https://doi.org/10.5194/acp-2020-237
Preprint. Discussion started: 9 July 2020

Atmospheric Chemistry and Physics
Discussions

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

Ling Huang[1], Yangjun Wang[1], Hehe Zhai[1], Shuhui Xue[1], Tianyi Zhu[1], Yun Shao[1], Ziyi Liu[1], Chris Emery[2], Joshua Fu[3], Kun Zhang[1], Greg Yarwood[2], 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 compliance associated with deteriorated air quality in China for the past decade. Solid evaluation of model performances guarantees the robustness and reliability of the baseline modelling results, so subsequent applications are built on top of it; thus, model performance evaluation (MPE) is a critical step of any PGM applications. MPE procedures and associated benchmarks have been proposed for PGM applications in the United States and Europe. However, with numerous input data needed, diverse model configurations, and evolution of the model itself, no two PGM applications are exactly the same. Therefore, those MPE benchmarks proposed based on studies outside China may not be appropriate for evaluation of the increasing number of PGM applications in China. Here we follow an established approach as published in previous literatures, to recommend statistical benchmarks for evaluation of simulated particulate matter (PM) concentrations in China. A total of 128 peer-reviewed articles published between 2006 and mid-2019 that applied one of four 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. Influences of different model configurations, including modelling regions and seasons, spatial resolution of modelling grids, temporal resolution of MPE, etc., on the range of reported statistics are discussed. Benchmarks for four frequently used evaluation metrics are provided for 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 (i.e. two thirds) of studies can meet. Our proposed benchmarks are further compared with those developed for United States and Europe. Additional recommendations for MPE practices are also given. Results from this study shall help the ever-growing modelling community in China to have a better objective assessment of how well their simulation results are compared with previous studies and to better demonstrate the credibility and robustness of their PGM applications prior to subsequent regulatory assessments.

1 Introduction

Along with the rapid economic development and fast urbanization in China for the past several decades, serious air pollution problems have frequently occurred in many regions of China. The infamous 2013 January severe haze pollution in Beijing and surrounding areas with record-breaking hourly concentrations of $PM_{2.5}$ (particular matter with an aerodynamic diameter less than 2.5 μm) has attracted numerous attention (e.g. Tao et al., 2014; Quan et al., 2014; M. Gao et al., 2015; etc.). Tremendous efforts have been spent to mitigate air pollution situations in China, including the "Air Pollution Prevention and Control Action Plan" in 2013 (The State Council of the People's Republic of China, 2013), "Three-year Plan on Defending the Blue Sky" in 2018 (The State Council of the People's Republic of China, 2018), "Action Plan for Comprehensive Control of Air Pollution in Autumn and Winter" (The Ministry of Ecological Environment, 2018a). Annual $PM_{2.5}$

concentrations and the number of heavy haze days have been reduced in many regions across China during the past several years (Q. Zhang et al., 2019; The Ministry of Ecological Environment, 2018b). Among these efforts, photochemical grid models (PGMs) that numerically simulate the spatial and temporal distributions of air pollutants including ozone, particulate matter (PM), air toxics, and their precursors and/or products, is a key component of linking scientific researches with regulatory applications. With its 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, evaluation of air quality impacts on public health and ecosystems, developing effective control strategies towards meeting national air quality standards, and etc.

The use of PGMs is much less constrained in the sense that there are no such "uniform" settings for PGM applications. First and foremost, there exist different photochemical models developed by different groups. To give a few examples, the GEOS-Chem by Harvard University at global scale (http://www.geos-chem.org), the Comprehensive Air Quality Model with Extensions (CAMx) by Ramboll (Ramboll Environment and Health, 2018) and the Community Multiscale Air Quality (CMAQ) model (Foley et al., 2010) by United States (U.S.) Environmental Protection Agency (EPA) at regional scale. On top of that, a PGM application requires various inputs including time-variant meteorology, hourly and gridded emissions inventory, initial/boundary conditions (for example, from global models, or static assumptions), and land use dataset. Model configurations include chemical mechanism, vertical diffusion scheme, planetary boundary layer scheme, numerical solver, dry deposition scheme (e.g. L. Zhang et al. 2003 vs. Wesely 1989), etc. In addition, PGMs are applied with different spatial scales (from urban to regional, super-regional and even global) over different temporal scales (from episodic to monthly, seasonal, yearly or even multi-yearly). All these variations lead to a rich compilation of PGM applications that differ from each other in one way or more.

A critical step of all PGM applications is model performance evaluation (MPE); that is to demonstrate how well modelling results can replicate the observed magnitude as well as the spatial and temporal variations of the target pollutant. Comprehensive and solid MPE practices ensure the accuracy and reliability of modelling results of a baseline PGM simulation and therefore the subsequent applications that are built on top of it. In U.S., four tiers of MPE were proposed as regulatory modelling guidance (EPA, 2014; see full description by Dennis et al. 2010): (1) **operational evaluation**, in which quantitative, statistical and graphical comparisons are performed based on paired modelled and observed data; (2) **dynamic evaluation**, in which "*the accuracy of the model in characterizing the sensitivity of ozone and/or $PM_{2.5}$ to changes in emissions*" is analysed; (3) **diagnostic evaluation**, in which individual physical and chemical process of the model system is evaluated based on process-oriented analysis; and (4) **probabilistic evaluation**, in which "*the level of confidence in the model predictions is assessed through techniques such as ensemble model simulations*". In most cases, only the operational evaluation is being applied for MPE and only few applications also conducted dynamic evaluation (e.g., Foley et al., 2015). The first modelling guidance document issued by EPA provided a set of ozone MPE metrics for ozone attainment demonstration (EPA, 1991). Later, Boylan and Russell (2006) introduced 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. They recommended mean fraction error (MFE, <=50% for goal and <=75% for criteria) and mean fraction bias (MFB, within 30% for goal and within 60% for criteria) as the metrics for PM species evaluation. Several years later, Simon et al. (2012) conducted a comprehensive review of operational MPE results reported in peer-reviewed journals published between 2006 and 2012 on PGM applications across North America (mostly U.S.) and presented quantile distribution of most commonly reported MPE statistics. Emery et al. (2017) later expanded the literature compiled by Simon et al. (2012) and developed an updated set of MPE benchmarks for both ozone and PM species following the concept of "goals" and "criteria" proposed by Boylan and Russell (2006). In Europe, the Forum for Air Quality Modelling in Europe (FAIRMODE) model evaluation methodology is developed to support unified model evaluation process of air quality models used by European Union Member States (Janssen et al., 2017).

The approach of FAIRMODE also relies on various statistical indicators and diagrams based on paired modelled and observed data to offer diagnostics of model performance. Many PGM applications in China used these U.S. based benchmarks to demonstrate 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.) and no doubtfully these U.S. oriented studies provide invaluable information. Nevertheless, it should be noted that all these benchmark studies were based on PGM applications mostly for US and may not be suitable for model evaluation of PGM applications in China, given the complex interactions of various model inputs and availability of local dataset (i.e. emission inventory, speciation database). Therefore, a set of statistics and benchmarks that is specifically targeted to evaluate PGM applications in China is urgently needed but is currently missing to our knowledge.

In this study, a comprehensive review of operational model evaluations of criteria air pollutants including gaseous pollutants (e.g. SO_2, NO_2, ozone) and particulate matters (e.g. PM_{10}, total $PM_{2.5}$, and speciated $PM_{2.5}$) based on model evaluations results of PGM applications in China published in peer reviewed journals between 2006 and 2019 (latest journal published on July 22, 2019, Du et al. 2019) was conducted. The ultimate goal of this work is to develop and recommend a set of quantitative and objective MPE benchmarks that are suitable for PGM applications in China so that the modelling community can have an objective assessment of how well their simulation results compared with historical studies and to better demonstrate the credibility and robustness of PGM applications prior to subsequent regulatory assessments. The work done by Simon et al (2012) and Emery et al (2017) provide excellent examples of methodology and thereby was mostly adopted in this study. We divided this whole work into three parts: the first part (i.e. the current one) gives a general overview of air quality modelling studies in China compiled in this study and results for $PM_{2.5}$ and speciated components are presented; results for ozone will be discussed in the second part; results for other criteria pollutants including PM_{10}, SO_2, NO_2, and CO, etc. will be discussed in the last part. Same as Emery et al. (2017), our proposed benchmarks should not be considered as pass/fail tests but "*simple references to the range of recent historical performance for commonly reported statistics*" (Emery et al., 2017). Evaluation of performances of meteorological inputs for PGM application is also critical, especially for applications focused on source attribution; this will be discussed in a separate study as future work.

2 Methodology

2.1 Data compilation

Over 160 peer-reviewed articles that applied regional air quality models in China and published from 2006 to mid-2019 were first compiled in this work. These studies address diverse air quality issues over entire or certain regions of China, including quantifying source contributions during heavy haze episodes, evaluating emission control schemes, accessing impact of air pollution on health effects and crop yields, etc. Four photochemical models - CMAQ, CAMx, 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 covered in this compilation. While the former three 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. Similar to Simon et al. (2012), we excluded studies that did not report any MPE results or only reported MPE results in graphical form, which leads to a final set of 128 articles included in this review (see summary in Table S1). We defined ten regions 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), Northwest, Northeast, Southeast, and Southwest (see Table S2 for provinces covered in this region).

2.2 Metrics evaluated

A total of 20 performance metrics was used in the 128 articles compiled in this study (see Supplemental Table S3 for a complete list of the 20 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 rare 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 128 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 128 articles included in this work, WRF-Chem was the most frequently used PGM (used in 56 studies), followed by CAMx (31 studies), CMAQ (27 studies), and NAQPMS (18 studies). One study evaluated model performances for CAMx, CMAQ, and NAQPMS (Q. Wu et al. 2012). In terms of regions, BTH (56 studies), YRD (35 studies), and PRD (25 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 or the Fifth Generation Penn State/NCAR Mesoscale Model (MM5) (Grell et al., 1994) 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., 2017b), 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 2 shows the formula of metrics that have been used in more than 10 studies. Same as Simon et al. (2012), the top three most frequently used metrics is correlative coefficient (R, 85 studies), normalized mean bias (NMB, 80 studies), and mean bias (MB, 58 studies). Other frequently used (>10 studies) metrics include root mean square error (RMSE, 54 studies), normalized mean error (NME, 50 studies), fraction bias (FB, 32 studies), index of agreement (IOA, 33 studies), fraction error (FE, 29 studies), and mean error (ME, 11 studies). Mean normalized bias (MNB) and mean normalized error (MNE) were only used in six and four studies, respectively, as mentioned in Simon et al. (2012) that these two metrics tends to give more weight to data at low values. About 71% of articles included in this work used at least three statistical metrics for model performance evaluation (Figure 2c); 13% of studies reported numerical values for only one metric; studies included more than five MPE metrics were less than 10%; three studies (X. Li et al., 2015; Kim et al., 2017; X. Li et al., 2018) used eight statistical metrics. In terms of number of pollutants evaluated in each study (Figure 2d), 55 studies (43%) evaluated only one pollutant and 96 studies (75%) evaluated less than or equal to three pollutants; one study (Tie et al., 2013) evaluated 12 pollutants (including several VOCs species).

