metric and the results are shown in Figure S4. The choice of emission inventory is shown to affect the model performances most, followed by grid resolution, aerosol and gas chemistry. Meteorological input and the choice of model itself is of least importance.

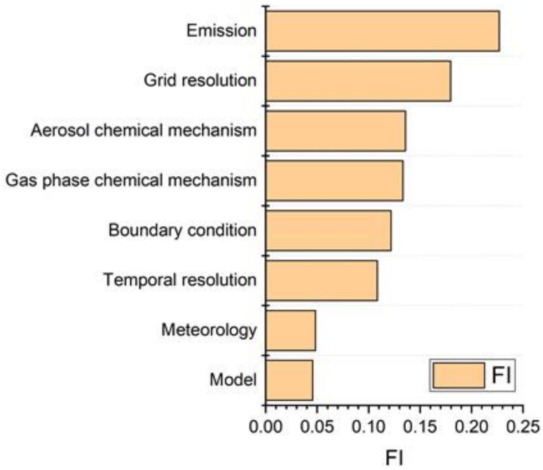

Figure S4: Ranking of key model inputs in terms of feature importance

- Finally, in addition to the 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 the contributions of uncertainties associated with chemistry, boundary concentrations, deposition and emissions to uncertainty in simulated ozone results. These discussions were added in the revised manuscript (see Section 3.3).

In revised manuscript:

As mentioned earlier, PGM 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. A preliminary analysis based on the Random Forest Method (Liu et al., 2012), a machine learning method suitable for classification and regression, suggests that emission inventory, grid resolution and boundary conditions are the top three factors that affect model performances results (see details in Supplemental information). 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.

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: We have added more descriptions for Figure 10 in the revised manuscript. We also modified the caption to include more details. 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 values reported is 47 (multiple IOA values could be reported in a single study). We sorted these 47 numbers from high to low and calculated the corresponding sample fraction for each individual number as its sorted rank value 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 10). Based on this plot, we can directly tell that one third (first dashed vertical line in Figure 10) 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 reader can compare their IOA results to Figure 9 and get a sense of where their results are located with respect to previous studies. An example is added to the caption for clarification.

References:

Dunker, A.M., Wilson, G., Bates, J.T. and Yarwood, G., 2020. Chemical Sensitivity Analysis and Uncertainty Analysis of Ozone Production in the Comprehensive Air Quality Model with Extensions Applied to Eastern Texas. Environmental Science & Technology, 54(9), pp.5391-5399.

Liu, Y., Wang, Y., & Zhang, J. (2012, September). New machine learning algorithm: Random forest. In International Conference on Information Computing and Applications (pp. 246-252). Springer, Berlin, Heidelberg.

Interactive comment on “Recommendations on benchmarks for photochemical grid model applications in China: Part I – $PM_{2.5}$ and chemical species” by Ling Huang et al.

Response to Reviewer’s Comments

Anonymous Referee #2

Received and published: 12 August 2020

The manuscript compiles 128 prior publications of chemical transport modeling studies on PM air pollution in China and summarizes model performances in commonly used statistics such as correlation, bias, quantile distribution, etc. I have three major concerns of the manuscript which makes it unsuitable for publication in ACP.

Response: We thank the reviewer’s comments. We have made substantial efforts to improve the existing manuscript and aim to address the reviewer’s concerns. Our major revisions include:

(1) re-write the Abstract/Introduction part to emphasize the importance and applications of results obtained from this work;

(2) expand our compiled studies to include GEOS-Chem studies and provided a detailed description of how we compiled the studies;

(3) to provide discussions on how model performances evolve over the past 15 years; and

(4) to provide a preliminary analysis on the importance of several model key inputs on model performance results.

Our responses to the reviewer’s comment are given below in blue. Revised manuscript with revisions highlighted in yellow is attached after the response.

First, treating it as a research article I do not find the manuscript contains new knowledge in its current form. All the graphs and tables are simple summaries of the results from published papers. To justify their study, the authors make an affirmative statement in the introduction that benchmark metrics developed based on US and European studies may not be suitable for model evaluation in China (pg 7, line 5-7) but they do not show scientific evidence or conduct their own analysis to support this claim. On the contrary, all the benchmark metrics the manuscript recommended have been proposed and used in the US or Europe and none of them is specific to China. The authors made an argument on the correlation coefficient being inconsistently used in prior studies (pg 4, line 10-15). I found this a trivial matter which can be easily reconciled by a careful reading of the reference of interest.

Response: Applications of PGMs in China have increased significantly over the past decade because of their unique features and capabilities that cannot be achieved via field observations or chamber studies. PGM applications are especially important in air quality management practices because they are extensively used to identify source (both sectoral and regional) contributions to air quality problems as well as support formulation of control strategies. A critical step of all PGM applications is a robust and comprehensive model performance evaluation (MPE) to ensure the accuracy of the simulated results. However, MPE results from different studies differ dramatically. For example, the two most commonly used MPE statistical metrics – normalized mean bias (NMB) and correlation coefficient (R) could range -87%~110.6% and -0.59~0.98, respectively. No unified guidance, references or contextual information are provided for the ever-growing modeling community in China to interpret how well or bad their model performance results are.

Therefore, there are two objectives of this study. The first one is to provide a general overview of the status of PGM applications in China over the past 15 years; for example, the most frequently used models, the most frequently investigated regions, the most frequently used statistical metrics for model performances, etc. This is presented in Section 3.1. The second objective, which is more important, is to recommend performance benchmarks for commonly used statistical metrics (not new metrics!) based on studies in China. The underlying reasons for this need include: (1) no guidelines on systematic and standard model performance criteria or goals are available in China to provide context for good vs. poor results relative to the growing number of applications in China; and (2) the benchmarks that were used for quantitative model performance evaluation in some of the studies are based on results of PGM applications in the U.S., which are outdated (for example, the “Guidance on the Use of Models and Other Analyses for Demonstrating Attainment of Air Quality Goals for Ozone, $PM_{2.5}$ and Regional Haze” developed by the U.S. EPA was published in 2007 (EPA, 2007), which is almost 15 years ago; the benchmark of FB and FE for PM introduced by Boylan and Russell, which is used in quite some studies for model performance evaluation, was published in 2006 (Boylan and Russell, 2006)) and may not be appropriate for PGM applications in China, due to different model configurations (e.g. unique environments, quality of emission inventories, availability of source emission profiles, etc.). It is appropriate to come up with benchmarks that are solely based on PGM applications in China for a more direct apple-to-apple comparison. The statistical metrics are not specific to China but the recommended benchmarks are. This part is presented in Section 3.3. This is the first time that a set of quantitative and objective MPE benchmarks that are suitable for PGM applications in China are recommended for the scientific and regulatory community. Results from the current study will support the ever-growing modelling community in China by allowing for objective assessments of how well their simulation results compare with historical studies and to better demonstrate the credibility and robustness of PGM applications prior to subsequent regulatory assessments. We have re-written the abstract and introduction part to emphasize the importance of this work.

Abstract

Photochemical grid models (PGMs) are being applied more frequently to address diverse scientific and regulatory issues associated with deteriorated air quality in China over the past decade. Thorough evaluation of model performance helps to illuminate the robustness and reliability of the baseline modelling results and subsequent scenarios that are built upon it. Thus, the model performance evaluation (MPE) is a critical step of any PGM application. However, with numerous input data requirements, diverse model configurations, and the scientific evolution of the models themselves, no two PGM applications are the same and their model performance results can differ significantly. Currently, a uniform set of MPE procedures and associated benchmarks is lacking in China, where air pollution problems have attracted extensive attention and the modelling community has grown substantially. Here we present a comprehensive review of PGM applications in China with the aim to propose a set of statistical benchmarks for evaluation of simulated fine particulate matter (PM2.5) concentrations in China. A total of 307 peer-reviewed articles published between 2006 and 2019, which applied five of the most frequently used PGMs in China, are compiled to summarize operational model performance results. Quantile distributions of common statistical metrics are presented for total PM2.5 and speciated components. We discuss 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. This is the first study to propose benchmarks of six frequently used evaluation metrics for PGM 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. 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 PGM applications prior to subsequent regulatory assessments.

1 Introduction

Photochemical grid models (PGMs) numerically simulate the spatial and temporal distributions of numerous chemically complex air pollutants and provide an essential component of atmospheric research by building a crucial bridge between field observations and chamber studies. With unique capabilities and features, PGMs have been utilized for a wide range of purposes, including, but not limited to, understanding the underlying formation mechanisms of secondary air pollutants and evaluating air quality impacts on public health and ecosystems. In particular, PGMs are important to air quality management programs because they are extensively used to identify source contributions as well as assist in the formulation and evaluation of control strategies. Over the past decade, tremendous efforts have been carried out by the Chinese central government to address the severe air pollution problems in China. Consequently, the number of PGM applications in China has increased tremendously.

A critical step in all PGM applications is the model performance evaluation (MPE); that is, to assess how well modelling results can replicate the observed magnitudes and spatial/temporal variations of the target pollutant. Comprehensive MPE practices help to illuminate the accuracy and reliability of modelling results from a baseline PGM simulation and therefore the reliability of subsequent applications built on top of it. However, PGMs are not constrained in the sense that there are no "uniform" settings for PGM applications (e.g. different models developed and evolved by different groups, multiple and diverse sources of input data, various model configurations and science treatments, etc.) thus MPE results from different studies vary significantly. For example, in China normalized mean bias (NMB) and correlation coefficient (R) are two of the most commonly used statistical metrics for total PM2.5 (particular matter with an aerodynamic diameter less than 2.5 μm). Reported NMB values for total PM2.5 range from large under-predictions of -73.6% (Zhang et al., 2016) to large over-estimates of 110.6% (Zhang et al., 2017); the reported R values used to reflect a model's ability to capture observed variations range from -0.59 (Gao et al., 2018) to as high as 0.98 (Feng et al., 2018). Unfortunately, the modelling community in China has no contextual references for how well or poor their model results are since there are no unified guidelines or benchmarks developed for PGM applications in China.

In the United States (U.S.) and Europe, efforts have resulted in guidance and/or benchmarks on MPE. For instance, the first modelling guidance document issued by the U.S. Environmental Protection Agency (EPA) provided a set of ozone MPE metrics for ozone attainment demonstration (EPA, 1991). Later, the concept of "goals" (*"the level of accuracy that is considered to be close to the best a model can be expected to achieve"*) and "criteria" (*"the level of accuracy that is considered to be acceptable for modelling applications"*) for model evaluation were first introduced by Boylan and Russell (2006) and later updated by Emery et al. (2017). In Europe, the Forum for Air Quality Modelling in Europe (FAIRMODE) developed a methodology to support a unified model evaluation process for modeling applied by European Union Member States (Janssen et al., 2017). Some PGM applications in China have used the U.S.-based benchmarks to assess their model robustness (e.g. J. Hu et al., 2017; D. Chen et al. 2017; Tao et al. 2018; J. Gao et al., 2017; etc.). However, it should be noted that some these benchmark studies might be outdated and all these studies are based on PGM applications in North America and may not necessarily be applicable or useful to provide needed context for applications in China, given the complex interactions of various model inputs and availability of local dataset (i.e. emission inventory, speciation database, etc.). Therefore, a set of statistics and benchmarks that are specifically targeted to evaluate PGM applications in China is urgently needed but is currently missing.

This study presents a comprehensive review of PGM applications in China over the past 15 years. The ultimate goal is to develop and recommend a set of quantitative and objective MPE benchmarks that are specifically formed from PGM applications in China. Model evaluations for criteria air pollutants including gaseous pollutants (e.g. SO_2, NO_2, ozone) and particulate matter (e.g. PM_{10}, total $PM_{2.5}$, and speciated $PM_{2.5}$) that have been published in peer reviewed journals between 2006 and 2019 were collected and analysed. We divided this work into three parts: the first part and the subject of this paper gives a general overview of air quality modelling studies in China and presents results for $PM_{2.5}$ and speciated components; results for ozone will be presented in the second part while results for other criteria pollutants including PM_{10}, SO_2, NO_2, and CO, etc. will be discussed in the third part. This is the first time that a set of quantitative and objective MPE benchmarks are recommended that are suitable for PGM applications in China. Results from this study will support the ever-growing modelling community in China by providing a more objective assessment, context for how well their results compare with previous studies, and to better demonstrate the credibility and robustness of their PGM applications prior to subsequent regulatory assessments.

We mentioned that "correlation coefficient being inconsistently used in prior studies" is simply an example of the situation where a same statistical metrics could be calculated in different ways and cautions need to be taken when interpreting the performance results.

Second, treating it as a review article I do not find the manuscript conducts an objective and comprehensive review. It does not provide any justification for the selection criteria of publications included in the review. For example, what keywords did the authors use to search those 128 papers included in the manuscript? Why was the period of publication limited to be between 2006 and 2019? Why were only four models included?

Response: We thank the reviewer for this question. As suggested by the other reviewers, we added GEOS-Chem studies, thus leading to a total of five models being investigated. We are confident that these five models represent the mainstream air quality models used in China. In the revised manuscript, we added a full description (Section 2.1) of how we conducted the selection process from the initial search on Web of Science, to screening an initial of 900+ studies down to 307 studies. The period of 2006 to 2019 was selected because few studies used air quality models prior to 2006.

In revised manuscript:

A total of five photochemical models – the Community Multiscale Air Quality (CMAQ, Foley et al., 2010), the Comprehensive Air Quality Model with Extensions (CAMx, Ramboll Environment and Health, 2018), the Goddard Earth Observing System (GEOS)-Chem (http://geos-chem.org), the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem, Grell et al., 2005), and the Nested Air Quality Prediction Modelling System (NAQPMS, Z. Wang et al. 2006) – are included in this compilation. While the former four models are developed by institutes and/or companies outside China, the NAQPMS is developed by the Institute of Atmospheric Physics of Chinese Academy of Sciences and has mostly been utilized for applications in China. GEOS-Chem is a global chemical transport model with coarser resolution (only 20% of complied GEOS-Chem studies has a grid resolution less than 50 km), as opposed to the other four regional models that are applied with finer spatial resolution at local scale (for example, less than 10 km). 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 gives 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). 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, 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. The next filtering stage needed substantial manual effort. The 464 records were downloaded and manually checked to exclude (1) studies that were accidentally included in the search but did not apply any of the models in their study; (2) studies that were intended for other purposes (for example, evaluating meteorological simulations); (3) studies that were not focused on China (for example, the target region was Korea, Japan, etc.); (4) studies that did not provide any air quality model performance evaluation or the evaluation results were referred to previous studies; (5) studies that did conduct model performance evaluation but no numerical values were given (for example, only graphical plots were given). The final selection included a total of 307 papers (see a complete list in Table S2). We defined ten regions of China as shown in Figure 1, namely Beijing-Tianjin-Hebei (BTH) region, Yangtze River Delta (YRD) region, Pearl River Delta (PRD) region, Sichuan Basin (SCB), North China Plain (NCP), Central, Northwest, Northeast, Southeast, and Southwest (see Table S3 for provinces covered in this region).

