Atmospheric Environment 319 (2024) 120290

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Recommendations on benchmarks for photochemical air quality model applications in China — NO_2, SO_2, CO and PM_{10}

Hehe Zhai[a], Ling Huang[a,**], Chris Emery[b], Xinxin Zhang[a], Yangjun Wang[a], Greg Yarwood[b], Joshua S. Fu[c], Li Li[a,*]

[a] School of Environmental and Chemical Engineering, Shanghai University, Shanghai, 200444, China
[b] Ramboll, Novato, CA, 94945, USA
[c] Department of Civil and Environmental Engineering, University of Tennessee, Knoxville, TN, 37996, USA

HIGHLIGHTS

- A set of benchmarks for evaluation of the NO_2, SO_2, CO and PM_{10} simulation over China are proposed.
- Three commonly used statistical indicators (NMB, NME, and R) were proposed for the validation of the four pollutants.
- The goal benchmarks of NMB for NO_2, SO_2, PM_{10}, and CO are <20%, <25%, <20% and <25%, respectively.

ARTICLE INFO

Keywords:
Model performance evaluation
Benchmarks
Air quality models
China

ABSTRACT

Photochemical air quality models (AQMs) are a vital tool for atmospheric pollution research and have been widely used in various applications, such as air quality prediction and evaluation of pollution control strategies. Before using these models for further studies, it is essential to thoroughly evaluate their reliability and accuracy. While previous guidelines and benchmarks have primarily focused on fine particulate matter ($PM_{2.5}$) and ozone (O_3), there is still a lack of benchmarks for evaluating the model performance on primary criteria pollutants such as sulfur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), and particulate matter with aerodynamic diameter less than or equal to 10 μm (PM_{10}). The use of air quality models in China has increased significantly in the past decades. However, there is still a lack of standardized benchmarks for the performance evaluation of these models. Building upon our previous work on $PM_{2.5}$ and its chemical species, we propose a set of benchmarks for evaluating the performance of the aforementioned four air pollutants. Initially, we identified a total of 475 papers published during 2007–2019 that utilized at least one of the five commonly used AQMs in China. From these papers, we selected 164 articles that provided model performance evaluation (MPE) results of the four primary air pollutants. The three most frequently used Model Performance Evaluation (MPE) metrics were selected to analyse the impact of different model configurations on the reported statistics, including modelling region, season, and emission inventory. Lastly, three commonly used statistical indicators, including normalized mean bias (NMB), normalized mean error (NME), and correlation coefficient (R), were proposed for the validation of simulated NO_2, SO_2, CO, and PM_{10}. Two sets of benchmarks are given, including the "goal" and "criteria". The "goal" represents the best range of performance that a model can be expected to achieve, and the "criteria" represents performance that the majority of studies have achieved. We recommend R values above 0.50, 0.35, 0.45, and 0.40 for NO_2, SO_2, PM_{10}, and CO, respectively, in order to meet the "criteria" benchmark. If the "goal" benchmark is to be achieved, the corresponding R values are 0.60, 0.55, 0.60, and 0.60. The "goal" benchmarks of NMB for NO_2, SO_2, PM_{10}, and CO are within ±20%, ±25%, ±20%, and ±25%, respectively; while the "goal" benchmarks of NME for the four pollutants are less than 40%, 45%, 45%, and 60%, respectively. These benchmarks supplement our previous benchmarks for $PM_{2.5}$ and its components and provide a more comprehensive guideline for the air quality modelling community in China.