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, PM_{10} and SO_2, 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 S2). For total $PM_{2.5}$, slightly more studies reported positive MB values and negative NMB values while approximately equivalent number of studies reported both positive and negative FB values. Reported bias for $PM_{2.5}$ ranges from as low as -40 $\mu g/m^3$ to as much as 50 $\mu g/m^3$ (outliers excluded) with median values around 5 $\mu g/m^3$. The bias range for speciated components is much smaller (within 20 $\mu g/m^3$) because the absolute magnitude of speciated components is much smaller. In terms of normalized bias, the range of $PM_{2.5}$ is comparable or smaller than speciated components. Speciated $PM_{2.5}$ tends to be dominantly under-estimated except for elemental carbon (EC), which is directly emitted from sources as opposed to other speciated components that could be both emitted directory from sources (i.e. primary) and formed via chemical reactions of precursors (i.e. secondary). 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 45%. For FE, median values for all PM species (except organic matters (OM)) are within 40~60%.

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 0.60 (R^2=0.36). Minimum IOA value reported for total $PM_{2.5}$ is 0.44 while minimum R value approaches to zero. Six studies (L. Li et al., 2018; Cheng et al., 2013; L. Chen et al., 2017; X. Li et al., 2018; Y. Liu et al., 2017; Shimadera et al., 2014) 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 (30 out of 32 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 elemental carbon (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 availablilty of data points) by season (Figure5). For total $PM_{2.5}$, number of data points reported for fall and winter is significantly higher than those reported for spring and summer as heavy haze episdoes generally occur in fall and winter. More studies reported negative bias of total $PM_{2.5}$ for all four seasons except spring. The underestimation of total $PM_{2.5}$ in summer and fall is accompanied by dominant understimation of PM components in these two seasons (except for ammoinium in summer). Sulfate tends to be overwhelmingly underesmiated 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). Nitrate is heavily underestimated in summer but since nitrate concentrations tend to be low under high temperature, this negative bias does not much affect the total mass of $PM_{2.5}$. In winter, nitrate is equivalently over- and under-estimated but over-estimation could be as much as 60% in terms of NMB. As opposed to sulfate and nitrate, ammonium could be overesimated in summer. However, it should be noted that the large positive NMB values of ammonium (> 20%) in summer reported here are from one single study that was conducted at a national nature reserve in Sichuan basin (Qian et al. 2015), where prettly low ammonium concentration (< 1 $\mu g/m^3$) was observed and shall not be considered as a representative case. OM also tends to be more underestimated, especially in summer and fall. The understimation 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).

Imapct 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. 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. More positive NMB values were reported for total $PM_{2.5}$ in YRD while the opposite trend is observed for BTH and PRD. In terms of NMB, PRD shows better performance results with smaller range of NMB (within ±25%) whereas ranges for the other two regions are within ±45%. For sulfate and ammonium, underestimation is observed for all three regions with most underestimation in YRD. For nitrate, studies in BTH and PRD reported both positive and negative NMB while nitrate in YRD is always underestimated.

Imapct 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, weekly, monthly and even annually-averaged pairs of modelling results and observations were used for model evaluation. Figure 7 shows the quantile distribution of R, RMSE, MB, NMB and NME for $PM_{2.5}$ presented by the temporal resolution used for model validation. Model seems to better capture observed variations when coarser temporal pairs of observations and model results are used, as indicated by higher R values as temporal resolution gets coarser. Hourly and daily results of bias metrics do not show much difference. However, NME significantly improves as temporal resolution gets coarser.

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 128 articles compiled in this study, a total of 20 grid resolutions was used, ranging from as coarse as 81 km to as fine as 1 km depending on the target region and the purpose of the application. While most of the studies only performed model evaluation for one modelling domain (usually the finest domain), four studies (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. Figure 8 shows the distribution of three statistical metrics (R, NMB, and FB) presented by model's horizontal resolution. To remove the impact of temporal resolution, results shown in Figure 8 are only based on hourly data and results with less than five data points were excluded. In terms of R values, finer spatial resolution does not necessarily improve the correlation performance between modelling results and observations. R values at the finest grid resolution (3km) range from as low as 0.12 to as high as 0.95 while at the coarsest resolution (80km) from 0.51 to 0.76. NMB seems to be moving from underestimation to overestimation as grid resolution gets coarser and no clear trend is observed for FB. The range of each statistical metrics seems to be more associated with the number of available points instead of the grid resolution. For example, the wider range of R and NMB at 3 km and 4 km resolution and that of FB at 12 km resolution is more likely due to more data points being available. 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. 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.

3.3 Recommended metrics and benchmarks

We presented similar diagrams as Emery et al. (2017) to develop metrics and benchmarks for model evaluation. Figure 9 shows the rank-ordered distribution of R, IOA, NMB and NME 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 33rd percentile value increases from around 0.5 for hourly and daily to 0.70 for monthly results; the 67th percentile increases from 0.64 to 0.91 as the total $PM_{2.5}$ is evaluated with coarser resolution. Secondary inorganic species (sulfate, nitrate and ammonium) show consistently higher correlation coefficient compared to total $PM_{2.5}$ with relative similar range of 0.65~0.75 to above 0.80 over the 33rd – 67th percentile interval. For OC/OM, the 33rd (0.51) and 67th (0.74) percentile value is similar to that of daily $PM_{2.5}$ while EC shows slightly lower 33rd (0.43) and 67th (0.66) percentile value compared to OC/OM. In terms of IOA, the 33rd – 67th percentile interval ranges from 0.69 to 0.91 for total $PM_{2.5}$, 0.6 to 0.83 for sulfate and nitrate, 0.73 to 0.77 for ammonium and 0.57 to 0.62 for OC/OM. 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 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 16.4% to 51.0% and NME from 46.5% to 63.5% for 33rd to 67th percentile interval). The 33rd to 67th range of NMB for EC (12.0% to 39.0%) is much lower than that for OC/OM (34.7% to 59.6%) while NME for OC/OM and EC is similar, ranging from ~43% to 58%.

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 10 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 one of our compiled studies used IOA; after 2013, 32 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, R and NME results for total $PM_{2.5}$ get better as temporal resolution gets coarser while no clear trend is observed for NMB. Therefore, different benchmarks are developed for R and NME. 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 ten data points were used to develop the benchmarks of R for ammonium and OC/OM and IOA for ammonium. For sulfate and nitrate, although the numbers of R data points are slightly fewer than that in Emery et al. (2017), 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 NMB and NME, we do observe sharp changes in rank-order values, for example, the NMB for nitrate and EC, and NME for EC. Therefore, we do not give benchmarks in this situation.

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 33th (or 67th) 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 daily $PM_{2.5}$ is 25%(45%) as opposed to 30%(50%) in Emery's study; "criteria" value for R benchmark is also higher (0.45) 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

https://doi.org/10.5194/acp-2020-237
Preprint. Discussion started: 9 July 2020

included in Emery et al. (2017) was published in 2015. However, our "goal" values for NMB and R benchmarks are less strict than that proposed by Emery et al. (2017). For speciated components, NMB and NME benchmarks for nitrate and EC are lower (i.e. stricter) than Emery's study while the opposite is true for sulfate, ammonium. However, it should be noted that the numbers of data points for NMB and NME results in our study are significantly lower than that used in Emery's study, thus a direct comparison would be inappropriate. For correlation coefficient, we were only able to make a direct comparison for sulfate because of data availability and our R benchmarks for sulfate are much higher (i.e. more strict) than those in Emery's study.

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 bewteen 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 uncertaities. Therefore, it is not diretly comparable between MPC and the benchamrks 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., 2017). 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. Over a quarter (33 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 1 study vs. 32 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 128 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 four commonly used metrics – NMB, NME, R and IOA based on the method employed by Emery et al. (2017). For total $PM_{2.5}$, we provided 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; H. Z., S. X, T. Z., and Y. S. complied studies and collected data with equal contributions; L.H. 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. 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), the Shanghai International Science and Technology Cooperation Fund (NO. 19230742500), and Chinese National Key Technology R&D Program (NO. 2014BAC22B03 and NO. 2018YFC0213800).

References

Bouarar, I., Brasseur, G., Petersen, K., Granier, C., Fan, Q., Wang, X., ... & Gao, W. (2019). Influence of anthropogenic emission inventories on simulations of air quality in China during winter and summer 2010. Atmospheric Environment, 198, 236-256.

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.

Chen, D., Liu, Z., Fast, J., & Ban, J. (2016). Simulations of sulfate-nitrate-ammonium (SNA) aerosols during the extreme haze events over northern China in October 2014. Atmospheric Chemistry & Physics, 16(16).

Chen, D., Liu, X., Lang, J., Zhou, Y., Wei, L., Wang, X., & Guo, X. (2017). 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.

Chen, D., Zhao, N., Lang, J., Zhou, Y., Wang, X., Li, Y., ... & Guo, X. (2018). Contribution of ship emissions to the concentration of $PM_{2.5}$: A comprehensive study using AIS data and WRF/Chem model in Bohai Rim Region, China. Science of the Total Environment, 610, 1476-1486.

Chen, D., Tian, X., Lang, J., Zhou, Y., Li, Y., Guo, X., ... & Liu, B. (2019). The impact of ship emissions on $PM_{2.5}$ and the deposition of nitrogen and sulfur in Yangtze River Delta, China. Science of the Total Environment, 649, 1609-1619.

Chen, H., Li, J., Ge, B., Yang, W., Wang, Z., Huang, S., ... & Zhu, L. (2015). Modeling study of source contributions and emergency control effects during a severe haze episode over the Beijing-Tianjin-Hebei area. Science China Chemistry, 58(9), 1403-1415.

Chen, L., Zhao, H., Han, B., & Bai, Z. (2014). Combined use of WEPS and Models-3/CMAQ for simulating wind erosion source emission and its environmental impact. Science of the Total Environment, 466, 762-769.

Chen, L., Zhang, M., Zhu, J., & Skorokhod, A. (2017). Model analysis of soil dust impacts on the boundary layer meteorology and air quality over East Asia in April 2015. Atmospheric Research, 187, 42-56.

Chen, Q., Fu, T. M., Hu, J., Ying, Q., & Zhang, L. (2017). Modelling secondary organic aerosols in China. National Science Review, 4(6), 806-809.

Chen, X., Situ, S., Zhang, Q., Wang, X., Sha, C., Zhouc, L., ... & Li, C. (2019). The synergetic control of NO_2 and O_3 concentrations in a manufacturing city of southern China. Atmospheric Environment, 201, 402-416.