Third, being a summary of prior modeling studies, the manuscript does not make any attempt to provide useful insights on why the published model performances on PM2.5 in China vary so much as shown in their figures. Is it due to different inventories, chemistry mechanisms, or meteorological fields used? Without this type of discussion, the manuscript would not provide much value to readers.

Response: The underlying reason for diverse model performance of $PM_{2.5}$ is an important question, but yet difficult to answer. As a group, these ~300 studies utilized diverse input data (e.g. emission inventory, boundary conditions, meteorological inputs) and model formulations (e.g. gas- and aerosol chemistry), and spatial resolutions. On top of this, these studies focused on different modeling periods and regions. Therefore, they were not designed as a set of controlled experiments (i.e., varying one input/algorithm while holding the others constant) with an objective of investigating sources of model uncertainty. In addition to that, because obtaining "good model performance" is an essential requirement for publishing, it is likely that researchers have already tested alternative model configurations (e.g., with different meteorology, emissions, chemistry) before choosing and publishing the best-performing base-case results (Reynolds et al., 1996). 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. Nevertheless, we tried our best to address these questions as follows:

- First, we added time series plots of commonly used statistical metrics for $PM_{2.5}$ and speciated components (depending on data availability) reported in the literature reviewed herein to illustrate an overall trend of model performance during the past decade (see Section 3.2)

  In revised manuscript:

Trends over the past decade

In an attempt to assess whether model performance results have evolved over the past decades, we present time series of selected statistical metrics for total PM2.5 in Figure 9 (plots for inorganic species are shown in Figure S3). Results published prior to 2013 were aggregated into one group because there were a limited number of studies prior to 2013. For total $PM_{2.5}$, reported R values have remained relatively consistent over the past decade with the median fluctuating within 0.6~0.8. The ranges of reported RMSE and MB become narrower in recent years even though the number of studies has increased substantially. Reported IOA and RMSE values fluctuated upward and downward over the period. On the other hand, there seems to be an improving trend in terms of FB, FE, and NME as the reported values for these three metrics shift towards zero. For instance, the median value of reported FE decreased from 56.9% prior to 2013 to around 33% in 2019. However, it is important not to over-interpret these results as the number of studies published each year could affect the results.

- Second, we did a preliminary analysis of 176 studies that reported model performance results for $PM_{2.5}$ using the Random Forest Method to rank the influence of key model attributes such as inputs, chemical and dispersion formulations, and grid resolution (see "Analysis of feature importance based on Random Forest Method" in revised supplemental information). Our results indicate that emission inventory, grid resolution, and boundary conditions are the top three factors influencing good vs. poor model results, while the choice of model and meteorology are least important. This is a very preliminary analysis and thus we decided to include these results in the supporting materials.

  In supplemental information:

  Analysis of feature importance based on Random Forest Method

  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. 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. Like other machine learning methods, random forest is also a black box. 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 collected detailed model configurations for studies that reported results of correlation coefficient (R), index of agreement (IOA), mean bias (MB), normalized mean bias (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. 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 did not include the study region and period for FI selection because we feel 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 bad performances. For MB and NMB, we used absolute values so that deviations from zero represent the performance level. We then classified these results 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. We then ran the random forest model using the 'sklearn' module in python to obtain the FI metric and the results are shown in Figure S4. The choice of emission inventory is shown to affect the model performances most, followed by grid resolution, aerosol and gas chemistry. Meteorological input and the choice of model itself is of least importance.

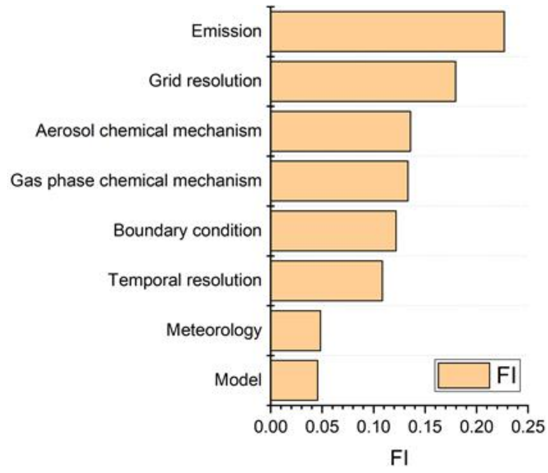

Figure S4: Ranking of key model inputs in terms of feature importance

- Last but not the least, separate from the need for model performance benchmarks, there is also 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 ozone model results. These discussions were added in Section 3.3.

In revised manuscript:

As mentioned earlier, PGM 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. A preliminary analysis based on the Random Forest Method (Liu et al., 2012), a machine learning method suitable for classification and regression, suggests that emission inventory, grid resolution and boundary conditions are the top three factors that affect model performances results (see details in Supplemental information). 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.

References:

Boylan, J. W., & Russell, A. G. (2006). PM and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models. Atmospheric Environment, 40(26), 4946-4959.

EPA, U. (2007). Guidance on the use of models and other analyses for demonstrating attainment of air quality goals for ozone, PM2. 5, and regional haze. US Environmental Protection Agency, Office of Air Quality Planning and Standards.

Dunker, A.M., Wilson, G., Bates, J.T. and Yarwood, G., 2020. Chemical Sensitivity Analysis and Uncertainty Analysis of Ozone Production in the Comprehensive Air Quality Model with Extensions Applied to Eastern Texas. Environmental Science & Technology, 54(9), pp.5391-5399.

Liu, Y., Wang, Y., & Zhang, J. (2012, September). New machine learning algorithm: Random forest. In International Conference on Information Computing and Applications (pp. 246-252). Springer, Berlin, Heidelberg.

Reynolds, S., Michaels, H., Roth, P., Tesche, T.W., Mcnally, D., Gardner, L. and Yarwood, G., 1996. Alternative base cases in photochemical modeling: their construction, role, and value. Atmospheric Environment, 30(12), pp.1977-1988.

Interactive comment on "Recommendations on benchmarks for photochemical grid model applications in China: Part I – $PM_{2.5}$ and chemical species" by Ling Huang et al.

Response to Reviewer's Comments

Anonymous Referee #3

Received and published: 6 August 2020

Severe air quality problem in China has recently attracted attention from the public as well as the scientific community. Photochemical grid models (PGMs) are frequently used to investigate the phenomenon and develop emission control strategies. The number of PGMs-based research articles for scientific and regulatory applications has surged. This study aimed to apply model performance evaluation (MPE) methodologies developed in U.S. to evaluate PGMs used in China for total PM2.5 and speciated PM components. A total of 128 recent peer reviewed articles based on one of four most popular PGMs were compiled and different model configurations as well as statistical metrics were evaluated. The benchmarks were developed and recommended for two tiers: "goals" and "criteria" for evaluating PGM applications in China. Although the methodologies and metrics used in this study are not novel, or adopted from several studies conducted in U.S., the derived results/recommendations/conclusions are scientifically sound and its logic or context is reasonable. This study is expected to provide guidance for future PGM evaluations in China.

The manuscript is well written and the logic and context are well presented and easy to follow. Several comments are provided below in hope that these will assist the authors strengthen the manuscript.

Response: We thank the reviewer for these positive comments and valuable suggestions. We have made substantial edits in the revised manuscript, which aims to strengthen the approach and findings of our study. Our major improvements include:

(1) adding GEOS-Chem results thus leading to a more complete picture of PGM applications in China;

(2) a detailed description of literature selection process is provided and the number of

peer-reviewed publications increased from 128 to 307 (Section 2.1);

3) more discussions were added to provide more in-depth insights into our findings, including trends of model performance results over the past decade (Section 3.2), application of Random Forest Model to rank feature importance of key model inputs (Section 3.3 and Supplemental information).

Our responses to the reviewer's comment are given below in blue. Revised manuscript with revisions highlighted in yellow is attached after the response.

Technical comments:

The methodologies and metrics adopted in this study are well established and published in several literature and 128 relevant modeling studies conducted in China were compiled in this study for the model performance evaluation. Although the information provided in this study is useful for the modeling community, the analysis was relatively straightforward so I would consider this study a critical literature review, instead of a novel study. To strengthen the scope of this study, one would expect that the authors go beyond what was accomplished in the U.S. studies and consider additional analyses such as the following.

Although the manuscript describes the reasons why China specific modeling performance evaluation are needed, there are many commonalities across the air quality modeling community worldwide so comparison can be made among studies conducted in China or elsewhere. Emery et al (2017) indicates that "While we primarily address U.S. modeling and regulatory settings, these recommendations are relevant to any such applications of state-of-the-science photochemical models." The comparison of benchmarks from this study with Emery et al (2017) shows similarities. Thus, it seems that the benchmarks developed in China in this study confirm their worldwide applicability for other super-regional, regional, or local modeling domains. It would be valuable if the authors discuss the broader implication of these findings.

Response: The benchmarks results are updated based on the updated compilation of studies. Currently, no guidelines on systematic and standard model performance criteria or goals are available in China to provide context for good vs. poor results relative to the growing number of applications in China. Benchmarks that were used for quantitative model performance evaluation in some of the studies are based on results of PGM applications in the U.S., some of which are outdated (for example, the "Guidance on the Use of Models and Other Analyses for Demonstrating Attainment of Air Quality Goals for Ozone, $PM_{2.5}$ and Regional Haze" developed by the U.S. EPA was published in 2007 (EPA, 2007), which is almost 15 years ago; the benchmark of FB and FE for PM introduced by Boylan and Russell, which is used in quite some studies for model performance evaluation, was published in 2006 (Boylan and Russell, 2006)) and may not be appropriate for PGM applications in China, due to different model configurations (e.g. unique environments, quality of emission inventories, availability of source emission profiles, etc.). It is appropriate to come up with benchmarks that are solely based on PGM applications in China for a more direct apple-to-apple comparison. More discussions with respect to the broader implications have been added to the revised manuscript in Section 3.3:

In revised manuscript:

As mentioned earlier, PGM 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. A preliminary analysis based on the Random Forest Method (Liu et al., 2012), a machine learning method suitable for classification and regression, suggests that emission inventory, grid resolution and boundary conditions are the top three factors that affect model performances results (see details in Supplemental information). 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.

A total of 128 peer-reviewed articles were compiled for this study. Are there articles or studies that were excluded from this study but could be potentially included by reapplying the metrics used in this study? Some of the studies may not report any MPE results but could be recalculated to get MPE results if needed. Please add some discussion on those studies, especially on those with speciated PM components since the number of these studies is very limited. If applicable, please include any additional studies so the dataset is larger or more meaningful. In addition to peer-reviewed articles, there may be non-peer-reviewed reports which deal with PGM applications (e.g., US EPA's PGM reports). I wonder if there are such reports published by Chinese central or provincial government or NGOs that can be included in this study.

Response: In our revised manuscript, we expanded our data compilation to include (1) studies that applied GEOS-Chem and (2) studies that were published in late 2019. We added a detailed description of how we selected the samples from initially 900+ Web of Science records down to a final of 307 papers used in this study. Studies that are excluded in our compilation are studies that either do not report model performance results or results are presented in graphical format (we did not want to “estimate” values from figures). It was just not feasible to recalculate MPE metrics without having the source data (i.e. simulated results and observations used). The number of studies that reported MPE results for speciated PM components has been increased to 169 studies.

As far as we understand, there are no official reports published by Chinese central or provincial governments that address PGM applications, especially with respect to model performance results. The "Manual for Compilation of Urban Air Quality Standards Planning[1]" is issued by the Clean Air Innovation Center, the Ministry of Environmental Protection, the Chinese Academy of Environmental Sciences, Tsinghua University and other institutions with the purpose of guiding cities in the preparation of air quality compliance plans and to establish a systematic air quality compliance management model accordingly. The manual includes the specific methods and steps for the preparation of the plan to achieve the standards, and air quality models (including CMAQ, CAMx, WRF-Chem, NAQPMS) were only briefly recommended as related tools to help the city complete the preparation of the air quality plan.

On the other hand, are there cases (excluded in this study) that the authors can reapply the benchmarks recommended in this study to demonstrate the improved model robustness or validity? For example, there may be PGM studies that did not use the metrics adopted in this study but the evaluation may be improved after these metrics or benchmarks are applied.

Response: It is difficult to answer this question, given that any response here or in the manuscript would be strictly hypothetical given that we don’t know what those cases entail that are excluded from this study. The point of this work and the recommended benchmarks is to provide context so that future modelers understand more directly and objectively where their simulation results stand in the universe of Chinese PGM applications. There may be some historical studies that can benefit from this new information, especially those that comprise the outer 67% of results reported here as they would now know that their simulations need to be reworked to improve performance. But it is difficult to say that the benchmarks would provide evidence for directions in which to pursue improvements as there are just too many variables involved.

Some benchmarks for speciated PM components are questionable due to the number of available studies, which may lead to biased or inconclusive results. Although caution is warranted, I wonder if the dataset can be enriched by including some studies elsewhere (e.g., U.S. studies) since benchmarks for speciated PM components were not studied in Emery et al (2017). I understand the focus of this study is in China, but it seems that

1 https://www.efchina.org/Reports-zh/report-cemp-20170928-zh

benchmarks developed in China and U.S. are valid, comparable in both countries.

Response: We have expanded the number of studies to 307 and thus the resulting dataset for speciated PM components are enriched substantially. To be consistent with total $PM_{2.5}$ as well as follow-up work on ozone and other criteria pollutants, we decide to only use results from PGM applications in China.

In-depth discussion on statistical metrics in "Impact of temporal and spatial resolution" (page 7, lines 35-37) is needed. This is counter-intuitive that the wider ranges are associated with the larger number of data points. What data is needed to improve the confidence on the benchmarks developed for speciated PM components?

Response: We included GEOS-Chem studies in the revised manuscript and the spatial resolution range from as fine as 1km to as coarse as over 100 km. Instead of presenting results by specific resolution, we grouped results into five categories of spatial resolution range: (0, 5 km], (5 km, 10 km], (10 km, 25 km], (25, 50 km], and (50 km, 100 km]. It is not necessary that winder ranges are associated with large number of data points. For example, reported values for (10, 25 km] is associated widest range but not necessarily the large number of data points. As mentioned in the manuscript, there are a lot of parameters that could influence the model performance results. Based on our analysis of feature importance using the random forest method (see "Analysis of feature importance based on Random Forest Method" in Supplemental information), grid resolution turns out to be the second most important factor for model performances. These discussions have been added to the revised manuscript.