* Corresponding author.
** Corresponding author.
E-mail addresses: linghuang@shu.edu.cn (L. Huang), lily@shu.edu.cn (L. Li).

https://doi.org/10.1016/j.atmosenv.2023.120290
Received 27 July 2023; Received in revised form 25 October 2023; Accepted 10 December 2023
Available online 13 December 2023

1. Introduction

Over the past decades, air pollution has become a growing concern in China, leading to increased use of air quality models (AQMs) to investigate causes of air pollution, chemistry and transport, the effectiveness of emission reduction strategies, and to provide air quality forecast to the public (Cheng et al., 2019; Huang et al., 2019; Shi et al., 2021). Evaluating the performance of these models against measured ambient pollutant concentrations recorded by monitoring networks is essential to determine whether modelling results align with expected outcomes. However, there is no universally accepted "good or bad" measure of model performance; rather, the quality of the model depends on the research objectives, such as chemical sensitivity, pollution source analysis, or air quality prediction. The reliability of subsequent research is reflected in the results of model performance evaluations. In 1991, the US Environmental Protection Agency (EPA) issued a guidance document that provided indicators for evaluating the performance of models in simulating ozone levels. Boylan and Russell (2006) first proposed the concept of "goals" and "criteria" for assessing model performance, in which "goals" represent the maximum extent of model-observation agreement models can technically achieve given known inherent limitations, and "criteria" represent an acceptable level of model-observation agreement for modelling applications. These concepts were later updated by Emery et al. (2017) and extended to recommend evaluation procedures and benchmarks for ozone and PM evaluation in North America. Similarly, the Forum for Air Quality Modelling In Europe (FAIRMODE) provided recommendations and guidance for assessing model performance related to a given air quality model application in the frame of the Air Quality Directive (AQD) based on the experience and elaborations in the FAIRMODE community (Janssen et al., 2022). However, it is important to note that different regions may have unique factors that can impact model performance, such as meteorology, local topography, regulatory settings, etc. The availability and accuracy of local emission inventories, including the magnitude of emissions and their speciation, as well as details of temporal and spatial variations, can greatly influence model outcomes. Therefore, the benchmarks that are applicable to one region may not be appropriate for other regions. Some regions, particularly those lacking local emission inventories, may find the benchmarks too stringent, while others that have invested significant efforts in improving their emission inventories may find them too lenient.

In our previous work (Huang et al., 2021), we introduced a set of benchmarks to evaluate the performance of AQMs in reproducing the spatial and temporal variations of $PM_{2.5}$ and its chemical components in China. However, the existing guidelines and benchmarks for evaluating model performance have mainly focused on $PM_{2.5}$ and O_3 simulations (Huang et al., 2021; Yang and Zhao, 2023), neglecting other important criteria pollutants like SO_2, NO_2, CO, and PM_{10}. This gap in model performance evaluation has been largely overlooked thus far. This study aims to expand the benchmark framework for evaluating model performance with respect to primary air pollutants, specifically NO_2, SO_2, PM_{10}, and CO. Unlike $PM_{2.5}$, which consists of both primary and secondary portions, the accuracy of primary pollutants is heavily influenced by the precision (both temporal and spatial representation) of the emission inventory and simulated meteorological conditions. Our study follows a similar structure to our previous study (Huang et al., 2021). Section 2 outlines the data source and methodology employed, while Section 3 presents the major results, including the impact of different model configurations on model performance. The paper concludes by offering recommended evaluation metrics and associated benchmarks for each air pollutant. By expanding the benchmark framework for assessing the performance of AQMs for primary air pollutants, we aim to contribute to a more comprehensive evaluation of model performance and enhance the reliability of air quality models in China.