Cheng, S., Wang, F., Li, J., Chen, D., Li, M., Zhou, Y., & Ren, Z. (2013). Application of trajectory clustering and source apportionment methods for investigating trans-boundary atmospheric PM_{10} pollution. Aerosol Air Qual. Res, 13, 333-342.

Cheng, Z., Wang, S., Fu, X., Watson, J. G., Jiang, J., Fu, Q., ... & Hao, J. (2014). Impact of biomass burning on haze pollution in the Yangtze River delta, China: a case study in summer 2011. Atmos. Chem. Phys, 14(9), 4573-4585.

Du, H., Li, J., Chen, X., Wang, Z., Sun, Y., Fu, P., ... & Wei, Y. (2019). Modeling of aerosol property evolution during winter haze episodes over a megacity cluster in northern China: roles of regional transport and heterogeneous reactions of SO_2. Atmospheric Chemistry and Physics, 19(14), 9351-9370.

Emery, C., Liu, Z., Russell, A. G., Odman, M. T., Yarwood, G., & Kumar, N. (2017). Recommendations on statistics and benchmarks to assess photochemical model performance. Journal of the Air & Waste Management Association, 67(5), 582-598.

Feng, R., Wang, Q., Huang, C. C., Liang, J., Luo, K., Fan, J. R., & Cen, K. F. (2019). Investigation on air pollution control strategy in Hangzhou for post-G20/pre-Asian-games period (2018–2020). Atmospheric Pollution Research, 10(1), 197-208.

Feng, T., Bei, N., Huang, R. J., Cao, J., Zhang, Q., Zhou, W., ... & Lei, W. (2016a). Summertime ozone formation in Xi'an and surrounding areas, China. Atmospheric Chemistry & Physics, 16(7).

Feng, T., Li, G., Cao, J., Bei, N., Shen, Z., Zhou, W., ... & Tie, X. (2016b). Simulations of organic aerosol concentrations during springtime in the Guanzhong Basin, China. Atmospheric Chemistry & Physics, 16(15).

Feng, T., Bei, N., Zhao, S., Wu, J., Li, X., Zhang, T., ... & Li, G. (2018a). Wintertime nitrate formation during haze days in the Guanzhong basin, China: A case study. Environmental Pollution, 243, 1057-1067.

Feng, T., Zhou, W., Wu, S., Niu, Z., Cheng, P., Xiong, X., & Li, G. (2018b). Simulations of summertime fossil fuel CO2 in the Guanzhong basin, China. Science of the Total Environment, 624, 1163-1170.

Feng, X., Fu, T. M., Cao, H., Tian, H., Fan, Q., & Chen, X. (2019). Neural network predictions of pollutant emissions from open burning of crop residues: Application to air quality forecasts in southern China. Atmospheric Environment, 204, 22-31.

Foley, K. M., Hogrefe, C., Pouliot, G., Possiel, N., Roselle, S. J., Simon, H., & Timin, B. (2015). Dynamic evaluation of CMAQ part I: Separating the effects of changing emissions and changing meteorology on ozone levels between 2002 and 2005 in the eastern US. Atmospheric Environment, 103, 247-255.

Foley, K. M., Roselle, S. J., Appel, K. W., Bhave, P. V., Pleim, J. E., Otte, T. L., ... & Nolte, C. G. (2010). Incremental testing of the Community Multiscale Air Quality (CMAQ) modeling system version 4.7. Geoscientific Model Development, 3(1), 205.

Fu, X., Wang, T., Zhang, L., Li, Q., Wang, Z., Xia, M., ... & Zhou, Y. (2019). The significant contribution of HONO to secondary pollutants during a severe winter pollution event in southern China. Atmospheric Chemistry & Physics, 19(1).

Gao, J., Zhu, B., Xiao, H., Kang, H., Hou, X., & Shao, P. (2016). A case study of surface ozone source apportionment during a high concentration episode, under frequent shifting wind conditions over the Yangtze River Delta, China. Science of the Total Environment, 544, 853-863.

Gao, J., Zhu, B., Xiao, H., Kang, H., Hou, X., Yin, Y., ... & Miao, Q. (2017). Diurnal variations and source apportionment of ozone at the summit of Mount Huang, a rural site in Eastern China. Environmental Pollution, 222, 513-522.

Gao, M., Guttikunda, S. K., Carmichael, G. R., Wang, Y., Liu, Z., Stanier, C. O., ... & Yu, M. (2015). Health impacts and economic losses assessment of the 2013 severe haze event in Beijing area. Science of the Total Environment, 511, 553-561.

Gao, M., Carmichael, G. R., Saide, P. E., Lu, Z., Yu, M., Streets, D. G., & Wang, Z. (2016a). Response of winter fine particulate matter concentrations to emission and meteorology changes in North China. Atmospheric Chemistry and Physics, 16(18), 11837.

Gao, M., Carmichael, G. R., Wang, Y., Saide, P. E., Yu, M., Xin, J., ... & Wang, Z. (2016b). Modeling study of the 2010 regional haze event in the North China Plain. Atmospheric Chemistry and Physics, 16(3), 1673.

Gao, M., Ji, D., Liang, F., & Liu, Y. (2018). Attribution of aerosol direct radiative forcing in China and India to emitting sectors. Atmospheric Environment, 190, 35-42.

Ge, B. Z., Wang, Z. F., Xu, X. B., Wu, J. B., Yu, X. L., & Li, J. (2014). 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.

Grell, G. A., Dudhia, J., & Stauffer, D. R. (1994). A description of the fifth-generation Penn State/NCAR mesoscale model (MM5).

Grell, G. A., Peckham, S. E., Schmitz, R., McKeen, S. A., Frost, G., Skamarock, W. C., & Eder, B. (2005). Fully coupled “online” chemistry within the WRF model. Atmospheric Environment, 39(37), 6957-6975.

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P. I., & Geron, C. (2006). Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature). Atmospheric Chemistry & Physics, 6, 3181-3210.

Guo, J., He, J., Liu, H., Miao, Y., Liu, H., & Zhai, P. (2016). Impact of various emission control schemes on air quality using WRF-Chem during APEC China 2014. Atmospheric Environment, 140, 311-319.

Han, X., Zhu, L., Wang, S., Meng, X., Zhang, M., & Hu, J. (2018). Modeling study of impacts on surface ozone of regional transport and emissions reductions over North China Plain in summer 2015. Atmospheric Chemistry and Physics, 18(16), 12207-12221.

Hu, J., Chen, J., Qi, Y., & Zhang, H. (2016). One-year simulation of ozone and particulate matter in China using WRF/CMAQ modeling system. Atmospheric Chemistry and Physics, 16(16), 10333.

Hu, J., Li, X., Huang, L., Qi, Y., Zhang, Q., Zhao, B., ... & Zhang, H. (2017). Ensemble prediction of air quality using the WRF/CMAQ model system for health effect studies in China. Atmospheric Chemistry and Physics, 17(21), 13103.

Hu, J., Li, Y., Zhao, T., Liu, J., Hu, X. M., Liu, D., ... & Chang, L. (2018). An important mechanism of regional O_3 transport for summer smog over the Yangtze River Delta in eastern China. Atmospheric Chemistry and Physics, 18(22), 16239-16251.

Hu, Y., Wang, S., Yang, X., Kang, Y., Ning, G., & Du, H. (2019). Impact of winter droughts on air pollution over Southwest China. Science of the Total Environment, 664, 724-736.

Huang, L., An, J., Koo, B., Yarwood, G., Yan, R., Wang, Y., ... & Li, L. (2019). 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(22), 14311-14328.

Huang, X., Zhou, L., Ding, A., Qi, X., Nie, W., Wang, M., ... & Rusanen, A. (2016). Comprehensive modelling study on observed new particle formation at the SORPES station in Nanjing, China. Atmospheric Chemistry and Physics, 16(4), 2477.

Huang, Z., Ou, J., Zheng, J., Yuan, Z., Yin, S., Chen, D., & Tan, H. (2016). Process Contributions to Secondary Inorganic Aerosols during Typical Pollution Episodes over the Pearl River Delta Region, China. Aerosol and Air Quality Research, 16, 2129-2144.

Janssen, S., Guerreiro, C., Viane, P., Georgieva, E., Thunis, P., Cuvelier, K., ... & Stocker, J. (2017). 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).

Jia, J., Cheng, S., Liu, L., Lang, J., Wang, G., Chen, G., & Liu, X. (2017). An integrated WRF-CAMx Modeling approach for impact analysis of implementing the emergency $PM_{2.5}$ control measures during red alerts in Beijing in December 2015. Aerosol and Air Quality Research, 17, 2491-2508.

Jimenez, J. L., Canagaratna, M. R., Donahue, N. M., Prevot, A. S. H., Zhang, Q., Kroll, J. H., ... & Aiken, A. C. (2009). Evolution of organic aerosols in the atmosphere. Science, 326(5959), 1525-1529.

Kim, B. U., Bae, C., Kim, H. C., Kim, E., & Kim, S. (2017). Spatially and chemically resolved source apportionment analysis: Case study of high particulate matter event. Atmospheric Environment, 162, 55-70.

Kurokawa, J., Ohara, T., Morikawa, T., Hanayama, S., Janssens-Maenhout, G., Fukui, T., ... & Akimoto, H. (2013). Emissions of air pollutants and greenhouse gases over Asian regions during 2000–2008: Regional Emission inventory in ASia (REAS) version 2. Atmos. Chem. Phys, 13(21), 11019-11058.

Li, G., Bei, N., Cao, J., Wu, J., Long, X., Feng, T., ... & Tie, X. (2017). Widespread and persistent ozone pollution in eastern China during the non-winter season of 2015: observations and source attributions. Atmospheric Chemistry & Physics, 17(4).

Li, J., Wang, Z., Akimoto, H., Gao, C., Pochanart, P., & Wang, X. (2007). Modeling study of ozone seasonal cycle in lower troposphere over east Asia. Journal of Geophysical Research: Atmospheres, 112(D22).

Li, J., Wang, Z., Wang, X., Yamaji, K., Takigawa, M., Kanaya, Y., ... & Tanimoto, H. (2011). Impacts of aerosols on summertime tropospheric photolysis frequencies and photochemistry over Central Eastern China. Atmospheric Environment, 45(10), 1817-1829.

Li, J., Wang, Z., Zhuang, G., Luo, G., Sun, Y., & Wang, Q. (2012). Mixing of Asian mineral dust with anthropogenic pollutants over East Asia: a model case study of a super-duststorm in March 2010. Atmospheric Chemistry & Physics, 12(16).

Li, J., Wang, Z., Huang, H., Hu, M., Meng, F., Sun, Y., ... & Wang, Q. (2013). Assessing the effects of trans-boundary aerosol transport between various city clusters on regional haze episodes in spring over East China. Tellus B: Chemical and Physical Meteorology, 65(1), 20052.