In revised manuscript:

Impact of temporal and spatial resolution

Although PGM are usually conducted at hourly time step, validation of modelling results is not always performed with pairs of hourly data, which depends on the temporal resolution of observational data as well as the purpose of the application. Daily, monthly and even annually-averaged pairs of modelling results and observations were used for model evaluation. Of the 307 studies compiled in this work, 183 (60%) studies used hourly data for model validation, followed by 90 (29%) using daily, 31 (10%) using monthly, and only 12 (4%) studies using annual data. Due to the coarse resolution of GEOS-Chem studies in general, the finest validation is conducted in GEOS-Chem studies is using daily data. Figure 7 shows the quantile distribution of eight statistical metrics for total $PM_{2.5}$ presented by the temporal resolution used for model validation (plots for speciated components are shown in Figure S1; results for annual are not shown due to limited data). Model performances evaluated using daily-average values are similar or slightly better than hourly values but exhibit large improvements when monthly-average values are used. For instance, reported R values do not show much differences at hourly and daily scale (median values around 0.7) but exhibit a substantial improvement at monthly scale (median value around 0.85). A similar trend is also observed for reported error statistics (NME and FE), which show slight improvement as the validation resolution increase from hourly to daily but large improvement from daily to monthly. One study (Matsui et al., 2009) provided two sets of R values that calculated are based on hourly and daily-averaged data, respectively and R values for daily averages are always (12 out of 14 R values) higher.

Spatial resolution is a key setup for PGM applications. For applications at local or urban scale, PGM is usually configured with two or three nested domains that were downscaled from coarser outer domain to finer inner domain. Among the 307 articles compiled in this study, a total of 43 grid resolutions was used (for nested grids, we used the grid solution of the finest grid), ranging from as coarse as over 200 km (used by GEOS-Chem) to as fine as 1 km depending on the target region and the purpose of the application. GEOS-Chem is more often used with coarse resolution (>50 km) and the modelling grids are usually rectangular, which are converted to the side length of a square that has equivalent grid area. For simplicity, we classified these different grid resolutions into five categories: (0, 5 km], (5 km, 10 km], (10 km, 25 km], (25, 50 km], and (50 km, 100 km]. Figure 8 shows the distribution of eight statistical metrics of total $PM_{2.5}$ by these four categories (plots for speciated components are shown in Figure S2). It appears that finer spatial resolution does not necessarily improve model performances results. For example, the R values for the finest category range from as low as 0.47 to as high as 0.85 while for the coarsest category from 0.33 to 0.96. MB seems to be moving from underestimation to overestimation as grid resolution gets coarser and no clear trend is observed for FB and NMB. Reported NME and FE values seem to increase as the grid resolution gets coarser. As mentioned above, many factors could affect model performances. Thus it is difficult to solely evaluate whether there is a systematic improvement of model performances as the modelling resolution gets finer. While most of the studies only performed model evaluation for one modelling domain (usually the finest domain), a few studies (e.g. X. Qiao et al., 2015; L. Wang et al., 2015; X. Liu et al., 2010; S. Liu et al., 2018) calculated statistical results for multiple domains. L. Wang et al. (2015) reported results for evaluating hourly $PM_{2.5}$ at two spatial resolutions (12 km vs. 36 km) simultaneously. For this particular study, model over-predicted $PM_{2.5}$ at 12 km resolution (positive values of MB, NMB, and FB) but under-predicted $PM_{2.5}$ at 36 km resolution (negative values of MB, NMB, and FB). This is likely due to the dilution effect that makes model results lower at 36 km domain.

In supplemental information:

Analysis of feature importance based on Random Forest Method

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. 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. Like other machine learning methods, random forest is also a black box. 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 collected detailed model configurations for studies that reported results of correlation coefficient (R), index of agreement (IOA), mean bias (MB), normalized mean bias (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. 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 did not include the study region and period for FI selection because we feel 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 bad performances. For MB and NMB, we used absolute values so that deviations from zero represent the performance level. We then classified these results 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. We then ran the random forest model using the 'sklearn' module in python to obtain the FI metric and the results are shown in Figure S4. The choice of emission inventory is shown to affect the model performances most, followed by grid resolution, aerosol and gas chemistry. Meteorological input and the choice of model itself is of least importance.

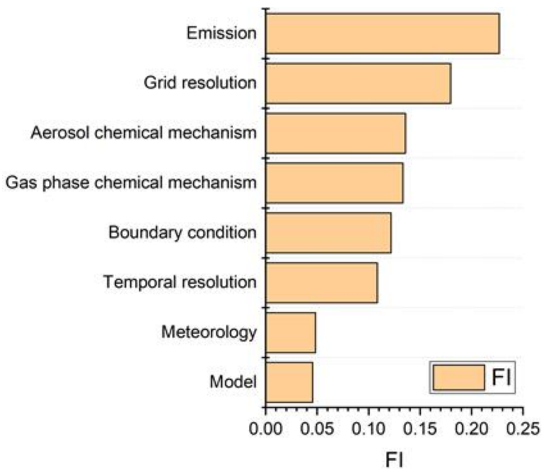

Figure S4: Ranking of key model inputs in terms of feature importance

The confidence on the benchmarks developed for speciated PM components are mostly limited by the number of available data points. Therefore, more results from studies that perform model validation against observed speciated PM would be needed to improve the confidence on developed benchmarks.

Minor comments:

Page 5, line 29: Table 2 should be Table 1.

Response: Corrected in revised manuscript.

Page 6, line 40: please check the number. It seems one single study is in spring, not summer. There are 5 studies in summer.

Response: Since the underlying dataset has been updated with more studies, this comment is not relevant.

Page 7, line 5-14: “Figure 6” is missing in this section and should appear somewhere.

Response: Added in the manuscript (Page 6, Line 39).

Page 7, line 4 and 15: “Impact” is misspelled.

Response: Corrected in revised manuscript.

Page 7, line 33-34: it seems the R values correctly correspond to the coarsest resolution (80km) but off to the finest resolution (3km).

Response: We have revised Figure 8 with updated results and grouped different resolutions into four categories. Thus this comment is not relevant.

Page 8, line 5-13: it seems that the R values in the text correspond to different percentile. For instance, the 33rd percentile value should be 0.64 for hourly to 0.91 for monthly results while the 67th percentile should be 0.5 for hourly to 0.70 for monthly. Please check the remaining values in the text against Figure 9.

Response: The reviewer is correct about this. We mistakenly reversed the values of 33rd and 67th percentile. We have corrected this in the revised manuscript with updated results.

References:

Boylan, J. W., & Russell, A. G. (2006). PM and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models. Atmospheric Environment, 40(26), 4946-4959.

EPA, U. (2007). Guidance on the use of models and other analyses for demonstrating attainment of air quality goals for ozone, PM2.5, and regional haze. US Environmental Protection Agency, Office of Air Quality Planning and Standards.

Revised manuscript

Recommendations on benchmarks for photochemical grid model applications in China: Part I – $PM_{2.5}$ and chemical species

Ling Huang[1], Yonghui Zhu[1], Hehe Zhai[1], Shuhui Xue[1], Tianyi Zhu[1], Yun Shao[1], Ziyi Liu[1], Chris Emery[2], Greg Yarwood[2], Yangjun Wang[1], Joshua Fu[3], Kun Zhang[1], Li Li[1*]

[1]School of Environmental and Chemical Engineering, Shanghai University, Shanghai, 200444, China

[2]Ramboll, Novato, California, 95995, USA

[3]Department of Civil and Environmental Engineering, University of Tennessee, Knoxville, TN 37996, USA

Correspondence to: Li Li (lily@shu.edu.cn)

Abstract

Photochemical grid models (PGMs) are being applied more frequently to address diverse scientific and regulatory issues associated with deteriorated air quality in China over the past decade. Thorough evaluation of model performance helps to illuminate the robustness and reliability of the baseline modelling results and subsequent scenarios that are built upon it. Thus, the model performance evaluation (MPE) is a critical step of any PGM application. However, with numerous input data requirements, diverse model configurations, and the scientific evolution of the models themselves, no two PGM applications are the same and their model performance results can differ significantly. Currently, a uniform set of MPE procedures and associated benchmarks is lacking in China, where air pollution problems have attracted extensive attention and the modelling community has grown substantially. Here we present a comprehensive review of PGM applications in China with the aim to propose a set of statistical benchmarks for evaluation of simulated fine particulate matter ($PM_{2.5}$) concentrations in China. A total of 307 peer-reviewed articles published between 2006 and 2019, which applied five of the most frequently used PGMs in China, are compiled to summarize operational model performance results. Quantile distributions of common statistical metrics are presented for total $PM_{2.5}$ and speciated components. We discuss 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. This is the first study to propose benchmarks of six frequently used evaluation metrics for PGM 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. 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 PGM applications prior to subsequent regulatory assessments.

1 Introduction

Photochemical grid models (PGMs) numerically simulate the spatial and temporal distributions of numerous chemically complex air pollutants and provide an essential component of atmospheric research by building a crucial bridge between field observations and chamber studies. With unique capabilities and features, PGMs have been utilized for a wide range of purposes, including, but not limited to, understanding the underlying formation mechanisms of secondary air pollutants and evaluating air quality impacts on public health and ecosystems. In particular, PGMs are important to air quality management programs because they are extensively used to identify source contributions as well as assist in the formulation and evaluation of control strategies. Over the past decade, tremendous efforts have been carried out by the Chinese central government to address the severe air pollution problems in China. Consequently, the number of PGM applications in China has increased tremendously.

A critical step in all PGM applications is the model performance evaluation (MPE); that is, to assess how well modelling results can replicate the observed magnitudes and spatial/temporal variations of the target pollutant. Comprehensive MPE practices help to illuminate the accuracy and reliability of modelling results from a baseline PGM simulation and therefore the reliability of subsequent applications built on top of it. However, PGMs are not constrained in the sense that there are no "uniform" settings for PGM applications (e.g. different models developed and evolved by different groups, multiple and diverse sources of input data, various model configurations and science treatments, etc.) thus MPE results from different studies vary significantly. For example, in China normalized mean bias (NMB) and correlation coefficient (R) are two of the most commonly used statistical metrics for total $PM_{2.5}$ (particular matter with an aerodynamic diameter less than 2.5 μm). Reported NMB values for total $PM_{2.5}$ range from large under-predictions of -73.6% (Zhang et al., 2016) to large over-estimates of 110.6% (Zhang et al., 2017); the reported R values used to reflect a model's ability to capture observed variations range from -0.59 (Gao et al., 2018) to as high as 0.98 (Feng et al., 2018). Unfortunately, the modelling community in China has no contextual references for how well or poor their model results are since there are no unified guidelines or benchmarks developed for PGM applications in China.

In the United States (U.S.) and Europe, efforts have resulted in guidance and/or benchmarks on MPE. For instance, the first modelling guidance document issued by the U.S. Environmental Protection Agency (EPA) provided a set of ozone MPE metrics for ozone attainment demonstration (EPA, 1991). Later, the concept of "goals" (*"the level of accuracy that is considered to be close to the best a model can be expected to achieve"*) and "criteria" (*"the level of accuracy that is considered to be acceptable for modelling applications"*) for model evaluation were first introduced by Boylan and Russell (2006) and later updated by Emery et al. (2017). In Europe, the Forum for Air Quality Modelling in Europe (FAIRMODE) developed a methodology to support a unified model evaluation process for modelling applied by European Union Member States (Janssen et al., 2017). Some PGM applications in China have used the U.S.-based benchmarks to assess their model robustness (e.g. J. Hu et al., 2017; D. Chen et al. 2017; Tao et al. 2018; J. Gao et al., 2017; etc.). However, it should be noted that some these benchmark studies might be outdated and all these studies are based on PGM applications in North America and may not necessarily be applicable or useful to provide needed context for applications in China, given the complex interactions of various model inputs and availability of local dataset (i.e. emission inventory, speciation database, etc.). Therefore, a set of statistics and benchmarks that are specifically targeted to evaluate PGM applications in China is urgently needed but is currently missing.

This study presents a comprehensive review of PGM applications in China over the past 15 years. The ultimate goal is to develop and recommend a set of quantitative and objective MPE benchmarks that are specifically formed from PGM applications in China. Model evaluations for criteria air pollutants including gaseous pollutants (e.g. SO_2, NO_2, ozone) and particulate matter (e.g. PM_{10}, total $PM_{2.5}$, and speciated $PM_{2.5}$) that have been published in peer reviewed journals between 2006 and 2019 were collected and analysed. We divided this work into three parts: the first part and the subject of this paper gives a general overview of air quality modelling studies in China and presents results for $PM_{2.5}$ and speciated components; results for ozone will be presented in the second part while results for other criteria pollutants including PM_{10}, SO_2, NO_2, and CO, etc. will be discussed in the third part. This is the first time that a set of quantitative and objective MPE benchmarks are recommended that are suitable for PGM applications in China. 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 PGM applications prior to subsequent regulatory assessments.

2 Methodology

2.1 Data compilation

A total of five photochemical models – the Community Multiscale Air Quality (CMAQ, Foley et al., 2010), the Comprehensive Air Quality Model with Extensions (CAMx, Ramboll Environment and Health, 2018), the Goddard Earth Observing System (GEOS)-Chem (http://geos-chem.org), the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem, Grell et al., 2005), and the Nested Air Quality Prediction Modelling System (NAQPMS, Z. Wang et al. 2006) – are included in this compilation. While the former four models are developed by institutes and/or companies outside China, the NAQPMS is developed by the Institute of Atmospheric Physics of Chinese Academy of Sciences and has mostly been utilized for applications in China. GEOS-Chem is a global chemical transport model with coarser resolution (only 20% of complied GEOS-Chem studies has a grid resolution less than 50 km), as opposed to the other four regional models that are applied with finer spatial resolution at local scale (for example, less than 10 km). 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 gives 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). 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, 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. The next filtering stage needed substantial manual effort. The 464 records were downloaded and manually checked to exclude (1) studies that were accidentally included in the search but did not apply any of the models in their study; (2) studies that were intended for other purposes (for example, evaluating meteorological simulations); (3) studies that were not focused on China (for example, the target region was Korea, Japan, etc.); (4) studies that did not provide any air quality model performance evaluation or the evaluation results were referred to previous studies; (5) studies that did conduct model performance evaluation but no numerical values were given (for example, only graphical plots were given). The final selection included a total of 307 papers (see a complete list in Table S2). We defined ten regions of China as shown in Figure 1, namely Beijing-Tianjin-Hebei (BTH) region, Yangtze River Delta (YRD) region, Pearl River Delta (PRD) region, Sichuan Basin (SCB), North China Plain (NCP), Central, Northwest, Northeast, Southeast, and Southwest (see Table S3 for provinces covered in this region).