2. Methodology

2.1. Data compilation

As in our previous work, published results from five AQMs were compiled, which are the Community Multiscale Air Quality (CMAQ, Foley et al., 2010), the Comprehensive Air Quality Model with Extensions (Campbell et al., 2017), the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem, Grell et al., 2005), the Nested Air Quality Prediction Modelling System (NAQPMS, Wang et al., 2006), and the Goddard Earth Observing System (GEOS)-Chem (http://geos-chem.org, last access: October 21, 2023). We conducted a comprehensive search of the Web of Science database using the keywords "China", "model name", and "pollutant name" (one of PM_{10}, SO_2, NO_2, CO) from 2007 to 2019. This yielded a total of 475 relevant articles. The same selection procedure as our previous work (Huang et al., 2021) was then applied, resulting in a final set of 164 articles (See Table S1 for details). During the selection process, conference articles and articles not published in English-language journals were excluded. Each article was manually reviewed to eliminate studies that did not utilize any of the models in their research, studies that focused on objectives other than air quality modelling (such as evaluating meteorological simulations), studies that were not cantered on China (for example, those targeting regions in Korea or Japan), studies that did not provide any evaluation of air quality model performance and studies that conducted model performance evaluation but did not present numerical values (only graphical plots were provided). Values of a total of 26 statistical indicators reported in the 164 articles were collected (see Table S2 for a list of statistical metrics used in the complied studies), together with the detailed model configurations, including the study region and period, the grid resolution of model simulation, the source of emission inventory used for model validation, etc. We focused on discussing the results of several metrics that have a sufficient number of reported values, which are the correlation coefficient (R), normalized mean bias (NMB), and normalized mean error (NME). The correlation coefficient reflects the model's ability to capture the observed spatial variations, while NMB and NME indicate how well models capture the magnitude of observations. The calculation formula is shown in Table S3 and all abbreviations are shown in Table S6.

2.2. Derivation of benchmarks

We derived the recommended MPE benchmarks for the four primary criteria pollutants following the methodology established by Emery et al. (2017) and employed in our previous study (Huang et al., 2021). The rank-ordered distributions of each selected MPE indicator were generated to determine the 33rd and 67th percentiles, representing the "goal" and "criteria" benchmarks, respectively. Studies within the 33rd percentile range represent the best model performance that models can be expected to achieve given known inherent limitations such as discretization, parameterizations, and input uncertainties. Studies within the 33rd and 67th percentile range indicate model performance that is typically achieved, and studies beyond the 67th percentile indicate relatively poor performance for that particular metric.

3. Results and discussions

3.1. General overview of air quality modelling studies for primary criteria pollutants in China

The number of modelling studies on the four primary air pollutants in China has seen a significant increase over the past years, as shown in Fig. 1a. The rise in awareness of environmental protection in China and increased research funding from the government have contributed to this growth. A breakdown of the number of studies conducted for each of the four pollutants reveals a gradual shift of research focus from PM_{10}

prior to 2015 to NO_2 after 2015 (Fig. 1b). This shift is likely associated with the increased attention given to coarse particulate matter pollution in earlier years and then to $PM_{2.5}$ and ozone, of which NO*x* is a critical precursor for both pollutants. In terms of the models used, CMAQ was the most frequently employed (72 out of 169), followed by WRF-Chem (69 studies). CAMx and NAQPMS were used less frequently, with 15 and 9 studies, respectively, while the GEOS-Chem global model was the least frequently used (4 studies). The limited use of GEOS-Chem can be attributed to its primary focus on global pollutant sources and transport, with fewer studies specifically examining China. Regarding the MPE metrics, the most commonly used indicators include correlation coefficient (R, reported in 115 studies), NMB (86 studies), root mean square error (RMSE, 73 studies), mean bias (MB, 67 studies), NME (59 studies), IOA (32 studies), fractional bias (FB, 31 studies), and fractional error (FE, 28 studies) (Fig. 1c). Same as studies for $PM_{2.5}$ (Huang et al., 2021), Beijing-Tianjin-Hebei (BTH, 64 studies) has been the most extensively studied, followed by Yangtze River Delta (YRD, 39 studies) and Pearl River Delta (PRD, 40 studies, see Table S4 for definition of regions). These regions are of particular interest due to their high level of economic development, population density, and the obvious air pollution in China.