Li, J. L., Zhang, M. G., Gao, Y., & Chen, L. (2016). Model analysis of secondary organic aerosol over China with a regional air quality modeling system (RAMS-CMAQ). Atmospheric and Oceanic Science Letters, 9(6), 443-450.

Li, J., Du, H., Wang, Z., Sun, Y., Yang, W., Li, J., ... & Fu, P. (2017). Rapid formation of a severe regional winter haze episode over a mega-city cluster on the North China Plain. Environmental Pollution, 223, 605-615.

Li, J., Zhang, M., Tang, G., Wu, F., Alvarado, L. M., Vrekoussis, M., ... & Burrows, J. P. (2018). Investigating missing sources of glyoxal over China using a regional air quality model (RAMS-CMAQ). Journal of Environmental Sciences, 71, 108-118.

Li, L., An, J. Y., Zhou, M., Yan, R. S., Huang, C., Lu, Q., ... & Zhu, S. H. (2015). Source apportionment of fine particles and its chemical components over the Yangtze River Delta, China during a heavy haze pollution episode. Atmospheric Environment, 123, 415-429.

Li, L., An, J. Y., Shi, Y. Y., Zhou, M., Yan, R. S., Huang, C., ... & Wu, J. (2016). Source apportionment of surface ozone in the Yangtze River Delta, China in the summer of 2013. Atmospheric Environment, 144, 194-207.

Li, L., An, J., Zhou, M., Qiao, L., Zhu, S., Yan, R., ... & Tao, S. (2018). An integrated source apportionment methodology and its application over the Yangtze River Delta region, China. Environmental Science & Technology, 52(24), 14216-14227.

Li, L., An, J., Huang, L., Yan, R., Huang, C., & Yarwood, G. (2019). Ozone source apportionment over the Yangtze River Delta region, China: Investigation of regional transport, sectoral contributions and seasonal differences. Atmospheric Environment, 202, 269-280.

Li, M., Song, Y., Mao, Z., Liu, M., & Huang, X. (2016). Impacts of thermal circulations induced by urbanization on ozone formation in the Pearl River Delta region, China. Atmospheric Environment, 127, 382-392.

Li, M., Wang, T., Han, Y., Xie, M., Li, S., Zhuang, B., & Chen, P. (2017a). Modeling of a severe dust event and its impacts on ozone photochemistry over the downstream Nanjing megacity of eastern China. Atmospheric Environment, 160, 107-123.

Li, M., Zhang, Q., Kurokawa, J. I., Woo, J. H., He, K., Lu, Z., ... & Cheng, Y. (2017b). MIX: a mosaic Asian anthropogenic emission inventory under the international collaboration framework of the MICS-Asia and HTAP. Atmospheric Chemistry and Physics (Online), 17(2).

Li, N., He, Q., Tie, X., Cao, J., Liu, S., Wang, Q., ... & Zhang, Q. (2016). Quantifying sources of elemental carbon over the Guanzhong Basin of China: A consistent network of measurements and WRF-Chem modeling. Environmental Pollution, 214, 86-93.

Li, N., He, Q., Greenberg, J., Guenther, A., Li, J., Cao, J., ... & Zhang, Q. (2018a). Impacts of biogenic and anthropogenic emissions on summertime ozone formation in the Guanzhong Basin, China. Atmospheric Chemistry and Physics, 18(10), 7489-7507.

Li, N., Lu, Y., Liao, H., He, Q., Li, J., & Long, X. (2018b). WRF-Chem modeling of particulate matter in the Yangtze River Delta region: Source apportionment and its sensitivity to emission changes. PloS one, 13(12).

Li, Q., Zhang, L., Tham, Y. J., Ahmadov, R., Xue, L., Zhang, Q., & Zheng, J. (2016). Impacts of heterogeneous uptake of dinitrogen pentoxide and chlorine activation on ozone and reactive nitrogen partitioning: improvement and application of the WRF-Chem model in southern China. Atmospheric Chemistry and Physics, 16(23), 14875.

Li, X., Zhang, Q., Zhang, Y., Zheng, B., Wang, K., Chen, Y., ... & He, K. (2015). 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.

Li, X., Zhang, Q., Zhang, Y., Zhang, L., Wang, Y., Zhang, Q., ... & Han, W. (2017). Attribution of $PM_{2.5}$ exposure in Beijing–Tianjin–Hebei region to emissions: implication to control strategies. Science Bulletin, 62(13), 957-964.

Li, X., Wu, J., Elser, M., Feng, T., Cao, J., El-Haddad, I., ... & Li, G. (2018). Contributions of residential coal combustion to the air quality in Beijing–Tianjin–Hebei (BTH), China: a case study. Atmospheric Chemistry and Physics, 18(14), 10675-10691.

Li, Y., Lau, A. K., Fung, J. C., Ma, H., & Tse, Y. (2013). Systematic evaluation of ozone control policies using an Ozone Source Apportionment method. Atmospheric Environment, 76, 136-146.

Liao, J., Wang, T., Wang, X., Xie, M., Jiang, Z., Huang, X., & Zhu, J. (2014). Impacts of different urban canopy schemes in WRF/Chem on regional climate and air quality in Yangtze River Delta, China. Atmospheric Research, 145, 226-243.

Liao, J., Wang, T., Jiang, Z., Zhuang, B., Xie, M., Yin, C., ... & Zhang, Y. (2015). WRF/Chem modeling of the impacts of urban expansion on regional climate and air pollutants in Yangtze River Delta, China. Atmospheric Environment, 106, 204-214.

Lin, J., An, J., Qu, Y., Chen, Y., Li, Y., Tang, Y., ... & Xiang, W. (2016). Local and distant source contributions to secondary organic aerosol in the Beijing urban area in summer. Atmospheric Environment, 124, 176-185.

Liu, H., Zhang, M., Han, X., Li, J., & Chen, L. (2019). Episode analysis of regional contributions to tropospheric ozone in Beijing using a regional air quality model. Atmospheric Environment, 199, 299-312.

Liu, S., Hua, S., Wang, K., Qiu, P., Liu, H., Wu, B., ... & Hao, Y. (2018). 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.

Liu, X. H., Zhang, Y., Cheng, S. H., Xing, J., Zhang, Q., Streets, D. G., ... & Hao, J. M. (2010). Understanding of regional air pollution over China using CMAQ, part I performance evaluation and seasonal variation. Atmospheric Environment, 44(20), 2415-2426.

Liu, Y., Hong, Y., Fan, Q., Wang, X., Chan, P., Chen, X., ... & Chen, X. (2017). Source-receptor relationships for $PM_{2.5}$ during typical pollution episodes in the Pearl River Delta city cluster, China. Science of the Total Environment, 596, 194-206.

Liu, Y., Li, L., An, J., Huang, L., Yan, R., Huang, C., ... & Zhang, W. (2018). Estimation of biogenic VOC emissions and its impact on ozone formation over the Yangtze River Delta region, China. Atmospheric Environment, 186, 113-128.

Long, X., Tie, X., Cao, J., Huang, R., Feng, T., Li, N., ... & Zhang, Q. (2016). Impact of crop field burning and mountains on heavy haze in the North China Plain: a case study. Atmospheric Chemistry & Physics, 16(15).

Lu, M., Tang, X., Wang, Z., Gbaguidi, A., Liang, S., Hu, K., ... & Shen, L. (2017). Source tagging modeling study of heavy haze episodes under complex regional transport processes over Wuhan megacity, Central China. Environmental Pollution, 231, 612-621.

Lu, X., & Fung, J. C. (2016a). Source apportionment of sulfate and nitrate over the Pearl River Delta region in China. Atmosphere, 7(8), 98.

Lu, X., Yao, T., Li, Y., Fung, J. C., & Lau, A. K. (2016b). Source apportionment and health effect of NO_x over the Pearl River Delta region in southern China. Environmental Pollution, 212, 135-146.

Lu, X., Chen, Y., Huang, Y., Lin, C., Li, Z., Fung, J. C., & Lau, A. K. (2019). Differences in concentration and source apportionment of $PM_{2.5}$ between 2006 and 2015 over the PRD region in southern China. Science of the Total Environment, 673, 708-718.

Ma, X., Sha, T., Wang, J., Jia, H., & Tian, R. (2018). Investigating impact of emission inventories on $PM_{2.5}$ simulations over North China Plain by WRF-Chem. Atmospheric Environment, 195, 125-140.

Mao, J., Yu, F., Zhang, Y., An, J., Wang, L., Zheng, J., ... & Huang, C. (2018). High-resolution modeling of gaseous methylamines over a polluted region in China: source-dependent emissions and implications of spatial variations. Atmospheric Chemistry and Physics, 18(11), 7933-7950.

Meng, L., Yang, X., Zhao, T., He, Q., Lu, H., Mamtimin, A., ... & Liu, C. (2019). Modeling study on three-dimensional distribution of dust aerosols during a dust storm over the Tarim Basin, Northwest China. Atmospheric Research, 218, 285-295.

The Ministry of Ecological Environment of the People's Republic of China, (2018a), http://www.mee.gov.cn/xxgk2018/xxgk/xxgk03/201811/t20181112_673371.html (accessed February 23, 2020)

The Ministry of Ecological Environment of the People's Republic of China, (2018b), http://www.gov.cn/xinwen/2018-02/01/content_5262720.htm (accessed February 23, 2020)

Peng, W., Yang, J., Wagner, F., & Mauzerall, D. L. (2017). Substantial air quality and climate co-benefits achievable now with sectoral mitigation strategies in China. Science of the Total Environment, 598, 1076-1084.

Peng, Y. P., Chen, K. S., Wang, H. K., Lai, C. H., Lin, M. H., & Lee, C. H. (2011). Applying model simulation and photochemical indicators to evaluate ozone sensitivity in southern Taiwan. Journal of Environmental Sciences, 23(5), 790-797.

Peng, Z., Liu, Z., Chen, D., & Ban, J. (2017). Improving $PM_{2.5}$ forecast over China by the joint adjustment of initial conditions and source emissions with an ensemble Kalman filter. Atmospheric Chemistry and Physics, 17(7), 4837.

Peng, Z., Lei, L., Liu, Z., Sun, J., Ding, A., Ban, J., ... & Chu, K. (2018). The impact of multi-species surface chemical observation assimilation on air quality forecasts in China. Atmospheric Chemistry and Physics, 18(23), 17387-17404.

Pereira, H. R., Meschiatti, M. C., Pires, R. C. D. M., & Blain, G. C. (2018). On the performance of three indices of agreement: an easy-to-use r-code for calculating the Willmott indices. Bragantia, 77(2), 394-403.

Qiao, X., Tang, Y., Hu, J., Zhang, S., Li, J., Kota, S. H., ... & Ying, Q. (2015). 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.

Qiu, Y., Ma, Z., & Li, K. (2019). A modeling study of the peroxyacetyl nitrate (PAN) during a wintertime haze event in Beijing, China. Science of the Total Environment, 650, 1944-1953.