2.2 Metrics evaluated

A total of 25 performance metrics was used in the 307 articles compiled in this study (see Supplemental Table S4 for a complete list of the 25 metrics). In general, these statistical metrics could be divided into two types: one is to indicate how well model captures the magnitude of observations. Examples of this type include mean bias (MB), normalized mean bias (NMB), fractional bias (FB), etc. The other type of statistical metrics is used to indicate how the model captures the variations of observations and most commonly used metrics are “correlation coefficient” or “index of agreement”. While some of the compiled studies explicitly provide mathematical formula of the MPE metrics used in their paper, quite many did not. This causes ambiguity when a common terminology or abbreviation was used but no explicit formula is provided. For example, the term of “correlation coefficient” (or “correlative coefficient”) is frequently used in many studies but turned out to be calculated using different mathematical formula in different studies. In some studies, the “correlation coefficient” refers to the Pearson correlation coefficient (R), which indicates the strength of linear relationship between observations and predictions; while in some studies, it refers to the coefficient of determination (R^2) that represents the fractions of predicted variations explained by observations. In these two cases, R^2 value is simply the square of R value. In two studies (X. Wang et al., 2018; H. Zhang et al., 2018), the term of “correlation coefficient” is used but formulated as the root mean square error (RMSE). To make things even more complicated, this correlation coefficient is used to indicate model’s capability of capturing temporal variations in most of the studies but also spatial variations in some cases (e.g. Ge et al., 2014). For temporal variations, this "correlation coefficient" is calculated based on temporally (hourly or daily) matched observation and modelled results at a single monitoring site (or averages across multiple monitoring sites in many cases). For spatial variations, this "correlation coefficient" is calculated based on pairs of observations and modelled results at multiple sites and its value is used to demonstrate spatial performance. To have better comparability among studies, we converted R^2 values to R. "Index of Agreement" (IOA) is another example that cautions must be taken when collecting data since the definition of IOA is not unique among these studies. Most of the studies use the definition of IOA (*d*) shown in Table 1 and only one study used the formula in Table 3. The use of IOA is discussed more in section 3.4 and we dropped the second formula for developing IOA benchmarks.

2.3 Derivation of benchmarks

In this study, the method established by Simon et al. (2012) and Emery et al. (2017) was mostly adopted. Quartile distribution for each statistical metrics (depending on the data availability) was first presented and the influences of several model key inputs on these metrics were discussed. Rank-ordered distribution for selected metrics was then used to pick out the 33rd and 67th percentiles. According to Emery et al. (2017), the 33rd and 67th percentile separates the whole distribution into three performance range: studies that fall within the 33rd percentile can be considered as successfully meeting the goals that the best a model is currently expected to achieve; studies that fall between 33rd and 67th quantiles indicate successfully meeting the criteria that the majority of studies could achieve; studies that fall outside the 67th quantile indicate relative poor performance for that specific metric. A summary table with values of 33rd and 67th quantile values for recommended statistical metrics is provided at the end this work and is compared with U.S. benchmarks proposed by Emery et al. (2017).

3. Results

3.1 General overview of air quality modelling studies in China

A total of 307 articles with PGM applications published between 2006 and 2019 were compiled in this work. Figure 2a shows the number of articles published in each year during the past 14 years. Prior to 2013, number of studies that utilized PGMs in China was generally limited. A noticeable increase of number of studies was apparent in 2013 with doubled or even tripled studies each year during 2016-2019. This sharp increase coincides with the infamous record-breaking haze event in January 2013 that attracted numerous attentions to air pollution issues in China. Since then, series of air pollution related actions were carried out due to increasing funding that became available for the research community to perform various studies related to air pollution. Of the 307 articles included in this work, CMAQ was the most frequently used PGM (used in 124 studies), followed by WRF-Chem (111 studies), CAMx (36 studies), GEOS-Chem (20 studies), and NAQPMS (18 studies). Several studies evaluated model performances for multiple models (e.g. Q. Wu et al. 2012; Zhang et al., 2016; Wang et al., 2017). In terms of regions, BTH (122 studies), YRD (84 studies), and PRD (65 studies) are the top three most evaluated regions (Figure 1) (note that we excluded studies that cover entire China for this count).

Meteorological data are needed to drive air quality simulations and the performance of meteorology modelling is one of uncertainties for air quality modelling performance. Meteorological data are dominantly 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) or the Regional Atmospheric Modelling system (RAMS) were used in a few studies. Model performances of meteorological results should be also evaluated in addition to air quality simulation results. However, we do find a few studies that did not report any results with respect to their meteorological simulations. The model performances of meteorological results used to drive air quality simulations will be discussed as a future work.

Emission inventory is another critical input for PGM applications and the accuracy of emission inventory being used no doubtfully directly affects the model performance. Most frequently used emission inventory for anthropogenic sources include the MEIC developed by Tsinghua University (http://www. meicmodel.org), Regional Emission Inventory in Asia (REAS, Kurokawa et al., 2013), Intercontinental Chemical Transport Experiment-Phase B (INTEX-B) emissions (Q. Zhang et al., 2009), MIX Asian anthropogenic emissions developed by the Model Inter-Comparison Study for Asia (MICS-Asia) emission group (M. Li et al., 2017), and many locally developed emission inventory at regional or city-scale. For biogenic emissions, the Model of Emissions of Gases and Aerosols from Nature (MEGAN, Guenther et al., 2006) is the dominant one being used.

The national monitoring stations from the China National Environmental Monitoring Center (CNEMC) are the dominant observational data source used for model validation. The coverage of the national monitoring system increased from 74 major cities in 2013 to 338 cities across China in 2018. However, since only criteria pollutants (namely $PM_{2.5}$, PM_{10}, SO_2, O_3, NO_2 and CO) are measured at the national monitoring sites, model validation of speciated $PM_{2.5}$, ammonia, volatile organic compounds (VOCs) species (e.g. isoprene, formaldehyde), and etc. are based on measurements obtained from local monitoring sites or field observations conducted by individual research groups or institutes.

Figure 2b shows the frequency of use for each statistical metric compiled in this study. Table 1 shows the formula of metrics that have been used in more than 20 studies. Same as Simon et al. (2012), the top three most frequently used metrics is correlative coefficient (R, 223 studies), normalized mean bias (NMB, 170 studies), and mean bias (MB, 132 studies). Other frequently used (>20 studies) metrics include root mean square error (RMSE, 118 studies), normalized mean error (NME, 111 studies), fraction bias (FB, 66 studies), fraction error (FE, 62 studies), index of agreement (IOA, 57 studies) and mean error (ME, 27 studies). Mean normalized bias (MNB) and mean normalized error (MNE) were only used in 15 and 10 studies, respectively, as mentioned in Simon et al. (2012) that these two metrics tends to give more weight to data at low values. About 65% of articles included in this work used at least three statistical metrics for model performance evaluation (Figure 2c); 16% of studies reported numerical values for only one metric; studies included more than five MPE metrics were less than 10%; four studies (X. Li et al., 2015; Kim et al., 2017; X. Li et al., 2018; Z. Zhang et al., 2017) used eight statistical metrics. In terms of number of pollutants evaluated in each study (Figure 2d), 132 studies (43%) evaluated only one pollutant and 223 studies (73%) evaluated less than or equal to three pollutants; one study (Ying et al., 2018) evaluated 17 pollutants (including elemental $PM_{2.5}$ components).

Figure 3 shows the number of studies broken down by pairs of pollutants and PGM models and pairs of pollutants and metrics. As expected, $PM_{2.5}$ is the most frequently evaluated pollutant, followed by ozone, NO_2, SO_2 and PM_{10}, all of which are criteria pollutants included in China's National Ambient Air Quality Standards (NAAQS). Evaluation of speciated PM species, including nitrate, sulfate, ammonium and organic carbon (OC) is about one fourth frequent as total $PM_{2.5}$ and was only covered in applications for certain regions due to limited observations.

3.2 Quantile distributions of $PM_{2.5}$ and speciated components

Figure 4 shows quantile distribution of eight most frequently used model performance metrics for $PM_{2.5}$ and speciated components (corresponding values are listed in Table S5). For total $PM_{2.5}$, slightly more negative values of MB, NMB, and FB were reported. Absolute bias for $PM_{2.5}$ ranges from as low as -50 μg/m^3 to over 50 μg/m^3 (outliers excluded) with median values around 3 μg/m^3. The bias range for speciated components is much smaller (within 20 μg/m^3) because the absolute magnitude of speciated components is much smaller. In terms of the normalized bias (i.e. NMB and FB), the range of $PM_{2.5}$ is comparable or smaller than speciated components, partly due to the compensating errors from speciated components. Speciated $PM_{2.5}$ tends to be dominantly under-estimated except for nitrate (in terms of NMB) and EC (in terms of FB). Widest range of normalized bias was reported for nitrate, suggesting substantial uncertainties in simulated nitrate concentrations. Some early reported large negative NMB values of nitrate were partly due to missing formation of coarse-mode nitrate implemented in early version of the model (Kwok et al., 2010). As the model evolves over the time, the large negative bias of nitrate disappeared (see Figure S3). Unlike other speciated components that could be both emitted directory from sources (i.e. primary) and formed via chemical reactions of precursors (i.e. secondary), elemental carbon (EC) is solely emitted from sources directly thus the performance of simulated OC concentration is more associated with the accuracy of the emission inventory. Model under-estimations of secondary species (organic and inorganic) have been reported in numerous studies with explanations of missing formation mechanisms, uncertainties with the emission inventory, and meteorology errors that were carried over, etc. For error metrics, total $PM_{2.5}$ performs better than speciated components in terms of NME, with a median NME value around 40%. For FE, median values for total $PM_{2.5}$ and carbonaceous species are within 40~60% while inorganic secondary species have relatively large (>60%) median FE.

R and IOA are used to indicate how well the model could capture the variations of observed values and both values are within the range of 0~1. We converted R^2 values to R for better comparability. For total $PM_{2.5}$, median IOA value is 0.76 while median R is 0.69 (R^2=0.4761). Minimum IOA value reported for total $PM_{2.5}$ is 0.4 while minimum R value could be negative. Eleven studies reported both R and IOA values that enable inter-comparisons of the two metrics based on identical sets of data points. It is found that IOA values always tend to be higher than R values (38 out of 40 data pairs). Compared to total $PM_{2.5}$, secondary inorganic aerosols (i.e. sulfate, nitrate, and ammonium) demonstrate better performances in terms of R values but slightly poorer performances in terms of IOA values. OM and EC show lower values for both R and IOA compared to total $PM_{2.5}$.

Impact of season

There are numerous factors that could affect model performances results, to give a few examples, the study region and period, source of emission inventory, model grid resolution, the temporal resolution of paired observations and modelling results used for model evaluation, etc. We first look at NMB results of total $PM_{2.5}$ and selected species (due to availability of data points) by season (Figure 5). For total $PM_{2.5}$, number of data points reported for winter is significantly higher than those reported for other seasons as heavy haze episodes generally occur in winter. Reported NMB for total $PM_{2.5}$ was dominantly negative except for winter where positive NMB is also reported. Most of the large positive NMB values (>50%) reported for winter is from one single study (Zhang et al., 2017), for which the author explained that the over-estimation may be associated with the inconsistency of emissions between the base year and the modeling year. Sulfate tends to be overwhelmingly underestimated regardless of season, which is commonly reported in literatures with potential causes of missing formation mechanisms (e.g. heterogeneous reactions, Ye et al., 2018; L. Huang et al. 2019; Shao et al., 2019; Chen et al., 2019). However, a few large NMB values (>50%) were reported in fall (Cheng et al., 2019). Nitrate and ammonium exhibits equivalent over- and under-estimation for all seasons and differences among seasons are minor. OM also tends to be more underestimated, especially in summer and fall. The underestimation of organic components, especially the secondary organic aerosols (SOA), is well documented by many studies (e.g. Jimenez et al., 2009; Q. Chen et al., 2017; B. Zhao et al., 2016). The two positive NMB values reported for winter is from one study (Li et al., 2018) and uncertainties in anthropogenic emissions were explained for the model bias.

Impact of region

We also look at whether there are any regional differences in these statistical metrics. Constrained by number of data points, we only compared results of R and NMB for total $PM_{2.5}$ and secondary inorganic species over three key regions in China, that is the Beijing-Tianjin-Hebei (BTH) region in north China, the Yangtze River Delta (YRD) region in eastern China, and the Pearl River Delta (PRD) region in south China (Figure 6). These three regions represent the most populated, economically developed and urbanized city clusters in China. With respect to the total $PM_{2.5}$, R and NMB values for the three regions do not exhibit substantial differences. All three regions have more negative NMB values in terms of $PM_{2.5}$ with median NMB around -10%. Reported R values for PRD are slightly lower compared with the other two regions. For sulfate and ammonium, reported R values for YRD are significantly lower than the other two regions. Underestimation of sulfate and ammonium is more severe in YRD and PRD. For nitrate, PRD shows the lowest R values and model bias shifts from positive to negative as the target region gets warmer.

Impact of temporal and spatial resolution

Although PGM are usually conducted at hourly time step, validation of modelling results is not always performed with pairs of hourly data, which depends on the temporal resolution of observational data as well as the purpose of the application. Daily, monthly and even annually-averaged pairs of modelling results and observations were used for model evaluation. Of the 307 studies compiled in this work, 183 (60%) studies used hourly data for model validation, followed by 90 (29%) using daily, 31 (10%) using monthly, and only 12 (4%) studies using annual data. Due to the coarse resolution of GEOS-Chem studies in general, the finest validation is conducted in GEOS-Chem studies is using daily data. Figure 7 shows the quantile distribution of eight statistical metrics for total $PM_{2.5}$ presented by the temporal resolution used for model validation (plots for speciated components are shown in Figure S1; results for annual are not shown due to limited data). Model performances evaluated using daily-average values are similar or slightly better than hourly values but exhibit large improvements when monthly-average values are used. For instance, reported R values do not show much differences at hourly and daily scale (median values around 0.7) but exhibit a substantial improvement at monthly scale (median value around 0.85). A similar trend is also observed for reported error statistics (NME and FE), which show slight improvement as the validation resolution increase from hourly to daily but large improvement from daily to monthly. One study (Matsui et al., 2009) provided two sets of R values that calculated are based on hourly and daily-averaged data, respectively and R values for daily averages are always (12 out of 14 R values) higher.