3.2. Quantile distributions of selected MPE metrics

Fig. 2 illustrates the distributions of the three most frequently reported model performance metrics for the four primary air pollutants. The median value, representing the middle point of the data, is indicated by the centre line, with 50% of the data falling above and below the median. The mean value is represented by the small box, where the upper and lower limits correspond to the 25th and 75th percentile values. The whisker line extends to 1.5 times the interquartile range. Outliers were excluded. The median R values were 0.58 for NO_2, 0.48 for SO_2, 0.58 for PM_{10}, and 0.52 for CO. Some studies reported negative R values for all pollutants except NO_2. For PM_{10}, negative R values may have resulted from the model underestimating precipitation rates (Bouarar et al., 2019). In terms of NMB, all four pollutants were generally underestimated, with median values of −4.0% for NO_2, -1.4% for SO_2, -19.0% for PM_{10}, and -32.9% for CO. In particular, PM_{10} and CO were particularly underestimated, which could be due to the underestimation of local CO emissions (Zhang et al., 2016) or the inability of the model grid to adequately resolve specific sources, such as traffic-oriented monitoring or large point sources. For NME, the median values ranged from 47.0% (NO_2) to 61.0% (CO). SO_2 exhibited the widest range of NMB (−84.2%–140.5%) and NME (0.3%–147.0%) compared to the other three pollutants. This might be due to (1) SO_2 concentration is the lowest in terms of absolute magnitude; (2) SO_2 emissions are more strongly associated with high stacks, so underestimating or overestimating emissions can lead to poor simulations (Liu et al., 2010; Zhang et al., 2019); and (3) uncertainties in the meteorological modelling, calculation of dry deposition rates, and oxidation rates of SO_2 (Biswas et al., 2020; Matsui et al., 2009).

Compared to secondary air pollutants (e.g., $PM_{2.5}$, O_3), the model performances of these primary pollutants are more influenced by the location and sensitivity of monitors and the accuracy of the emission inventory. For example, PM_{10} was sensitive to deposition on vegetation surfaces surrounding the monitor (Langner et al., 2011). Monitor sensitivity limitations could be important for NO_2 as PAN and other compounds could be detected as NO_2 (Gaffney et al., 1998). PM_{10} could also have intense sources that were less intense for other pollutants, such as quarries. In addition, construction dust, an important source of PM_{10}, exhibited stronger annual variations due to project completion. Thus, the quality of PM_{10} emission inventory may be decoupled from that of other pollutants (Hong et al., 2017; Zheng et al., 2009). The evaluation uncertainties may also arise from using a monitoring station to validate the simulated results from the nearest model grid (Kumar et al., 2022). AQMs can be applied with different spatial resolutions, from less than a few km to more than 50 km. As shown in Fig. S1, simulations conducted with a spatial resolution finer than 10 km generally show higher R values than spatial resolution coarser than 10 km, especially for NO_2. The NMB, however, does not differ significantly among simulations using different spatial resolutions. Furthermore, for species like NO_2, estimation of solar intensity and photolysis rate by different models can also affect the simulated results. For example, CAMQ adopts the JPROC photolysis mechanism derived from the Regional Acid Deposition Model (RADM) (Byun and Ching, 1999), while the WRF-Chem applies the Fast-J mechanism (Wild et al., 2000).

3.3. Variations in model performance

As shown in Fig. 2, the model performance of a single air pollutant could vary substantially in different applications with different model configurations. The analyses below provide a supplemental reference for the performance ranges for the specific factors presented. Note that we only present results with a sufficient number (>20, except CO) of data points for discussions.

3.3.1. Impact of season

Fig. 3 shows the R and NMB distributions for four pollutants across different seasons. In terms of seasonal distribution, the amount of data for PM_{10} was significantly higher than other pollutants in spring due to the presence of more dusty weather in spring (Bao et al., 2021). In winter, the amounts of PM_{10}, SO_2, and NO_2 are higher, due to the higher use of coal in winter (Fan et al., 2020; Zhang et al., 2017). In terms of R values, the simulated NO_2 exhibited the lowest R for spring (median value: 0.46) compared to other seasons. NO_2 tended to be more underestimated in spring and fall (median NMB values were −8.7% and −12.4%, respectively), while overestimated in winter (median NMB = 8.3%) and summer (median NMB = 9.8%). When evaluating NO_2 on an