Qu, Y., An, J., Li, J., Chen, Y., Li, Y., Liu, X., & Hu, M. (2014). Effects of NO x and VOCs from five emission sources on summer surface O 3 over the Beijing-Tianjin-Hebei region. Advances in Atmospheric Sciences, 31(4), 787-800.

Quan, J., Tie, X., Zhang, Q., Liu, Q., Li, X., Gao, Y., & Zhao, D. (2014). Characteristics of heavy aerosol pollution during the 2012–2013 winter in Beijing, China. Atmospheric Environment, 88, 83-89.

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

Shimadera, H., Hayami, H., Ohara, T., Morino, Y., Takami, A., & Irei, S. (2014). Numerical simulation of extreme air pollution by fine particulate matter in China in winter 2013. Asian Journal of Atmospheric Environment, 8(1), 25-34.

Shao, J., Chen, Q., Wang, Y., Lu, X., He, P., Sun, Y., ... & Zhao, Y. (2019) 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(9), 6107-6123

Simon, H., Baker, K. R., & Phillips, S. (2012). Compilation and interpretation of photochemical model performance statistics published between 2006 and 2012. Atmospheric Environment, 61, 124-139.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Wang, W., & Powers, J. G. (2005). 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.

Song, S., Gao, M., Xu, W., Sun, Y., Worsnop, D. R., Jayne, J. T., ... & Cheng, C. (2019). Possible heterogeneous chemistry of hydroxymethanesulfonate (HMS) in northern China winter haze. Atmospheric Chemistry and Physics, 19(2), 1357-1371.

The State Council of China. Air Pollution Prevention and Control Action Plan. 2013. http://www.gov.cn/jrzg/201309/12/content_2486918.htm (accessed February 23, 2020)

The State Council of China. Air Pollution Prevention and Control Action Plan. 2018. http://www.mee.gov.cn/ywdt/hjywnews/201807/t20180704_446065.shtml (accessed February 23, 2020)

Sun, X., Cheng, S., Li, J., & Wen, W. (2017). An Integrated Air Quality Model and Optimization Model for Regional Economic and Environmental Development: A Case Study of Tangshan, China. Aerosol and Air Quality Research, 17, 1592-1609.

Tan, J., Zhang, Y., Ma, W., Yu, Q., Wang, Q., Fu, Q., ... & Chen, L. (2017). Evaluation and potential improvements of WRF/CMAQ in simulating multi-levels air pollution in megacity Shanghai, China. Stochastic Environmental Research and Risk Assessment, 31(10), 2513-2526.

Tang, X., Zhu, J., Wang, Z. F., Wang, M., Gbaguidi, A., Li, J., ... & Ji, D. S. (2013). Inversion of CO emissions over Beijing and its surrounding areas with ensemble Kalman filter. Atmospheric Environment, 81, 676-686.

Tao, H., Xing, J., Zhou, H., Chang, X., Li, G., Chen, L., & Li, J. (2018). 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.

Tao, M., Chen, L., Xiong, X., Zhang, M., Ma, P., Tao, J., & Wang, Z. (2014). Formation process of the widespread extreme haze pollution over northern China in January 2013: Implications for regional air quality and climate. Atmospheric Environment, 98, 417-425.

Tie, X., Geng, F., Guenther, A., Cao, J., Greenberg, J., Zhang, R., ... & Cai, C. (2013). Megacity impacts on regional ozone formation: observations and WRF-Chem modeling for the MIRAGE-Shanghai field campaign. Atmospheric Chemistry and Physics, 13(11), 5655-5669.

U.S. EPA. (1991). Guideline for regulatory application of the Urban Airshed Model (No. PB-92-108760/XAB). Environmental Protection Agency, Research Triangle Park, NC (United States).

U.S. EPA. (2014). Draft Modeling Guidance for Demonstrating Attainment of Air Quality Goals for Ozone, $PM_{2.5}$, and Regional Haze. U.S. Environmental Protection Agency, Research Triangle Park, NC (December).

Valbuena, R., Hernando, A., Manzanera, J. A., Görgens, E. B., Almeida, D. R., Silva, C. A., & García-Abril, A. (2019). Evaluating observed versus predicted forest biomass: R-squared, index of agreement or maximal information coefficient?. European Journal of Remote Sensing, 52(1), 345-358.

Wang, D., Jiang, B., Lin, W., & Gu, F. (2019). Effects of aerosol-radiation feedback and topography during an air pollution event over the North China Plain during December 2017. Atmospheric Pollution Research, 10(2), 587-596.

Wang, J., Mo, J., Li, J., Ling, Z., Huang, T., Zhao, Y., ... & Ma, J. (2017). OMI-measured SO2 in a large-scale national energy industrial base and its effect on the capital city of Xinjiang, Northwest China. Atmospheric Environment, 167, 159-169.

Wang, L. T., Wei, Z., Yang, J., Zhang, Y., Zhang, F. F., Su, J., ... & Zhang, Q. (2013). The 2013 severe haze over the southern Hebei, China: model evaluation, source apportionment, and policy implications. Atmospheric Chemistry & Physics Discussions, 13(11).

Wang, L., Wei, Z., Wei, W., Fu, J. S., Meng, C., & Ma, S. (2015). Source apportionment of $PM_{2.5}$ in top polluted cities in Hebei, China using the CMAQ model. Atmospheric Environment, 122, 723-736.

Wang, L., Zhang, Y., Wang, K., Zheng, B., Zhang, Q., & Wei, W. (2016). Application of Weather Research and Forecasting Model with Chemistry (WRF/Chem) over northern China: Sensitivity study, comparative evaluation, and policy implications. Atmospheric Environment, 124, 337-350.

Wang, N., Guo, H., Jiang, F., Ling, Z. H., & Wang, T. (2015). Simulation of ozone formation at different elevations in mountainous area of Hong Kong using WRF-CMAQ model. Science of the Total Environment, 505, 939-951.

Wang, Q., Liu, S., Li, N., Dai, W., Wu, Y., Tian, J., ... & Zhang, R. (2019). Impacts of short-term mitigation measures on $PM_{2.5}$ and radiative effects: a case study at a regional background site near Beijing, China. Atmospheric Chemistry and Physics, 19(3), 1881-1899.

Wang, X., Wei, W., Cheng, S., Li, J., Zhang, H., & Lv, Z. (2018). Characteristics and classification of $PM_{2.5}$ pollution episodes in Beijing from 2013 to 2015. Science of the Total Environment, 612, 170-179.

Wang, X., Wei, W., Cheng, S., Zhang, C., & Duan, W. (2019). A monitoring-modeling approach to SO_4^{2-} and NO_3^- secondary conversion ratio estimation during haze periods in Beijing, China. Journal of Environmental Sciences, 78, 293-302.

Wang, Y., Bao, S., Wang, S., Hu, Y., Shi, X., Wang, J., ... & Russell, A. G. (2017). Local and regional contributions to fine particulate matter in Beijing during heavy haze episodes. Science of the Total Environment, 580, 283-296.

Wang, Z., Li, J., Wang, X., Pochanart, P., & Akimoto, H. (2006). Modeling of regional high ozone episode observed at two mountain sites (Mt. Tai and Huang) in East China. Journal of Atmospheric Chemistry, 55(3), 253-272.

Wang, Z., Zhang, D., Li, X., Li, Y., Chen, T., Liu, B., ... & Pan, L. (2016). Multi-method observation and numerical simulation of a $PM_{2.5}$ pollution episode in Beijing in October, 2014. Aerosol Air Qual. Res, 16, 1403-1415.

Wang, Z., Itahashi, S., Uno, I., Pan, X., Osada, K., Yamamoto, S., ... & Wang, Z. (2017). Modeling the long-range transport of particulate matters for January in East Asia using NAQPMS and CMAQ. Aerosol Air Qual. Res, 17, 3065-3078.

Wang, Z., Pan, X., Uno, I., Chen, X., Yamamoto, S., Zheng, H., ... & Wang, Z. (2018). Importance of mineral dust and anthropogenic pollutants mixing during a long-lasting high PM event over East Asia. Environmental pollution, 234, 368-378.

Wang, Z. F., Xie, F. Y., Wang, X. Q., An, J., & Zhu, J. (2006). Development and application of nested air quality prediction modeling system. Chinese Journal of Atmospheric Sciences-Chinese Edition, 30(5), 778.

Wei, Y., Li, J., Wang, Z. F., Cchen, H. S., Wu, Q. Z., Li, J. J., ... & Wang, W. (2017). Trends of surface $PM_{2.5}$ over Beijing–Tianjin–Hebei in 2013–2015 and their causes: emission controls vs. meteorological conditions. Atmospheric and Oceanic Science Letters, 10(4), 276-283.

Wei, W., Li, Y., Wang, Y., Cheng, S., & Wang, L. (2018a). Characteristics of VOCs during haze and non-haze days in Beijing, China: Concentration, chemical degradation and regional transport impact. Atmospheric Environment, 194, 134-145.

Wei, W., Lv, Z. F., Li, Y., Wang, L. T., Cheng, S., & Liu, H. (2018b). A WRF-Chem model study of the impact of VOCs emission of a huge petro-chemical industrial zone on the summertime ozone in Beijing, China. Atmospheric Environment, 175, 44-53.

Wen, W., Cheng, S., Liu, L., Wang, G., & Wang, X. (2016). Source apportionment of PM 2.5 in Tangshan, China—Hybrid approaches for primary and secondary species apportionment. Frontiers of Environmental Science & Engineering, 10(5), 6.

Wesely, M. L. (1989). Parameterization of surface resistances to gaseous dry deposition in regional-scale numerical models. Atmospheric Environment (1967), 23(6), 1293-1304.

Willmott, C. J. (1981). On the validation of models. Physical Geography, 2(2), 184-194.

Willmott, C. J. (1982). Some comments on the evaluation of model performance. Bulletin of the American Meteorological Society, 63(11), 1309-1313.

Willmott, C. J., Ackleson, S. G., Davis, R. E., Feddema, J. J., Klink, K. M., Legates, D. R., ... & Rowe, C. M. (1985). Statistics for the evaluation of model performance. J. Geophys. Res, 90(C5), 8995-9005.

Willmott, C. J., Robeson, S. M., & Matsuura, K. (2012). A refined index of model performance. International Journal of Climatology, 32(13), 2088-2094.

Wu, Q. Z., Wang, Z. F., Gbaguidi, A., Gao, C., Li, L. N., & Wang, W. (2011). A numerical study of contributions to air pollution in Beijing during CAREBeijing-2006. Atmospheric Chemistry and Physics, 11(12), 5997.

Wu, Q., Wang, Z., Chen, H., Zhou, W., & Wenig, M. (2012). An evaluation of air quality modeling over the Pearl River Delta during November 2006. Meteorology and Atmospheric Physics, 116(3-4), 113-132.

Wu, D., Fung, J. C. H., Yao, T., & Lau, A. K. H. (2013). A study of control policy in the Pearl River Delta region by using the particulate matter source apportionment method. Atmospheric Environment, 76, 147-161.