Spatial resolution is a key setup for PGM applications. For applications at local or urban scale, PGM is usually configured with two or three nested domains that were downscaled from coarser outer domain to finer inner domain. Among the 307 articles compiled in this study, a total of 43 grid resolutions was used (for nested grids, we used the grid solution of the finest grid), ranging from as coarse as over 200 km (used by GEOS-Chem) to as fine as 1 km depending on the target region and the purpose of the application. GEOS-Chem is more often used with coarse resolution (>50 km) and the modelling grids are usually rectangular, which are converted to the side length of a square that has equivalent grid area. For simplicity, we classified these different grid resolutions into five categories: (0, 5 km], (5 km, 10 km], (10 km, 25 km], (25, 50 km], and (50 km, 100 km]. Figure 8 shows the distribution of eight statistical metrics of total $PM_{2.5}$by these four categories (plots for speciated components are shown in Figure S2). It appears that finer spatial resolution does not necessarily improve model performances results. For example, the R values for the finest category range from as low as 0.47 to as high as 0.85 while for the coarsest category from 0.33 to 0.96. MB seems to be moving from underestimation to overestimation as grid resolution gets coarser and no clear trend is observed for FB and NMB. Reported NME and FE values seem to increase as the grid resolution gets coarser. As mentioned above, many factors could affect model performances. Thus it is difficult to solely evaluate whether there is a systematic improvement of model performances as the modelling resolution gets finer. While most of the studies only performed model evaluation for one modelling domain (usually the finest domain), a few studies (e.g. X. Qiao et al., 2015; L. Wang et al., 2015; X. Liu et al., 2010; S. Liu et al., 2018) calculated statistical results for multiple domains and results based on finer spatial resolution are generally better than those based on coarser resolution. For instance, L. Wang et al. (2015) reported results for evaluating hourly $PM_{2.5}$ at two spatial resolutions (12 km vs. 36 km) simultaneously. For this particular study, model over-predicted $PM_{2.5}$ at 12 km resolution (positive values of MB, NMB, and FB) but under-predicted $PM_{2.5}$ at 36 km resolution (negative values of MB, NMB, and FB). This is likely due to the dilution effect that makes model results lower at 36 km domain.

Trends over the past decade

In an attempt to assess whether model performance results have evolved over the past decades, we present time series of selected statistical metrics for total $PM_{2.5}$ in Figure 9 (plots for inorganic species are shown in Figure S3). Results published prior to 2013 were aggregated into one group because there were a limited number of studies prior to 2013. For total $PM_{2.5}$, reported R values have remained relatively consistent over the past decade with the median fluctuating within 0.6~0.8. The ranges of reported RMSE and MB become narrower in recent years even though the number of studies has increased substantially. Reported IOA and RMSE values fluctuated upward and downward over the period. On the other hand, there seems to be an improving trend in terms of FB, FE, and NME as the reported values for these three metrics shift towards zero. For instance, the median value of reported FE decreased from 56.9% prior to 2013 to around 33% in 2019. However, it is important not to over-interpret these results as the number of studies published each year could affect the results.

3.3 Recommended metrics and benchmarks

We presented similar diagrams as Emery et al. (2017) to develop metrics and benchmarks for model evaluation. Figure 10 shows the rank-ordered distribution of R, IOA, NMB, NME, FB, and FE results for total $PM_{2.5}$ and speciated components from all studies compiled in this work. Results of R for total $PM_{2.5}$ are further split into hourly (h), daily (d) and monthly (m) resolution since it increases as temporal resolution changes from hourly to monthly. The top 33rd percentile value increases from around 0.76 for hourly and daily to 0.92 for monthly results; the top 67th percentile increases from 0.60 to 0.70 as the total $PM_{2.5}$ is evaluated with coarser resolution. Secondary inorganic species (sulfate, nitrate and ammonium) show similar range (0.65 ~ 0.75) over the 33rd – 67th percentile interval. For OC/OM and EC, the 33rd and 67th percentile R value is lower compared to inorganic species; the 33rd to 67th percentile for OC/OM is 0.56~0.69 for OC/OM and 0.48~0.65 for EC. In terms of IOA, the 33rd – 67th percentile interval ranges from 0.73 to 0.83 for total $PM_{2.5}$ and lower for speciated components. Values for EC were not shown due to limited data. For bias and error, total $PM_{2.5}$ exhibits smaller values compare with speciated components, due to potential compensating effects from different components. The 33rd percentile of absolute NMB for total $PM_{2.5}$ is less than 10% while the 67th percentiles less than 20%. Among these three secondary inorganic species, the bias and error of nitrate exhibits largest variability (NMB ranges from 17.4% to 55.2% and NME from 47.0% to 71.2% for 33rd to 67th percentile interval). The 33rd to 67th range of NMB for EC (16.0% to 32.4%) is much lower than that for OC/OM (34.7% to 55.0%) while NME for OC/OM and EC is similar, ranging from ~40% to 55%. Number of FB and FE data is considerably less than NMB and NME for speciated components and nitrate exhibits largest variability in terms of FB and FE.

Based on our analysis above as well as previous conclusions from Emery et al. (2017), we propose recommended statistical metrics and associated benchmarks for total $PM_{2.5}$ and speciated component as shown in Table 2. Shaded values indicate that less than 20 data points were available to develop the benchmarks. Values for "goal" indicate that roughly the top one third of studies could meet the benchmarks and represent the best that a model is currently expected to achieve. Values for "criteria" indicate that roughly the top two thirds of studies meet the benchmarks and represent results from the majority of studies. Our table differs from Emery et al. (2017) in three aspects. Firstly, we added benchmarks for IOA in addition to the correlation coefficient. We found a general increasing trend of using IOA for model performance evaluation since 2013 (prior to 2013, only six of our compiled studies used IOA; after 2013, 51 studies used IOA). Thus we added IOA for future reference. Secondly, we presented benchmarks for different temporal resolution of total $PM_{2.5}$ when possible. As mentioned above, reported R values for total $PM_{2.5}$ get better as temporal resolution gets coarser while no strong trend is observed for other metrics. Therefore, different benchmarks are developed for R. Thirdly, Emery et al. (2017) did not present benchmarks for the correlation coefficient of speciated PM components due to large uncertainties. Here we presented benchmarks for R and IOA of speciated PM components (except IOA for EC is not available), but cautions should be taken comparing to these benchmarks. For example, less than twenty data points were used to develop the benchmarks of IOA for ammonium and sulfate. For sulfate and ammonium, we do not observe sudden changes in the rank-order distribution as observed in Emery et al. (2017). Thus, we keep these values for future references. For bias and error metrics, we do observe sharp changes in rank-order values, for example, the NMB/FB for nitrate, FB for EC. Therefore, we do not give benchmarks in this situation. We also presented benchmarks for FB and FE.

We further compared our results with benchmarks proposed by Emery et al. (2017). Values with an asterisk in Table 2 indicate that our benchmarks are stricter than corresponding values in Emery et al. (2017), which means results from a study would be more difficult to be considered within 33^{th} (or 67^{th}) percentiles if our benchmarks are used. For total $PM_{2.5}$, our proposed benchmarks are generally stricter than that in Emery et al. (2017). For example, our NMB (NME) "criteria" value for $PM_{2.5}$ is 20% (40%) as opposed to 30% (50%) in Emery's study; "criteria" value for R benchmark is also higher (0.55) than those based on U.S. studies (0.40). This might partially reflect the systematic improvements in model applications (e.g. incorporation of newly discovered mechanisms) during the past several years since the latest study included in Emery et al. (2017) was published in 2015. Our "goal" values for NMB, NME and R benchmarks are same as that proposed by Emery et al. (2017). For speciated components, NME benchmarks for nitrate are lower (i.e. stricter) than Emery's study while the opposite is true for sulfate and ammonium. For correlation coefficient, our criteria benchmarks for sulfate and ammonium (0.60) are much higher (i.e. more strict) than those in Emery's study (0.40).

As mentioned earlier, PGM 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. A preliminary analysis based on the Random Forest Method (Liu et al., 2012), a machine learning method suitable for classification and regression, suggests that emission inventory, grid resolution and boundary conditions are the top three factors that affect model performances results (see details in Supplemental information). 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.

3.4 Additional discussions and recommendations

Benchmarks for European modeling community - FAIRMODE

The air quality model benchmarking practise for PGM applications by the FAIRMODE community is somehow different from the U.S. benchmarks. The main modeling performance indicator is called the modeling quality indicator (MQI), which is calculated based on RMSE and measurement uncertainties (function of mean value and standard deviation of observations) (Janssen et al., 2017). The modeling quality objective (MQO) is the criteria value for MQI and is said to be met if MQI is less than or equal to one. In addition to the main MQI, three statistical indicators that describe certain aspects of the differences between observed and modeled results – namely bias, correlation, and standard deviation are proposed as the modelling performance indicators (MPI). For each MPI, the model performance criterion (MPC) that individual MPI is expected to meet is also given. However, unlike fixed values given in this study and Emery et al. (2017), MPC is dependent on observation uncertainties. Therefore, it is not directly comparable between MPC and the benchmarks proposed in this study or the ones in Emery et al. (2017).

The use of "index of agreement"

The concept of "index of agreement" is originally proposed by Willmott in the 1980s and has since then been widely used to "*reflect the degree to which the observed variate is accurately estimated by the simulated variate*" (Willmott, 1981) in a variety of fields. IOA has gone through several modifications (together referred as Willmott indices) since it was proposed in the original formula (Willmott 1982; Willmott et al., 1985, 2012). The formula of the original one (d) is shown in Table 2 (presented again in Table 3) and the other three (d_1, d_1' and d_r) shown in Table 3. The first version of IOA is proposed over the correlation coefficient for its ability to "*discern differences in proportionality and/or constant additive differences between the two variables*" (Willmott, 1981) and this version is also the most widely used version in our compiled studies.

Compared with R^2 values, the original IOA results systematically higher values (Valbuena et al., 2019) thus is being adopted in an increasing number of studies partially because it makes results appear “better”. However, the original and also being the most widely used IOA is problematic in that too much weight is given to the large errors when squared (Willmott et al., 2012) and relatively high IOA values could be obtained even when a model is performing poorly (Willmott et al., 1985; Pereira et al., 2018). Newer versions as later proposed by Willmott overcome this problem by removing the squaring and are recommended over the original one (Willmott et al., 1985, 2012). Valbuena et al. (2019) suggested using d^2 instead of d, at least for estimating forest biomass based on remote sensing to facilitate comparison with studies using correlation coefficient.

Near 20% (57 studies) of our compiled studies used the “index of agreement” for MPE but only one study (Y. Peng et al. 2011) used the second formula (d_1) while the rest studies all used the original formula. There seems to be an increasing trend of using IOA (the original formula) as a model performance indicator for PGM applications in China (prior to 2013 only 6 study vs. 51 studies after 2013), we decided to keep IOA based results and discussions in this work for future reference but cautions should be taken when using and interpreting IOA values. It should be noted that the value of IOA alone does not necessarily tell how well the modelling results are.

Additional recommendations

Other than the recommended metrics and associated benchmarks listed in Table 2, we list additional recommendations for validation practices that would enable a complete and comprehensive picture of model performances.

(1) Provide explicit mathematical formula of statistical metrics being used to avoid any confusion. As mentioned earlier, quite many studies did not give explicit formula of used metrics in their studies. This would sometimes cause ambiguity when a common name (for example, correlation coefficient, or index of agreement) is used but calculated using different formula.

(2) Provide as much details as possible with respect to how observation and modelling results are used to obtain the statistical results. For example, how observed data and modelled results are paired in space and time? Is any averaging performed prior to calculating statistical metrics? Specify the number of observation sites and the number of available data points being used. This would enable a further comparison of model performances based on the amount of available data points. It should be noted that large averaging (i.e. more pairing of observed and modelled results) usually result in better statistics, but do not convey any more meaning.

(3) It is always good practise to present model performance results of meteorological fields, usually including but not limited to temperature, humidity, wind speed, and wind direction. Performance results of meteorological model could also help explain potential causes of unsatisfactory PGM simulated results.

(4) Metrics used should always include two types of statistical metrics for model evaluation, one for magnitude evaluation (e.g. MB, NMB or FB) and one for variation evaluation (e.g. R or IOA). According to Simon et al. (2012), a minimum

set of MPE statistical metrics should include "*mean observation, mean prediction, MB, ME (or RMSE) and a normalized bias and error (NMB/NME or FB/FE)*". Cautions need to be taken when presenting values of fractional metrics, for example, NMB/NME, FB/FE. Double check if the values presented are before or after multiplied with 100%. We do find studies that present extremely small values of NMB (<1%) but should be multiplied by 100 based on the results of other evaluation metrics.

(5) Try to evaluate multiple pollutants even if the study focuses on one single pollutant. It is obvious that opposite biases in speciated PM components could compensate each other and falsely lead to a good performance of the total $PM_{2.5}$.

(6) 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. Spatial plots with modelling results as background and observation data as dots could help demonstrate how model performs spatially.

4 Conclusions

With the increasing number of PGM applications in China over the past decade, a review of the model performance is needed to help understand how well these models are currently performing compared with observations and how reliable the future model applications are compared with existing studies. Following an established method used in the U.S., a total of 307 peer-reviewed studies that applied PGMs in China was compiled in this work and key information, including model applied, study region, grid resolution, evaluated metrics, and etc., were collected. As an initial attempt, operational MPE results for total $PM_{2.5}$ and speciated components reported in the compiled studies are presented in this study; results for other pollutants and meteorological simulations will be discussed as follow-up studies. Quantile distributions of common statistical metrics used in the literature were presented and the impacts of different model configurations, including study region, study period, spatial and temporal resolutions on performance results are discussed. With the concept of "goals" and "criteria", we proposed benchmarks for six commonly used metrics – NMB, NME, FB, FE, R and IOA based on the method employed by Emery et al. (2017). For total $PM_{2.5}$, we provided R benchmarks with different temporal resolutions; for component species, we did not split results by temporal resolution due to limited number of data points. We kept results for index of agreement while recognizing it should be used and interpreted with cautions. Additional recommendations on good evaluation practices are provided at the end. Results from this study could help the ever-growing modelling community in China to have a better understanding of how their model performances are compared with existing studies and also help modellers to conduct model evaluation in a more consistent fashion, which would in turn improve the comparability among different studies.

Date availability. All data is available upon request from the corresponding author.

Competing interest. The authors declare that they have no conflict of interest.

Special issue statement. This article is part of the special issue "Regional assessment of air pollution and climate change over East and Southeast Asia: results from MICS-Asia Phase III". It is not associated with a conference.