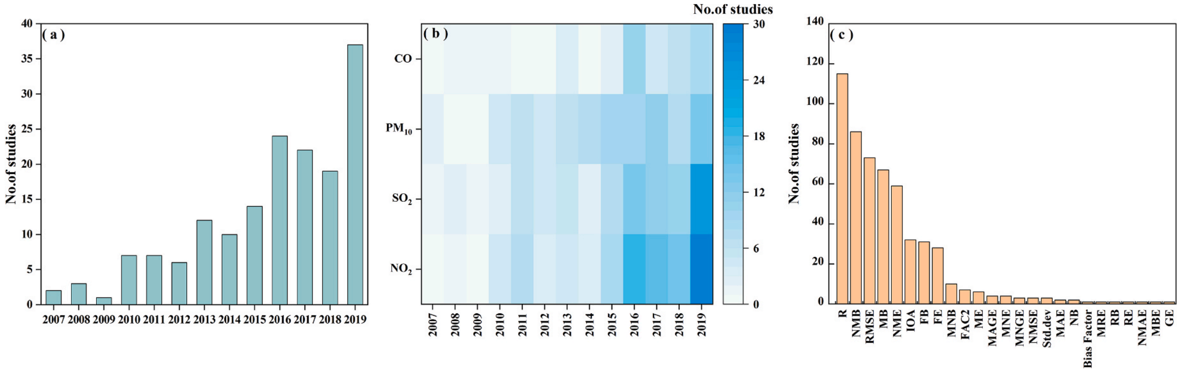

Fig. 1. (a) Number of studies published during 2007–2019, (b) number of studies evaluating each pollutant and AQM model pair, and (c) frequency of the use of each metric.

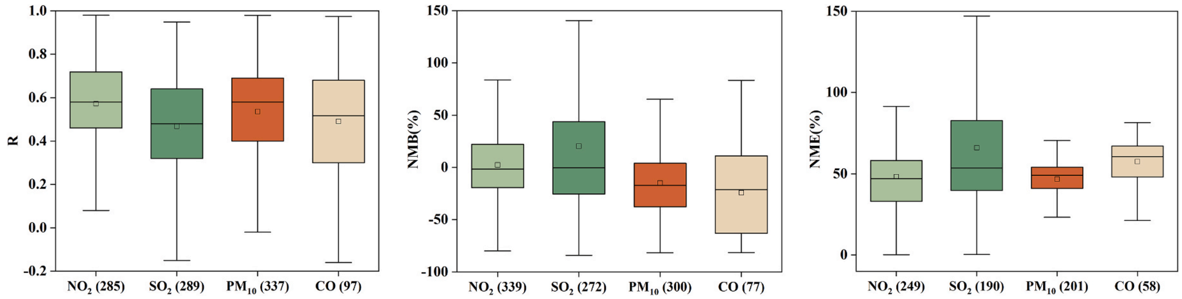

Fig. 2. Quantile distribution of NO_2, SO_2, PM_{10}, and CO performance metrics compiled.