Wu, J., Li, G., Cao, J., Bei, N., Wang, Y., Feng, T., ... & Tie, X. (2017a). Contributions of trans-boundary transport to summertime air quality in Beijing, China. Atmospheric Chemistry & Physics, 17(3).

Wu, J. B., Wang, Z., Wang, Q., Li, J., Xu, J., Chen, H., ... & Chang, L. (2017b). Development of an on-line source-tagged model for sulfate, nitrate and ammonium: A modeling study for highly polluted periods in Shanghai, China. Environmental Pollution, 221, 168-179.

Wu, J., Bei, N., Li, X., Cao, J., Feng, T., Wang, Y., ... & Li, G. (2018). Widespread air pollutants of the North China Plain during the Asian summer monsoon season: a case study. Atmospheric Chemistry and Physics, 18(12), 8491-8504.

Xie, M., Zhu, K., Wang, T., Feng, W., Gao, D., Li, M., ... & Liao, J. (2016). Changes in regional meteorology induced by anthropogenic heat and their impacts on air quality in South China. Atmos. Chem. Phys, 16(23), 15011-15031.

Xu, J., Chang, L., Qu, Y., Yan, F., Wang, F., & Fu, Q. (2016). The meteorological modulation on PM2. 5 interannual oscillation during 2013 to 2015 in Shanghai, China. Science of the Total Environment, 572, 1138-1149.

Xu, Y., Xue, W., Lei, Y., Zhao, Y., Cheng, S., Ren, Z., & Huang, Q. (2018). Impact of meteorological conditions on PM$_{2.5}$ Pollution in China during winter. Atmosphere, 9(11), 429.

Yang, J., Kang, S., Chen, D., Ji, Z., Tripathee, L., Chen, X., ... & Qiu, G. (2019). Quantifying the contributions of various emission sources to black carbon and assessment of control strategies in western China. Atmospheric Research, 215, 178-192.

Yang, W., Chen, H., Wang, W., Wu, J., Li, J., Wang, Z., ... & Chen, D. (2019). Modeling study of ozone source apportionment over the Pearl River Delta in 2015. Environmental Pollution, 253, 393-402.

Yao, T., Fung, J. C. H., Ma, H., Lau, A. K. H., Chan, P. W., Yu, J. Z., & Xue, J. (2014). Enhancement in secondary particulate matter production due to mountain trapping. Atmospheric Research, 147, 227-236.

Yao, H., Song, Y., Liu, M., Archer-Nicholls, S., Lowe, D., McFiggans, G., ... & Hu, M. (2017). Direct radiative effect of carbonaceous aerosols from crop residue burning during the summer harvest season in East China. Atmospheric Chemistry and Physics, 17(8), 5205.

Ye, C., Liu, P., Ma, Z., Xue, C., Zhang, C., Zhang, Y., ... & Mu, Y. (2018). High H_2O_2 concentrations observed during haze periods during the winter in Beijing: importance of H2O2 oxidation in sulfate formation. Environmental Science & Technology Letters, 5(12), 757-763.

Yin, X., Huang, Z., Zheng, J., Yuan, Z., Zhu, W., Huang, X., & Chen, D. (2017). Source contributions to PM2. 5 in Guangdong province, China by numerical modeling: Results and implications. Atmospheric Research, 186, 63-71.

Zhai, S., An, X., Liu, Z., Sun, Z., & Hou, Q. (2016). Model assessment of atmospheric pollution control schemes for critical emission regions. Atmospheric Environment, 124, 367-377.

Zhang, H., Cheng, S., Wang, X., Yao, S., & Zhu, F. (2018). Continuous monitoring, compositions analysis and the implication of regional transport for submicron and fine aerosols in Beijing, China. Atmospheric Environment, 195, 30-45.

Zhang, J., An, J., Qu, Y., Liu, X., & Chen, Y. (2019). Impacts of potential HONO sources on the concentrations of oxidants and secondary organic aerosols in the Beijing-Tianjin-Hebei region of China. Science of the Total Environment, 647, 836-852.

Zhang, L., Brook, J. R., & Vet, R. (2003). A revised parameterization for gaseous dry deposition in air-quality models. Atmos. Chem. Phys, 3, 2067-2082.

Zhang, L., Wang, T., Lv, M., & Zhang, Q. (2015). On the severe haze in Beijing during January 2013: Unraveling the effects of meteorological anomalies with WRF-Chem. Atmospheric Environment, 104, 11-21.

Zhang, L., Li, Q., Wang, T., Ahmadov, R., Zhang, Q., Li, M., & Lv, M. (2017). Combined impacts of nitrous acid and nitryl chloride on lower-tropospheric ozone: new module development in WRF-Chem and application to China. Atmospheric Chemistry and Physics, 17(16), 9733.

Zhang, L., Zhao, T., Gong, S., Kong, S., Tang, L., Liu, D., ... & Zhang, Y. (2018). Updated emission inventories of power plants in simulating air quality during haze periods over East China. Atmospheric Chemistry & Physics, 18(3).

Zhang, L., Guo, X., Zhao, T., Gong, S., Xu, X., Li, Y., ... & Yin, X. (2019). A modelling study of the terrain effects on haze pollution in the Sichuan Basin. Atmospheric environment, 196, 77-85.

Zhang, Q., Streets, D. G., Carmichael, G. R., He, K. B., Huo, H., Kannari, A., ... & Chen, D. (2009). Asian emissions in 2006 for the NASA INTEX-B mission. Atmospheric Chemistry and Physics, 9(14), 5131-5153.

Zhang, Q., Zheng, Y., Tong, D., Shao, M., Wang, S., Zhang, Y., ... & Ding, Y. (2019). Drivers of improved PM$_{2.5}$ air quality in China from 2013 to 2017. Proceedings of the National Academy of Sciences, 116(49), 24463-24469.

Zhang, Y., Zhang, X., Wang, L., Zhang, Q., Duan, F., & He, K. (2016). Application of WRF/Chem over East Asia: Part I. Model evaluation and intercomparison with MM5/CMAQ. Atmospheric Environment, 124, 285-300.

Zhang, Y., Li, X., Nie, T., Qi, J., Chen, J., & Wu, Q. (2018a). Source apportionment of $PM_{2.5}$ pollution in the central six districts of Beijing, China. Journal of Cleaner Production, 174, 661-669.

Zhang, Y., Shen, J., & Li, Y. (2018b). An atmospheric vulnerability assessment framework for environment management and protection based on CAMx. Journal of Environmental Management, 207, 341-354.

Zhang, Z., Xu, X., Qiao, L., Gong, D., Kim, S. J., Wang, Y., & Mao, R. (2018). Numerical simulations of the effects of regional topography on haze pollution in Beijing. Scientific Reports, 8(1), 1-11.

Zhao, B., Wang, S., Donahue, N. M., Jathar, S. H., Huang, X., Wu, W., ... & Robinson, A. L. (2016). Quantifying the effect of organic aerosol aging and intermediate-volatility emissions on regional-scale aerosol pollution in China. Scientific Reports, 6(1), 1-10.

Zhao, X., Zhao, Y., Chen, D., Li, C., & Zhang, J. (2019). Top-down estimate of black carbon emissions for city clusters using ground observations: a case study in southern Jiangsu, China. Atmospheric Chemistry and Physics, 19(4), 2095-2113.

Zheng, H., Cai, S., Wang, S., Zhao, B., Chang, X., & Hao, J. (2019). Development of a unit-based industrial emission inventory in the Beijing–Tianjin–Hebei region and resulting improvement in air quality modeling. Atmospheric Chemistry and Physics, 19(6), 3447-3462.

Zhou, G., Xu, J., Xie, Y., Chang, L., Gao, W., Gu, Y., & Zhou, J. (2017). Numerical air quality forecasting over eastern China: An operational application of WRF-Chem. Atmospheric Environment, 153, 94-108.

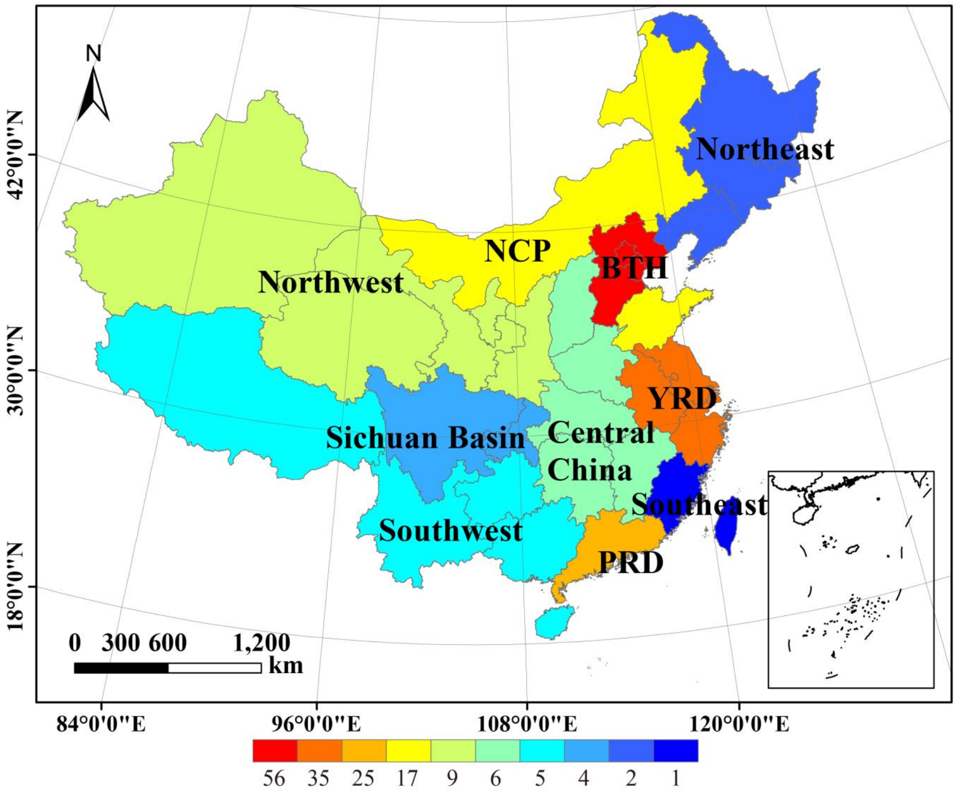


Figure 1: Map of regions defined in this study (see Table S2 for provinces covered by each region). Colour bar indicates the number of studies evaluating the region (studies covering entire China were excluded from counting)

https://doi.org/10.5194/acp-2020-237
Preprint. Discussion started: 9 July 2020

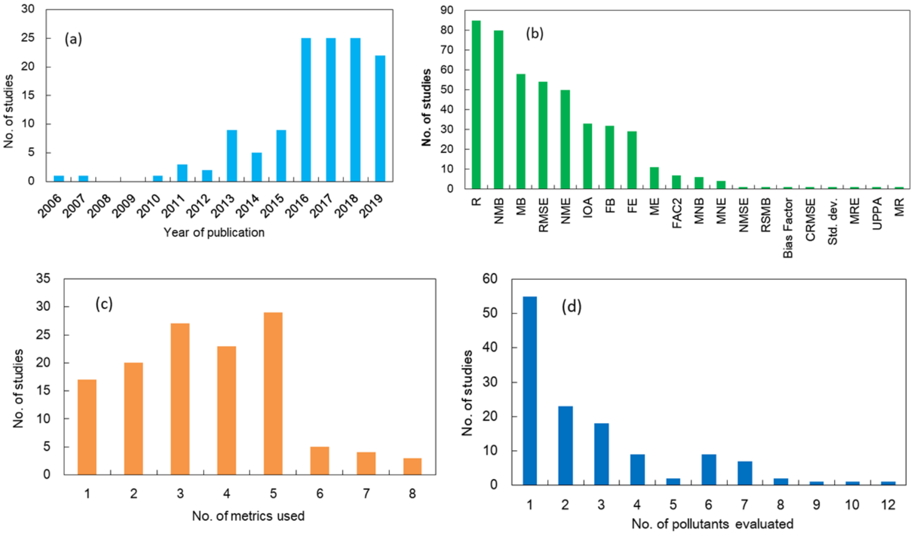


Figure 2: (a) number of studies published during 2006-2019; (b) frequency of use of each metrics; (c) number of metrics used in studies; (d) frequency of number of pollutants evaluated.