Author contribution. L.H., Y. W. and L.L. designed the research; Y. Z., H. Z., S. X, T. Z., and Y. S. complied studies and collected data with equal contributions; L.H. and Y.Z. reviewed and analyzed collected data; C. E, J. F., and G. Y. provided important academic guidance; L.H. wrote the paper with contributions from all authors.

Acknowledgement. The authors would like to recognize the contribution of Qian Wang, Hongli Li, Xin Yi, Lishu Shi, Liumei Yang, Ansheng Zhu, Hanqing Liu, Sen Jiang, Fangting Wang, and Jin Xue. This study was financially sponsored by the Shanghai Sail Program (NO. 19YF1415600), the Shanghai Science and Technology Innovation Plan (NO. 19DZ1205007), the National Natural Science Foundation of China (NO. 41875161, 42005112, 42075144), the Shanghai International Science and Technology Cooperation Fund (NO. 19230742500), and Chinese National Key Technology R&D Program (NO. 2014BAC22B03 and NO. 2018YFC0213800).

References

Boylan, J. W., and Russell, A. G.: PM and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models, Atmospheric Environment, 40, 4946-4959, https://doi.org/10.1016/j.atmosenv.2005.09.087, 2006.

Chen, D., Liu, X., Lang, J., Zhou, Y., Wei, L., Wang, X., and Guo, X.: Estimating the contribution of regional transport to $PM_{2.5}$ air pollution in a rural area on the North China Plain, Science of the Total Environment, 583, 280-291, https://doi.org/10.1016/j.scitotenv.2017.01.066, 2017.

Chen, L., Gao, Y., Zhang, M., Fu, J. S., Zhu, J., Liao, H., Li, J., Huang, K., Ge, B., Wang, X., Lam, Y. F., Lin, C.-Y., Itahashi, S., Nagashima, T., Kajino, M., Yamaji, K., Wang, Z., and Kurokawa, J.-i.: MICS-Asia III: multi-model comparison and evaluation of aerosol over East Asia, Atmospheric Chemistry and Physics, 19, 11911-11937, https://doi.org/10.5194/acp-19-11911-2019, 2019.

Chen, Q., Fu, T. M., Hu, J., Ying, Q., and Zhang, L.: Modelling secondary organic aerosols in China, National Science Review, 4, 806-809, https://doi.org/10.1093/nsr/nwx143, 2017.

Cheng, J., Su, J., Cui, T., Li, X., Dong, X., Sun, F., Yang, Y., Tong, D., Zheng, Y., Li, Y., Li, J., Zhang, Q., and He, K.: Dominant role of emission reduction in $PM_{2.5}$ air quality improvement in Beijing during 2013–2017: a model-based decomposition analysis, Atmospheric Chemistry and Physics, 19, 6125-6146, https://doi.org/10.5194/acp-19-6125-2019, 2019.

Dunker, A.M., Wilson, G., Bates, J.T. and Yarwood, G.: Chemical Sensitivity Analysis and Uncertainty Analysis of Ozone Production in the Comprehensive Air Quality Model with Extensions Applied to Eastern Texas, Environmental Science & Technology, 54, 5391-5399, https://doi.org/10.1021/acs.est.9b07543, 2020.

Emery, C., Liu, Z., Russell, A. G., Odman, M. T., Yarwood, G., and Kumar, N.: Recommendations on statistics and benchmarks to assess photochemical model performance, Journal of the Air & Waste Management Association, 67, 582-598, https://doi.org/10.1080/10962247.2016.1265027, 2017.

Feng, S., Jiang, F., Jiang, Z., Wang, H., Cai, Z., and Zhang, L.: Impact of 3DVAR assimilation of surface $PM_{2.5}$ observations on $PM_{2.5}$ forecasts over China during wintertime, Atmospheric Environment, 187, 34-49, https://doi.org/10.1016/j.atmosenv.2018.05.049, 2018.

Foley, K. M., Roselle, S. J., Appel, K. W., Bhave, P. V., Pleim, J. E., Otte, T. L., Mathur, R., Sarwar, G., Young, J. O., Gilliam, R. C., Nolte, C. G., Kelly, J. T., Gilliland, A. B., and Bash, J. O.: Incremental testing of the Community Multiscale Air Quality (CMAQ) modeling system version 4.7, Geoscientific Model Development, 3, 205, https://doi.org/10.5194/gmd-3-205-2010, 2010.

Gao, J., Zhu, B., Xiao, H., Kang, H., Hou, X., Yin, Y., Zhang, L., and Miao, Q.: Diurnal variations and source apportionment of ozone at the summit of Mount Huang, a rural site in Eastern China, Environmental Pollution, 222, 513-522, https://doi.org/10.1016/j.envpol.2016.11.031, 2017.

Gao, M., Ji, D., Liang, F., and Liu, Y.: Attribution of aerosol direct radiative forcing in China and India to emitting sectors, Atmospheric Environment, 190, 35-42, https://doi.org/10.1016/j.atmosenv.2018.07.011, 2018.

Ge, B. Z., Wang, Z. F., Xu, X. B., Wu, J. B., Yu, X. L., and Li, J.: Wet deposition of acidifying substances in different regions of China and the rest of East Asia: Modeling with updated NAQPMS, Environmental Pollution, 187, 10-21, https://doi.org/10.1016/j.envpol.2013.12.014, 2014.

Grell, G. A., Dudhia, J., and Stauffer, D. R.: A description of the fifth-generation Penn State/NCAR mesoscale model (MM5), https://doi.org/10.5065/D60Z716B, 1994.

Grell, G. A., Peckham, S. E., Schmitz, R., McKeen, S. A., Frost, G., Skamarock, W. C., and Eder, B.: Fully coupled "online" chemistry within the WRF model, Atmospheric Environment, 39, 6957-6975, https://doi.org/10.1016/j.atmosenv.2005.04.027, 2005.

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P. I., and Geron, C.: Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature), Atmospheric Chemistry and Physics, 6, 3181-3210, https://doi.org/10.5194/acp-7-4327-2007, 2006.

Hu, J., Li, X., Huang, L., Qi, Y., Zhang, Q., Zhao, B., Wang, S., and Zhang, H.: Ensemble prediction of air quality using the

WRF/CMAQ model system for health effect studies in China, Atmospheric Chemistry and Physics, 17, 13103, https://doi.org/10.5194/acp-17-13103-2017, 2017.

Huang, L., An, J., Koo, B., Yarwood, G., Yan, R., Wang, Y., Huang, C., and Li, L.: Sulfate formation during heavy winter haze events and the potential contribution from heterogeneous SO_2 + NO_2 reactions in the Yangtze River Delta region, China, Atmospheric Chemistry and Physics, 19, 14311-14328, https://doi.org/10.5194/acp-19-14311-2019, 2019.

Janssen, S., Guerreiro, C., Viane, P., Georgieva, E., Thunis, P., Cuvelier, K., ... and Stocker, J.: Guidance Document on Modelling Quality Objectives and Benchmarking– FAIRMODE WG1, https://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/Guidance_MQO_Bench_vs2.1.pdf (accessed on March 3, 2020), 2017.

Jimenez, J. L., Canagaratna, M. R., Donahue, N. M., Prevot, A. S. H., Zhang, Q., Kroll, J. H., ... and Aiken, A. C.: Evolution of organic aerosols in the atmosphere, Science, 326, 1525-1529, https://doi.org/10.1126/science.1180353, 2009.

Kim, B.-U., Bae, C., Kim, H. C., Kim, E., and Kim, S.: Spatially and chemically resolved source apportionment analysis: Case study of high particulate matter event, Atmospheric Environment, 162, 55-70, https://doi.org/10.1016/j.atmosenv.2017.05.006, 2017.

Kurokawa, J., Ohara, T., Morikawa, T., Hanayama, S., Janssens-Maenhout, G., Fukui, T., Kawashima, K., and Akimoto, H.: Emissions of air pollutants and greenhouse gases over Asian regions during 2000–2008: Regional Emission inventory in ASia (REAS) version 2, Atmospheric Chemistry and Physics, 13, 11019-11058, https://doi.org/10.5194/acp-13-11019-2013, 2013.

Kwok, R. H. F., Fung, J. C. H., Lau, A. K. H., and Fu, J. S.: Numerical study on seasonal variations of gaseous pollutants and particulate matters in Hong Kong and Pearl River Delta Region, Journal of Geophysical Research, 115, https://doi.org/10.1029/2009jd012809, 2010.

Li, M., Zhang, Q., Kurokawa, J. I., Woo, J. H., He, K., Lu, Z., ... and Cheng, Y.: MIX: a mosaic Asian anthropogenic emission inventory under the international collaboration framework of the MICS-Asia and HTAP, Atmospheric Chemistry and Physics (Online), 17, 935–963, https://doi.org/10.5194/acp-17-935-2017, 2017.

Li, X., Zhang, Q., Zhang, Y., Zheng, B., Wang, K., Chen, Y., Wallington, T. J., Han, W., Shen, W., Zhang, X., and He, K.: Source contributions of urban $PM_{2.5}$ in the Beijing–Tianjin–Hebei region: Changes between 2006 and 2013 and relative impacts of emissions and meteorology, Atmospheric Environment, 123, 229-239, https://doi.org/10.1016/j.atmosenv.2015.10.048, 2015.

Li, X., Wu, J., Elser, M., Feng, T., Cao, J., El-Haddad, I., Huang, R., Tie, X., Prévôt, A. S. H., and Li, G.: Contributions of residential coal combustion to the air quality in Beijing–Tianjin–Hebei (BTH), China: a case study, Atmospheric Chemistry and Physics, 18, 10675-10691, https://doi.org/10.5194/acp-18-10675-2018, 2018.

Liu, S., Hua, S., Wang, K., Qiu, P., Liu, H., Wu, B., Shao, P., Liu, X., Wu, Y., Xue, Y., Hao, Y., and Tian, H.: Spatial-temporal variation characteristics of air pollution in Henan of China: Localized emission inventory, WRF/Chem simulations and potential source contribution analysis, Science of the Total Environment, 624, 396-406, https://doi.org/10.1016/j.scitotenv.2017.12.102, 2018.

Liu, X.-H., Zhang, Y., Cheng, S.-H., Xing, J., Zhang, Q., Streets, D. G., Jang, C., Wang, W.-X., and Hao, J.-M.: Understanding of regional air pollution over China using CMAQ, part I performance evaluation and seasonal variation, Atmospheric Environment, 44, 2415-2426, https://doi.org/10.1016/j.atmosenv.2010.03.035, 2010.

Liu, Y., Wang, Y., and Zhang, J.: New machine learning algorithm: Random forest, In International Conference on Information Computing and Applications ,Springer, Berlin, Heidelberg, 14 September 2012, 246-252, 2012.

Matsui, H., Koike, M., Kondo, Y., Takegawa, N., Kita, K., Miyazaki, Y., Hu, M., Chang, S. Y., Blake, D. R., Fast, J. D., Zaveri, R. A., Streets, D. G., Zhang, Q., and Zhu, T.: Spatial and temporal variations of aerosols around Beijing in summer 2006: Model evaluation and source apportionment, Journal of Geophysical Research, 114, https://doi.org/10.1029/2008jd010906, 2009.

Peng, Y. P., Chen, K. S., Wang, H. K., Lai, C. H., Lin, M. H., and Lee, C. H.: Applying model simulation and photochemical indicators to evaluate ozone sensitivity in southern Taiwan, Journal of Environmental Sciences, 23, 790-797,https://doi.org/10.1016/S1001-0742(10)60479-2, 2011.

Pereira, H. R., Meschiatti, M. C., Pires, R. C. D. M., and Blain, G. C.: On the performance of three indices of agreement: an easy-to-use r-code for calculating the Willmott indices, Bragantia, 77, 394-403, 10.1590/1678-4499.2017054, 2018.

Qiao, X., Tang, Y., Hu, J., Zhang, S., Li, J., Kota, S. H., Wu, L., Gao, H., Zhang, H., and Ying, Q.: Modeling dry and wet deposition of sulfate, nitrate, and ammonium ions in Jiuzhaigou National Nature Reserve, China using a source-oriented CMAQ model: Part I. Base case model results, Science of the Total Environment, 532, 831-839, https://doi.org/10.1016/j.scitotenv.2015.05.108, 2015.

Ramboll Environment and Health. (2018). User’s Guide: Comprehensive Air quality Model with extensions, Version 6.50. Ramboll, Novato, CA (www.camx.com).

Shao, J., Chen, Q., Wang, Y., Lu, X., He, P., Sun, Y., ... and Zhao, Y.: Heterogeneous sulfate aerosol formation mechanisms during wintertime Chinese haze events: air quality model assessment using observations of sulfate oxygen isotopes in Beijing, Atmospheric Chemistry and Physics, 19, 6107-6123, https://doi.org/10.5194/acp-2018-1352, 2019.

Simon, H., Baker, K. R., and Phillips, S.: Compilation and interpretation of photochemical model performance statistics published between 2006 and 2012, Atmospheric Environment, 61, 124-139, https://doi.org/10.1016/j.atmosenv.2012.07.012, 2012.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Wang, W., and Powers, J. G.: A description of the advanced research WRF version 2 (No. NCAR/TN-468+ STR), National Center For Atmospheric Research Boulder Co Mesoscale and Microscale Meteorology Div, 2005.

Tao, H., Xing, J., Zhou, H., Chang, X., Li, G., Chen, L., and Li, J.: Impacts of land use and land cover change on regional meteorology and air quality over the Beijing-Tianjin-Hebei region, China, Atmospheric Environment, 189, 9-21, https://doi.org/10.1016/j.atmosenv.2018.06.033, 2018.

Valbuena, R., Hernando, A., Manzanera, J. A., Görgens, E. B., Almeida, D. R., Silva, C. A., and García-Abril, A.: Evaluating observed versus predicted forest biomass: R-squared, index of agreement or maximal information coefficient?, European Journal of Remote Sensing, 52, 345-358, https://doi.org/10.1080/22797254.2019.1605624, 2019.

Wang, L., Wei, Z., Wei, W., Fu, J. S., Meng, C., and Ma, S.: Source apportionment of $PM_{2.5}$ in top polluted cities in Hebei, China using the CMAQ model, Atmospheric Environment, 122, 723-736, https://doi.org/10.1016/j.atmosenv.2015.10.041, 2015.

Wang, X., Wei, W., Cheng, S., Li, J., Zhang, H., and Lv, Z.: Characteristics and classification of $PM_{2.5}$ pollution episodes in Beijing from 2013 to 2015, Science of the Total Environment, 612, 170-179, https://doi.org/10.1016/j.scitotenv.2017.08.206, 2018.