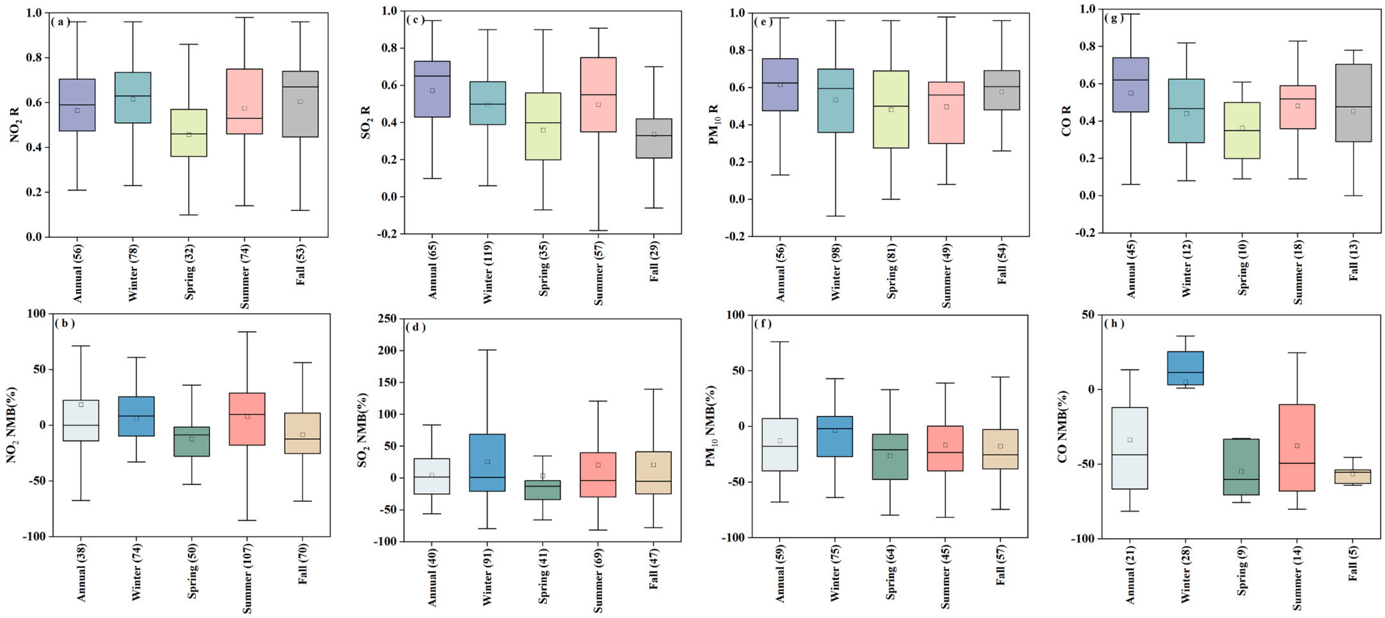

Fig. 3. Quantile distributions of R and NMB of NO_2, SO_2, PM_{10}, and CO by season.

annual scale, the median NMB value was close to zero, indicating an equal number of studies reporting negative and positive NMB values. Overall, NO_2 generally tended to be better simulated in summer and winter. For SO_2, the median R values for all seasons, except for fall, were higher than 0.50. The median NMB values for SO_2 were generally within ±5.0%, except for spring (median NMB = −12.9%). It is worth noting that the simulated SO_2 in winter exhibited a wider range of NMB (ranging −79.6%–124.0%), indicating higher uncertainties in wintertime SO_2 emissions, probably associated with a lack of several SO_2 reaction mechanisms, such as heterogeneous reactions (et al. Sha et al., 2019a; Zheng et al., 2019). Compared to NO_2 and SO_2, PM_{10} exhibited a better performance in terms of R values (median values > 0.5) but tended to be underestimated across all seasons (median NMB ranging from −25.6% to −1.9%). The underestimation observed in winter could be attributed to the absence of dust emissions in the simulations (Liu et al., 2018) and the perception that emissions are underestimated (Koo et al., 2015). The prevalence of dust weather in spring leads to higher PM_{10} concentrations (Filonchyk et al., 2018), which may explain the greater underestimation of NMB during this season. CO was largely underestimated across all seasons (median NMB: 60.1% to −50.0%), except for winter (median NMB: 11.5%). The performance in spring and fall was particularly poor, with NMB values indicating complete underestimation. Additionally, the R values for CO in spring were the lowest (median R: 0.35) among all seasons. This could be attributed to significant biomass burning activities in Southeast Asia during spring, resulting in high background CO concentrations in the modelled region (Lin et al., 2014). These activities may not be adequately accounted for in the emission inventory.