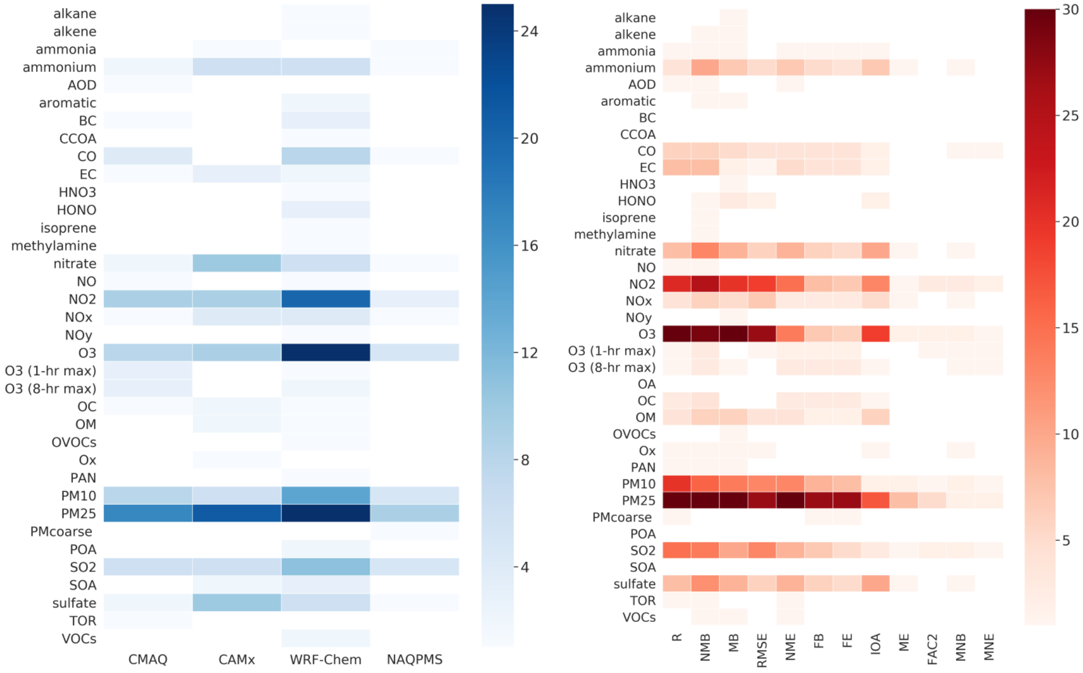


Figure 3: Number of studies evaluating each pair of a pollutant and PGM models (left); number of studies evaluating each pair of a pollutant and statistical metric (right). See Table S4 for species abbreviations.

https://doi.org/10.5194/acp-2020-237
Preprint. Discussion started: 9 July 2020

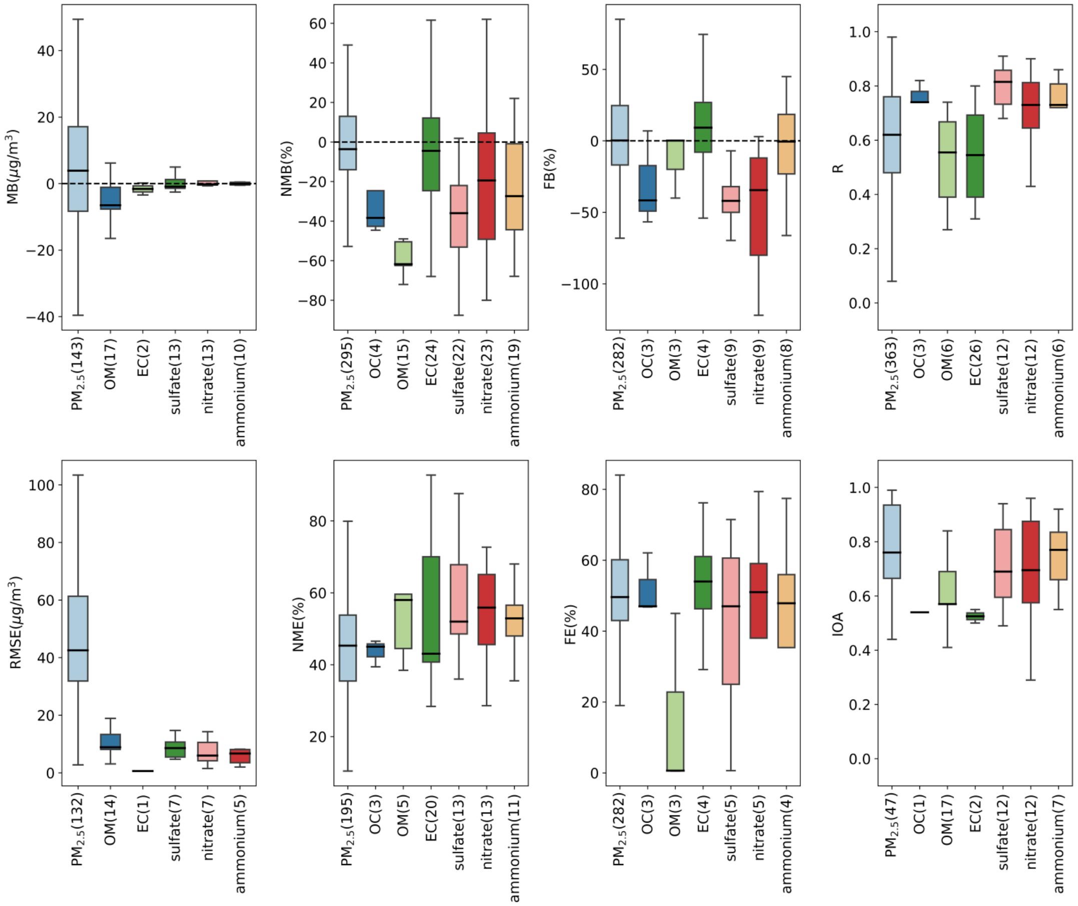


Figure 4: Quantile distribution of selected PM performance metrics compiled in this work. Median values are shown as centerlines; the upper and lower bound of boxes correspond to the 25th and 75th percentile values; whiskers extend to 1.5 times the interquartile range (outliers are excluded).

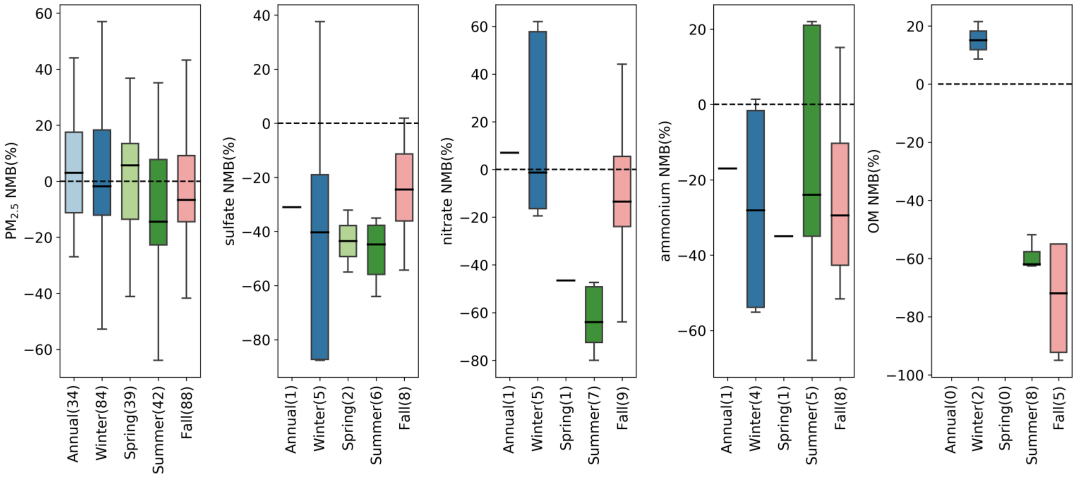


Figure 5: NMB of total $PM_{2.5}$ and speciated components split by season.