Wang, Z., Li, J., Wang, X., Pochanart, P., and Akimoto, H.: Modeling of regional high ozone episode observed at two mountain sites (Mt. Tai and Huang) in East China, Journal of Atmospheric Chemistry, 55, 253-272, https://doi.org/10.1007/s10874-006-9038-6, 2006.

Wang, Z., Itahashi, S., Uno, I., Pan, X., Osada, K., Yamamoto, S., Nishizawa, T., Tamura, K., and Wang, Z.: Modeling the Long-Range Transport of Particulate Matters for January in East Asia using NAQPMS and CMAQ, Aerosol and Air Quality Research, 17, 3064-3078, https://doi.org/10.4209/aaqr.2016.12.0534, 2017.

Willmott, C. J.: On the validation of models, Physical Geography, 2, 184-194, https://doi.org/10.1080/02723646.1981.10642213, 1981.

Willmott, C. J.: Some comments on the evaluation of model performance, Bulletin of the American Meteorological Society, 63, 1309-1313, https://doi.org/10.1175/1520-0477(1982)063<1309:SCOTEO>2.0.CO;2, 1982.

Willmott, C. J., Ackleson, S. G., Davis, R. E., Feddema, J. J., Klink, K. M., Legates, D. R., ... and Rowe, C. M.: Statistics for the evaluation of model performance. J. Geophys. Res, 90, 8995-9005, 1985.

Willmott, C. J., Robeson, S. M., and Matsuura, K.: A refined index of model performance, International Journal of Climatology, 32, 2088-2094, https://doi.org/10.1002/joc.2419, 2012.

Wu, Q., Wang, Z., Chen, H., Zhou, W., and Wenig, M.: An evaluation of air quality modeling over the Pearl River Delta during November 2006, Meteorology and Atmospheric Physics, 116, 113-132, https://doi.org/10.1007/s00703-011-0179-z, 2012.

Ye, C., Liu, P., Ma, Z., Xue, C., Zhang, C., Zhang, Y., Liu, J., Liu, C., Sun, X., and Mu, Y.: High H_2O_2 concentrations observed during haze periods during the winter in Beijing: importance of H_2O_2 oxidation in sulfate formation, Environmental Science & Technology Letters, 5, 757-763, https://doi.org/10.1021/acs.estlett.8b00579, 2018.

Ying, Q., Feng, M., Song, D., Wu, L., Hu, J., Zhang, H., Kleeman, M. J., and Li, X.: Improve regional distribution and source apportionment of $PM_{2.5}$ trace elements in China using inventory-observation constrained emission factors, Science of the Total Environment, 624, 355-365, https://doi.org/10.1016/j.scitotenv.2017.12.138, 2018.

Zhang, H., Cheng, S., Wang, X., Yao, S., and Zhu, F.: Continuous monitoring, compositions analysis and the implication of regional transport for submicron and fine aerosols in Beijing, China, Atmospheric Environment, 195, 30-45, https://doi.org/10.1016/j.atmosenv.2018.09.043, 2018.

Zhang, Q., Streets, D. G., Carmichael, G. R., He, K. B., Huo, H., Kannari, A., ... and Chen, D.: Asian emissions in 2006 for the NASA INTEX-B mission, Atmospheric Chemistry and Physics, 9, 5131-5153, https://doi.org/10.5194/acpd-9-4081-2009, 2009.

Zhang, Y., Zhang, X., Wang, L., Zhang, Q., Duan, F., and He, K.: Application of WRF/Chem over East Asia: Part I. Model evaluation and intercomparison with MM5/CMAQ, Atmospheric Environment, 124, 285-300, https://doi.org/10.1016/j.atmosenv.2015.07.022, 2016.

Zhang, Z., Wang, W., Cheng, M., Liu, S., Xu, J., He, Y., and Meng, F.: The contribution of residential coal combustion to $PM_{2.5}$ pollution over China's Beijing-Tianjin-Hebei region in winter, Atmospheric Environment, 159, 147-161, https://doi.org/10.1016/j.atmosenv.2017.03.054, 2017.

Zhao, B., Wang, S., Donahue, N. M., Jathar, S. H., Huang, X., Wu, W., Hao, J., and Robinson, A. L.: Quantifying the effect of organic aerosol aging and intermediate-volatility emissions on regional-scale aerosol pollution in China, Scientific Reports, 6, 1-10, https://doi.org/10.1038/srep28815, 2016.

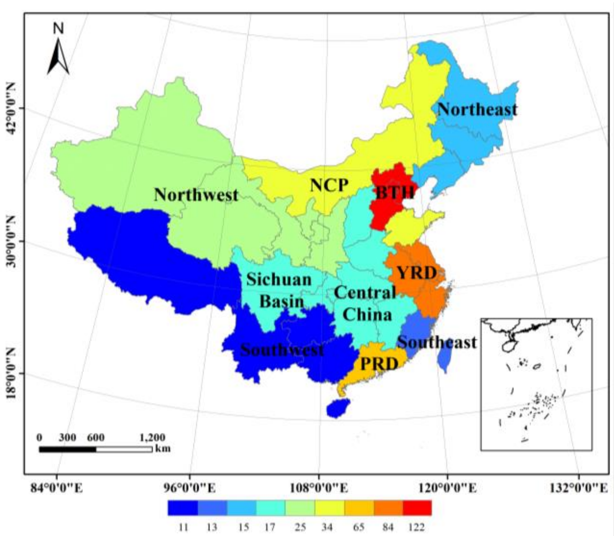

Figure 1: Map of regions defined in this study (see Table S2 for provinces covered by each region). Colour bar indicates the number of studies evaluating the region (studies covering entire China were excluded from counting)

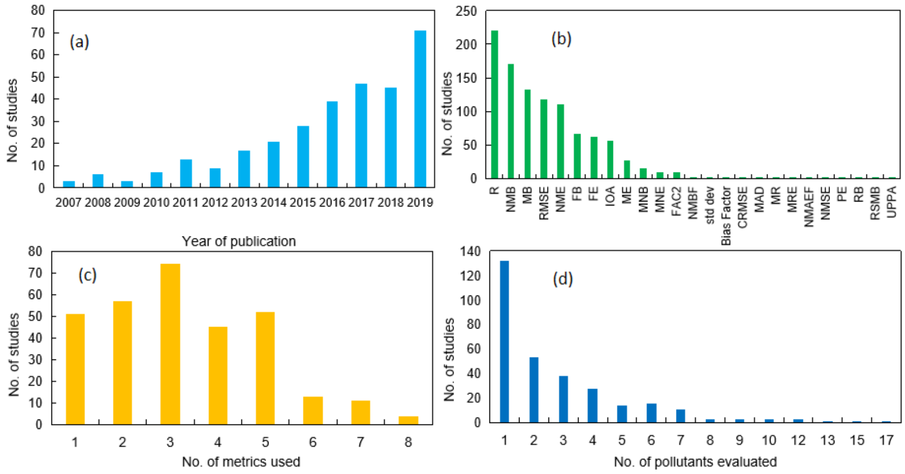

Figure 2: (a) number of studies published during 2006-2019; (b) frequency of use of each metrics; (c) number of metrics used in studies; (d) frequency of number of pollutants evaluated.

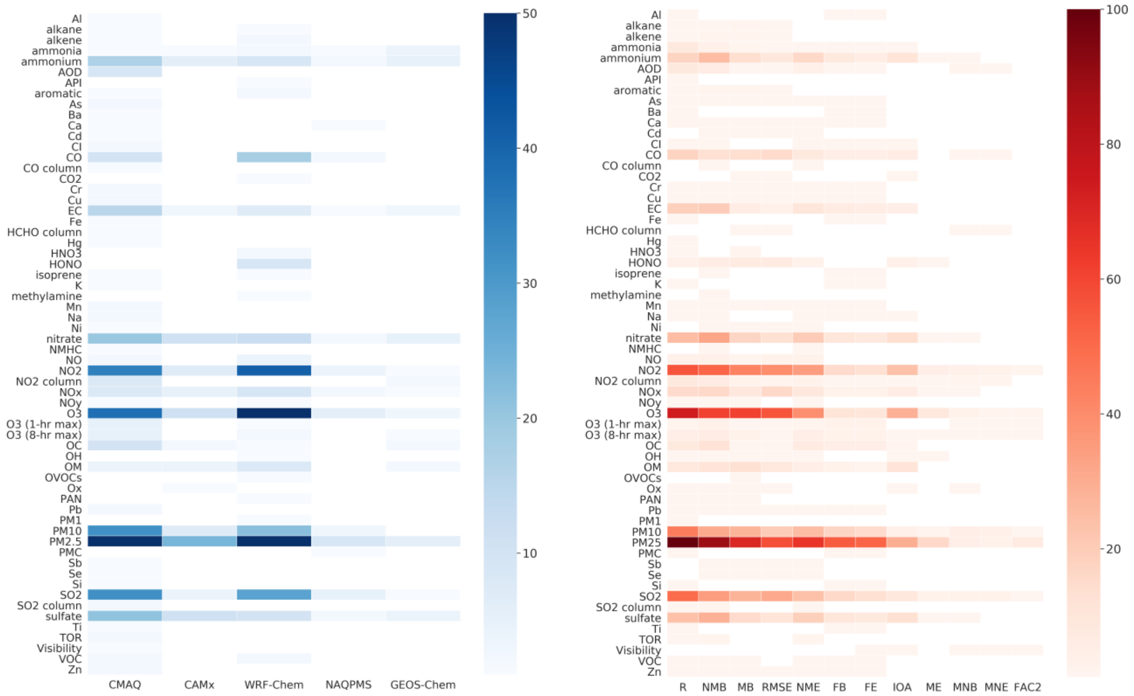

Figure 3: Number of studies evaluating each pair of a pollutant and PGM models (left); number of studies evaluating each pair of a pollutant and statistical metric (right). See Table S5 for species abbreviations.

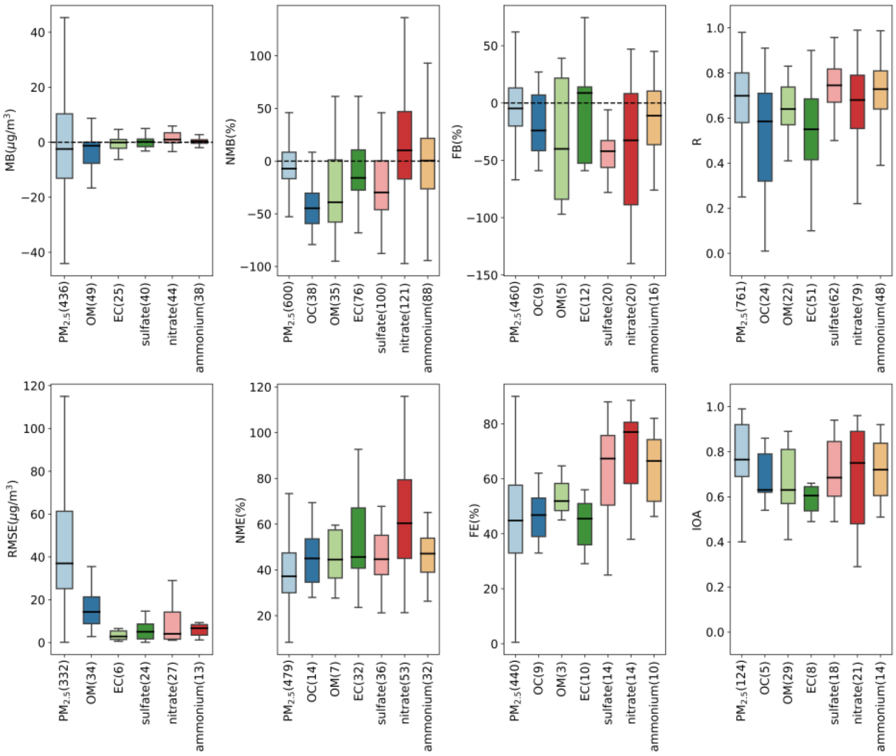

Figure 4: Quantile distribution of selected PM performance metrics compiled in this work. Median values are shown as centerlines; the upper and lower bound of boxes correspond to the 25th and 75th percentile values; whiskers extend to 1.5 times the interquartile range (outliers are excluded). The numbers in brackets indicate the number of data points available.

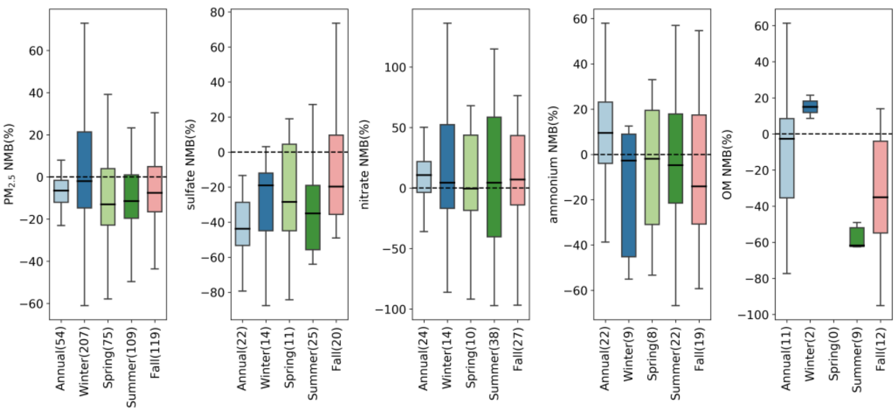

Figure 5: NMB of total $PM_{2.5}$ and speciated components split by season.