3.3.2. *Impact of modelling region*

The variations in model performance across different regions were also investigated. However, due to limitation in data availability, the comparison was only conducted for the three most extensively studied regions, namely the BTH, YRD, and PRD (Fig. 4). For NO_2, BTH shows a slightly higher median R value (0.67) compared to the other two regions (less than 0.60). PRD tended to underestimate NO_2, with a median NMB of −11.8%, whereas the other two regions had roughly equivalent underestimation and overestimation. For SO_2, the median R values were similar across the three regions (0.40–0.54), but the NMB shifted from overestimation in the north to underestimation in the south. This may be attributed to the overestimation of SO_2 emissions, as indicated by previous studies (Matsui et al., 2009; Yue et al., 2018). PM_{10} showed a slightly higher median R value for BTH (0.66) than YRD and PRD (0.50 and 0.58). The median NMB values were negative (−15.6% to −9.0%) in all three regions. The amount of data for CO was the least among the four pollutants. The lowest median R value (0.48) was found for YRD and the highest (0.67) for BTH. CO was significantly underestimated in all three regions, particularly in the PRD (median NMB: 64.1%). Overall, there was a trend towards lower correlations and more underestimations from north to south for SO_2 and CO. However, for NO_2 and PM_{10}, the three regions exhibit similar MPE results.

3.3.3. *Impact of emission inventory*

The accuracy of the emission inventory is especially crucial for the

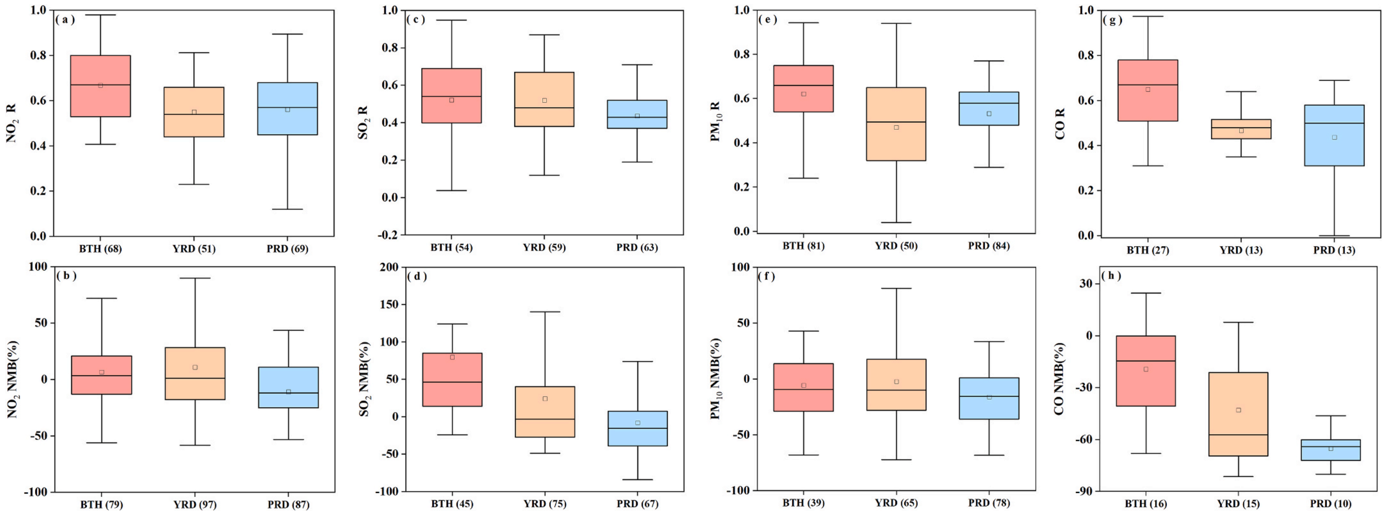

Fig. 4. Quantile distribution of R and NMB of NO_2, SO_2, PM_{10}, and CO for BTH, YRD, and PRD.

model performance of these primary air pollutants. A local emission inventory refers to one that is specifically developed for a particular region, such as a city or province (e.g., Zhou et al., 2017), as opposed to a national or global inventory. The development of local emission inventories involves the use of local emission factors, measured source profiles, and detailed emission activity data to improve the accuracy of the inventory (Lang et al., 2017; Liu et al., 2018). Thus, it is generally expected a locally developed emission inventory should result in better model performance.