https://doi.org/10.5194/acp-2020-237
Preprint. Discussion started: 9 July 2020

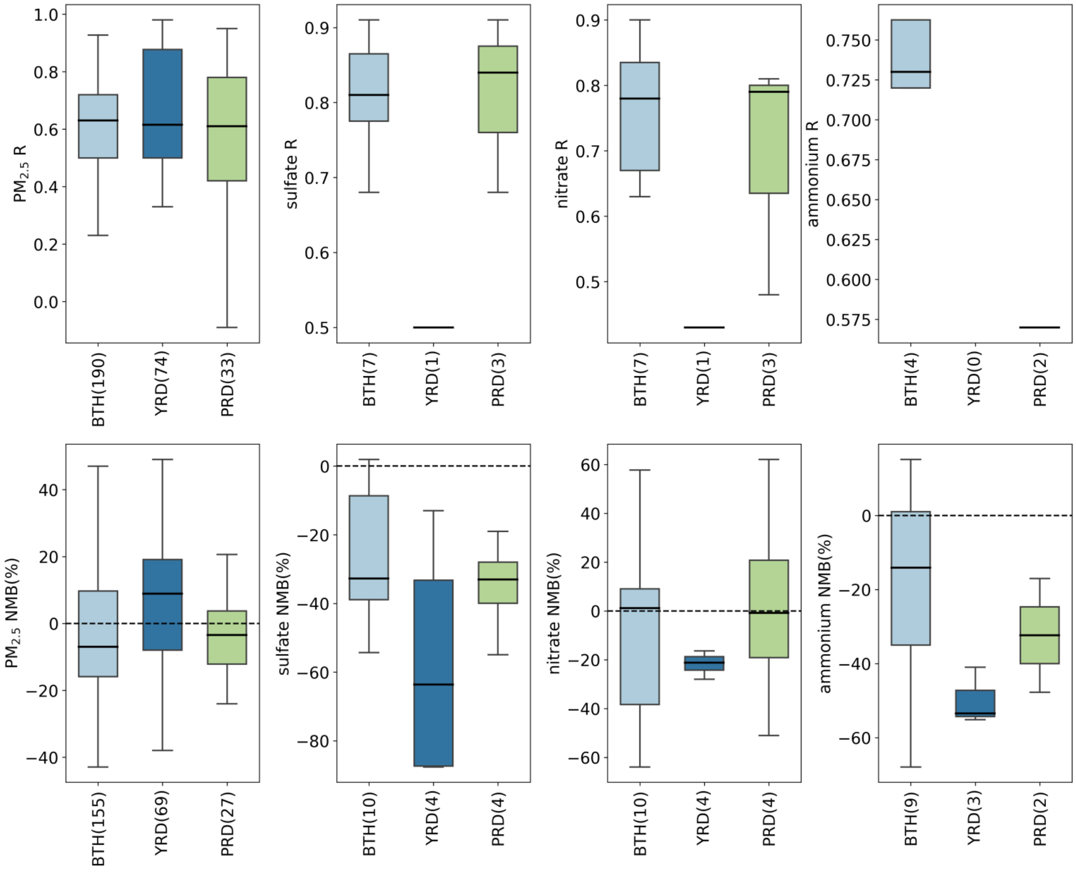


Figure 6: Quantile distribution of R and NMB of total $PM_{2.5}$ and speciated species in BTH, YRD and PRD

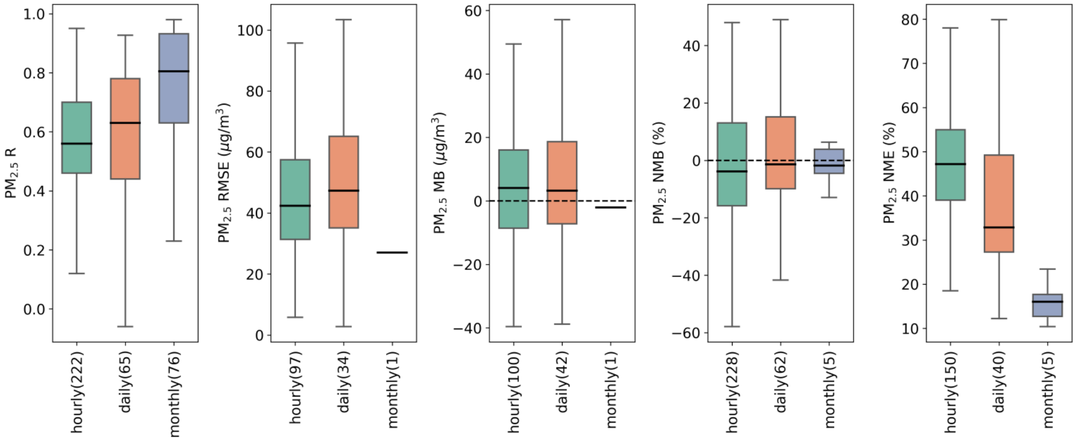


Figure 7: Quantile distributions of R, MB, NMB and NME of total $PM_{2.5}$ presented by temporal resolution for model validation

https://doi.org/10.5194/acp-2020-237
Preprint. Discussion started: 9 July 2020

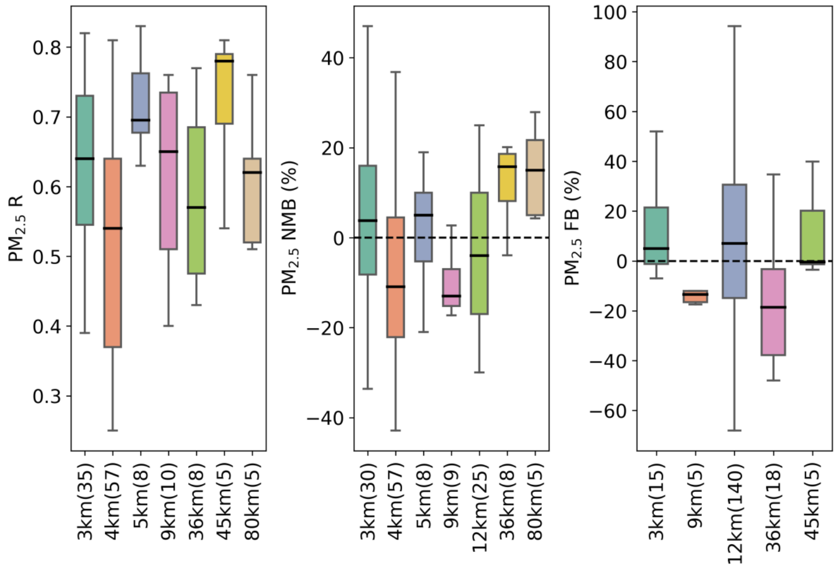
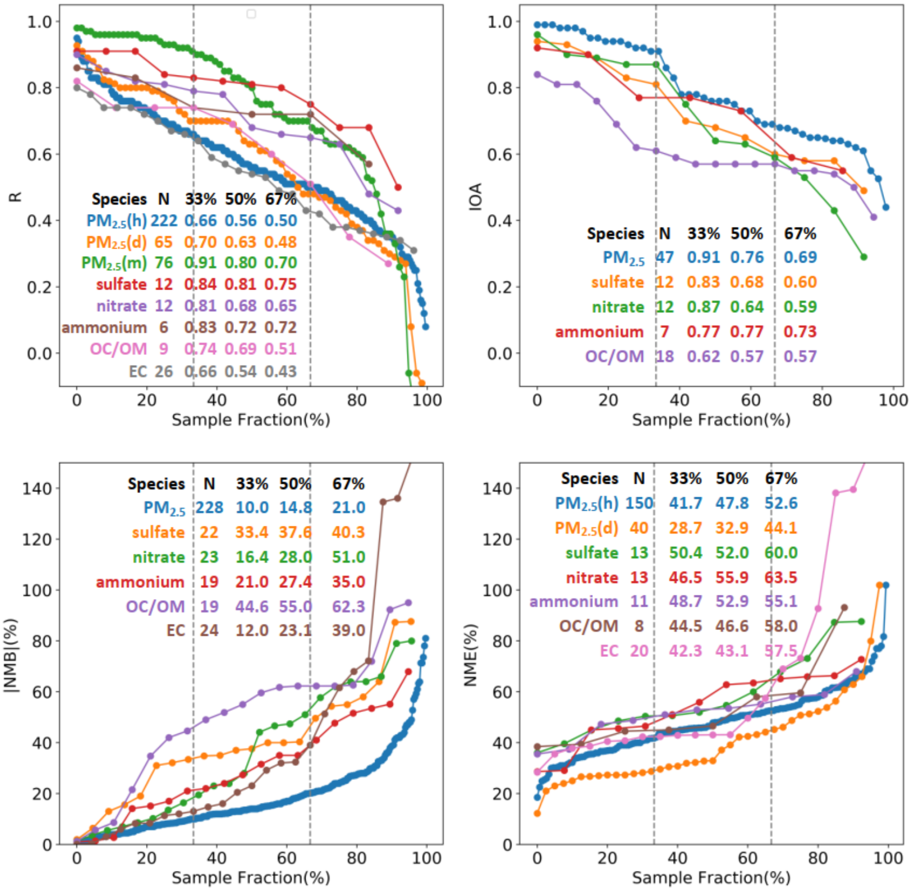


Figure 8: Quantile distributions of R, NMB and FB of total PM$_{2.5}$ presented by model grid resolution



Species	N	33%	50%	67%
PM$_{2.5}$(h)	222	0.66	0.56	0.50
PM$_{2.5}$(d)	65	0.70	0.63	0.48
PM$_{2.5}$(m)	76	0.91	0.80	0.70
sulfate	12	0.84	0.81	0.75
nitrate	12	0.81	0.68	0.65
ammonium	6	0.83	0.72	0.72
OC/OM	9	0.74	0.69	0.51
EC	26	0.66	0.54	0.43

Species	N	33%	50%	67%
PM$_{2.5}$	47	0.91	0.76	0.69
sulfate	12	0.83	0.68	0.60
nitrate	12	0.87	0.64	0.59
ammonium	7	0.77	0.77	0.73
OC/OM	18	0.62	0.57	0.57

Species	N	33%	50%	67%
PM$_{2.5}$	228	10.0	14.8	21.0
sulfate	22	33.4	37.6	40.3
nitrate	23	16.4	28.0	51.0
ammonium	19	21.0	27.4	35.0
OC/OM	19	44.6	55.0	62.3
EC	24	12.0	23.1	39.0

Species	N	33%	50%	67%
PM$_{2.5}$(h)	150	41.7	47.8	52.6
PM$_{2.5}$(d)	40	28.7	32.9	44.1
sulfate	13	50.4	52.0	60.0
nitrate	13	46.5	55.9	63.5
ammonium	11	48.7	52.9	55.1
OC/OM	8	44.5	46.6	58.0
EC	20	42.3	43.1	57.5

5 **Figure 9：Rank-ordered distributions of R, IOA, NMB and NME for total PM$_{2.5}$ and speciated components. The number of data points and the 33rd, 50th, and 67th percentile values are also listed.**

https://doi.org/10.5194/acp-2020-237
Preprint. Discussion started: 9 July 2020

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]

	NMB		NME		R		IOA	
Species	**Goal**	**Criteria**	**Goal**	**Criteria**	**Goal**	**Criteria**	**Goal**	**Criteria**
hourly $PM_{2.5}$	<±15%	<±25%	<45%	<55%	>0.60	>0.45	>0.90	>0.65
daily $PM_{2.5}$	<±15%	<±25%*	<30%*	<45%*	>0.65	>0.45*	>0.90	>0.65
monthly $PM_{2.5}$	<±15%	<±25%	<30%	<45%	>0.90	>0.65	>0.90	>0.65
sulfate	<±35%	<±45%	<55%	<65%	>0.80*	>0.70*	>0.80	>0.60
nitrate	<±20%	<±55%*	<50%*	<65%*	>0.80	>0.65	>0.85	>0.55
ammonium	<±25%	<±40%	<50%	<60%	>0.80*	>0.70*	>0.75	>0.70
OC/OM	<±45%	<±65%	<45%	<60%	>0.70	>0.50	>0.60	>0.50
EC	<±15%*	<±40%	<45%*	<60%*	>0.65	>0.40	none	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 10 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)