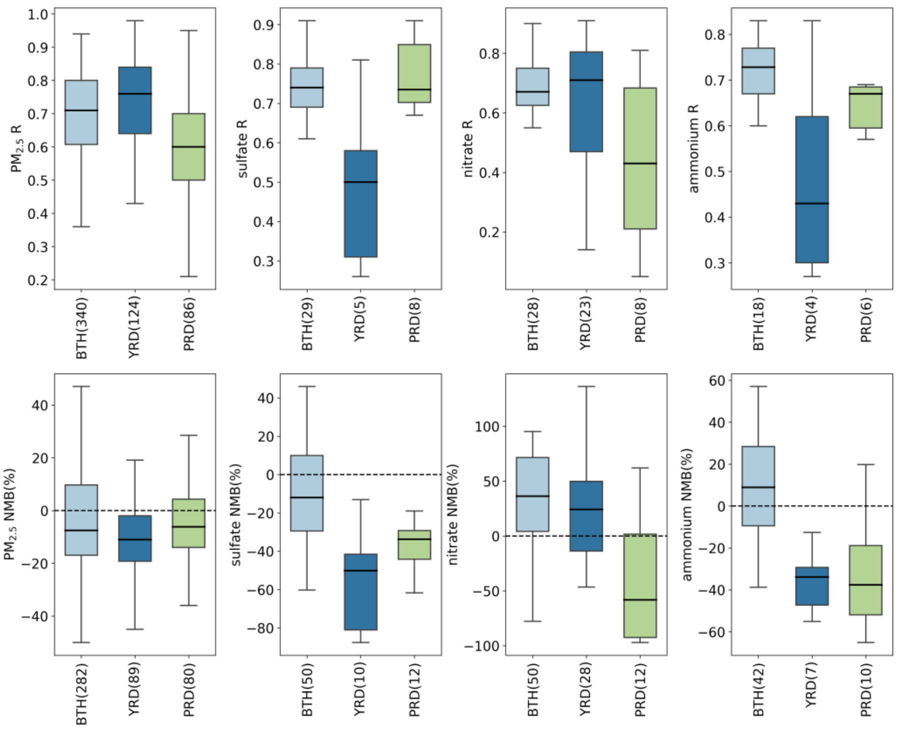

Figure 6: Quantile distribution of R and NMB of total $PM_{2.5}$ and speciated species in BTH, YRD and PRD

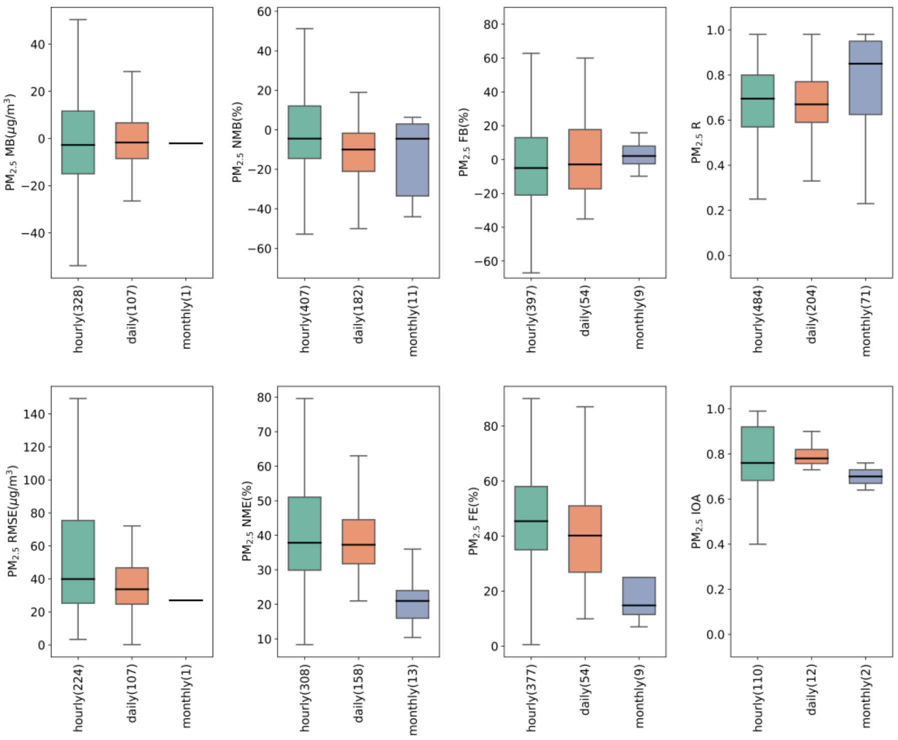

Figure 7: Quantile distributions of MB, RMSE, NMB, NME, FB, FE, R and IOA of total $PM_{2.5}$ presented by temporal resolution for model validation

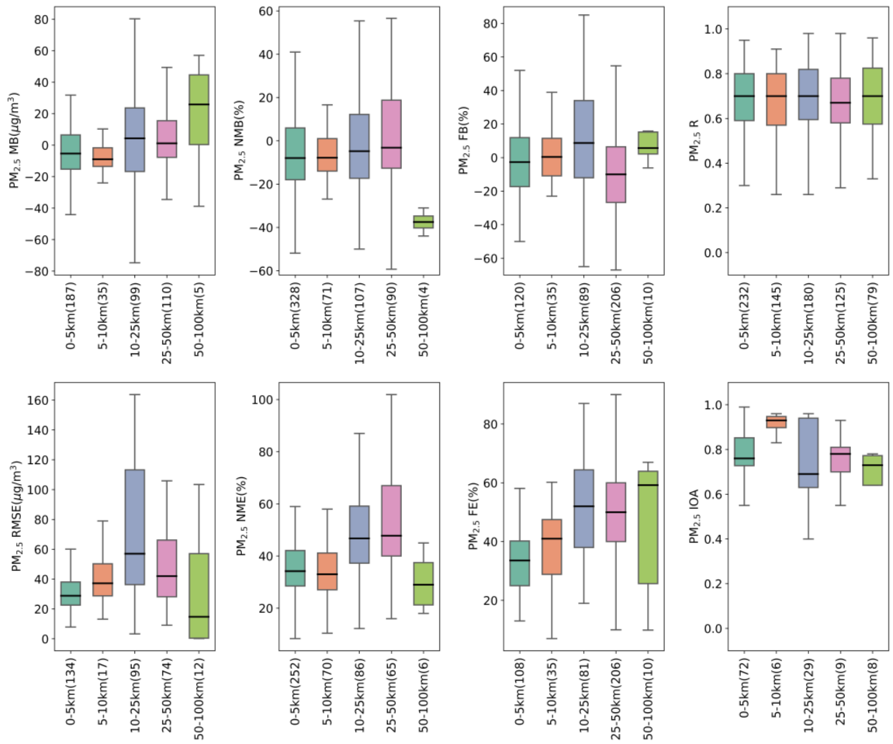

Figure 8: Quantile distributions of MB, RMSE, NMB, NME, FB, FE, R and IOA of total $PM_{2.5}$ presented by model grid resolution

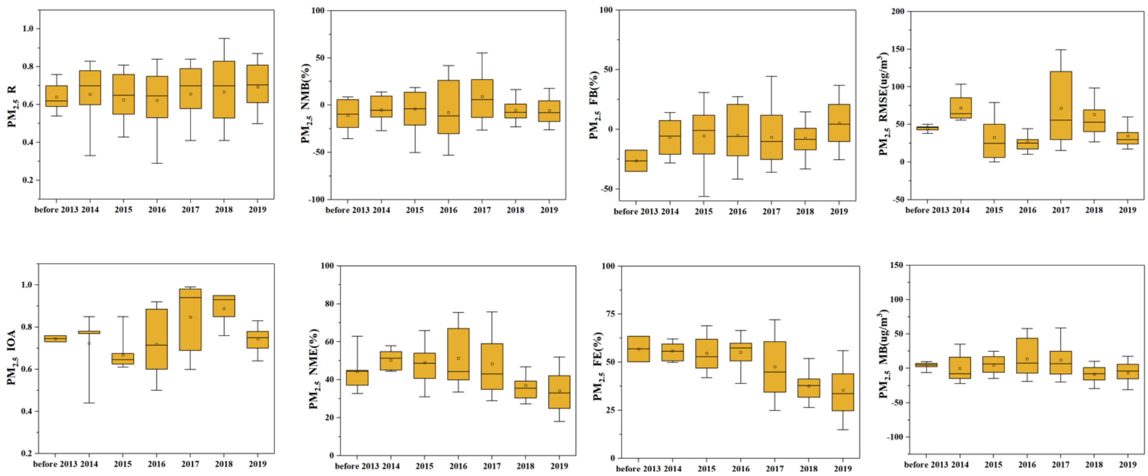

Figure 9: Quantile distribution of R, IOA, NMB, NME, FB, FE, MB and RMSE of total $PM_{2.5}$ presented by data published year

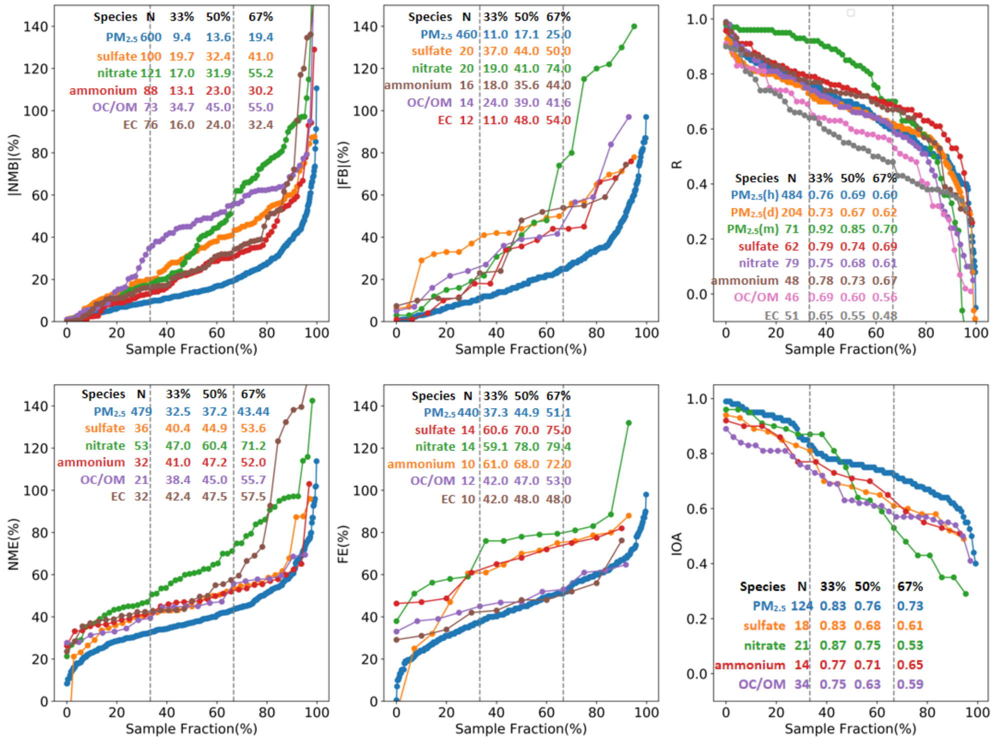

Figure 10: Rank-ordered distributions of NMB, NME, FB, FE, R and IOA for total $PM_{2.5}$ and speciated components. The number of data points and the 33rd, 50th, and 67th percentile values are also listed. For instance, one third of reported R value for predicted hourly $PM_{2.5}$ concentration is higher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

Table 1 Definition of statistical metrics used in more than ten studies complied in this work

No.	Statistics (abbreviation)	Definition	Note
1	Correlation coefficient (R)	$\frac{\sum[(P_j-\bar{P})\times(O_j-\bar{O})]}{\sqrt{\sum(P_j-\bar{P})^2\times\sum(O_j-\bar{O})^2}}$	Unitless, -1≤R≤1
2	Index of agreement (*d*)	$1-\frac{\sum(P_j-O_j)^2}{\sum(\lvert P_j-\bar{O}\rvert+\lvert O_j-\bar{O}\rvert)^2}$	Unitless, 0≤*d*≤1
3	Normalize mean bias (NMB)	$\frac{\sum(P_j-O_j)}{\sum O_j}\times 100$	-100%≤NMB≤+∞
4	Normalize mean error (NME)	$\frac{\sum\lvert P_j-O_j\rvert}{\sum O_j}\times 100$	0%≤NME≤+∞
5	Fractional bias (FB)	$\frac{2}{N}\frac{\sum(P_j-O_j)}{(P_j+O_j)}\times 100$	-200%≤FB≤+200%
6	Fractional error (FE)	$\frac{2}{N}\frac{\sum\lvert P_j-O_j\rvert}{(P_j+O_j)}\times 100$	0%≤FE≤+200%
7	Root mean square error (RMSE)	$\sqrt{\frac{\sum(P_j-O_j)^2}{N}}$	concentration unit
8	Mean bias (MB)	$\frac{\sum(\mathrm{P_j}-\mathrm{O_j})}{\mathrm{N}}$	concentration unit

9	Mean error (ME)	$\frac{\sum \lvert P_j - O_j \rvert}{N}$	concentration unit

Table 2: Recommended benchmarks for evaluating PGM 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
NMB	Goal	<±10%	<±20%	none	<±20%	<±40%	<±20%
	Criteria	<±20%*	<±45%	none	<±35%	<±60%	<±35%*
NME	Goal	<35%	<45%	<50%*	<45%	<40%*	<45%*
	Criteria	<45%*	<55%	<75%*	<55%	<60%*	<60%*
FB	Goal	<±15%	<±40%	none	<±20%	<±30%	none
	Criteria	<±30%	<±55%	none	<±50%	<±50%	None
FE	Goal	<40%	<65%	<65%	<65%	<45%	<45%
	Criteria	<55%	<80%	<80%	<80%	<55%	<55%
R	Goal	>0.70 (hourly/daily) >0.85 (monthly)	>0.70	>0.70	>0.70	>0.60	>0.60
	Criteria	>0.55* (hourly/daily) >0.85 (monthly)	>0.60*	>0.55	>0.60*	>0.50	>0.40
IOA	Goal	>0.80	>0.80	>0.80	>0.75	>0.70	None
	Criteria	>0.70	>0.55	>0.50	>0.60	>0.55	None

[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.

Table 3: List of different formulas for index of agreement

Formula	Range	Reference
$d = 1 - \frac{\sum (P_j - O_j)^2}{\sum (\lvert P_j - \bar{O} \rvert + \lvert O_j - \bar{O} \rvert)^2}$	[0,1]	Willmott (1981)
$d_1 = 1 - \frac{\sum \lvert P_j - O_j \rvert}{\sum (\lvert P_j - \bar{O} \rvert + \lvert O_j - \bar{O} \rvert)}$	[0,1]	Willmott (1982)
$d_1' = 1 - \frac{\sum \lvert P_j - O_j \rvert}{2 \sum \lvert O_j - \bar{O} \rvert)}$	(-∞,1)	Willmott et al. (1985)
$d_r = \begin{cases} 1 - \frac{\sum \lvert P_j - O_j \rvert}{2 \sum \lvert O_j - \bar{O} \rvert}, when \sum \lvert P_j - O_j \rvert \le 2 \sum \lvert O_j - \bar{O} \rvert \\ \frac{2 \sum \lvert O_j - \bar{O} \rvert}{2 \sum \lvert P_j - O_j \rvert} - 1, when \sum \lvert P_j - O_j \rvert > 2 \sum \lvert O_j - \bar{O} \rvert \end{cases}$	[0,1]	Willmott et al. (2012)