In this study, we divided MPE results into two groups (Fig. 5): studies that used locally developed emission inventories and studies that used non-local emission inventories. It is worth noting that locally developed emission inventories are typically used for nested domains with finer spatial resolutions, while national or regional emission inventories are used for coarser spatial resolutions and studies that cover multiple regions. Overall, the use of local emission inventories tends to lead to slightly higher and narrower distribution of R values for all four pollutants. However, the distribution of NMB values does not show noticeable improvement when using local emission inventories compared to non-local inventories. Specifically, for R values, using a local emission inventory generally leads to a narrower distribution for NO_2, SO_2, and PM_{10}, with the median R value ranging from 0.57 to 0.60. In contrast, the median R value for non-local emission inventories ranges from 0.44 to 0.57. This suggests the simulations using local emission inventories are more consistent with ground observations. For CO simulations, the median R values for non-local emission inventories (0.53) are slightly higher than that of the local emission inventory R value (0.52), but the former is more scattered. In terms of NMB, both PM_{10} and CO local emission inventories exhibit more scattered distributions.

3.4. Recommended metrics and benchmarks

Fig. 6 shows the rank-ordered distributions of R, NMB, and NME results for all pollutants examined in this study (for statistical purposes, all values were sorted after taking absolute values). The 33rd and 67th percentiles of R were calculated for each pollutant and metric. In terms of R, NO_2 shows the highest values for both percentiles, with 0.68 for the 33rd percentile and 0.50 for the 67th percentile. This was followed by PM_{10} (0.65 and 0.47), CO (0.64 and 0.42), and SO_2 (0.58 and 0.39). In other words, an R value of 0.60 for SO_2 simulation would be considered among the top 33rd percentile, whereas for NO_2, it would only be around the 50th percentile. For bias and error, NO_2 also demonstrated the best performance compared to other pollutants (except for the 67th NME value). The 33rd percentile of absolute NMB and NME for NO_2 were 13.0% and 38.0%, respectively, and the corresponding 67th percentile values were 27.0% and 54.4%. CO exhibited the greatest variability in terms of absolute NMB, ranging from 18.5% to 57.3% for the 33rd to 67th percentile interval. The overall trend of NME was similar to NMB but with much weaker variability. For PM_{10} and CO, the 33rd to 67th range of NME was 43.7%–52.0% and 55.7%–64.0%, with both showing a difference of less than 10%.

Based on the above analysis, the recommended statistical indicators and associated benchmarks for the four primary air pollutants are presented in Table 1. The first one-third of studies represent the "goal" values, indicating the best level that the current models are expected to achieve. The first two-thirds of the studies represent the "criteria" values,

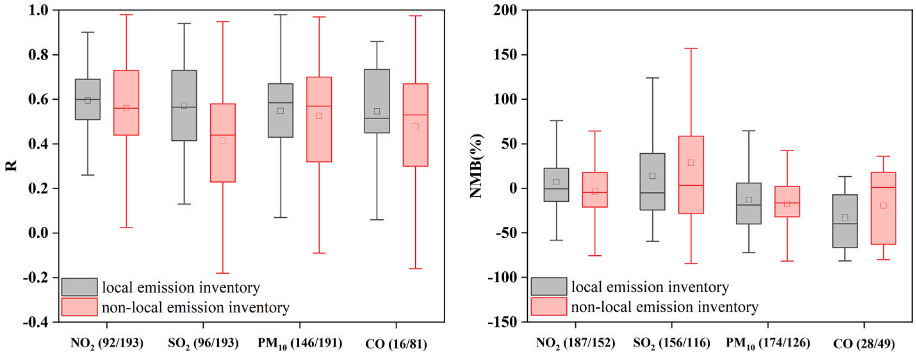

Fig. 5. Selection of emission inventory on model performance. (The left side of the number in parentheses represents the number of data points that used a local emission inventory and the right represents number of data points used a non-local emission inventory).

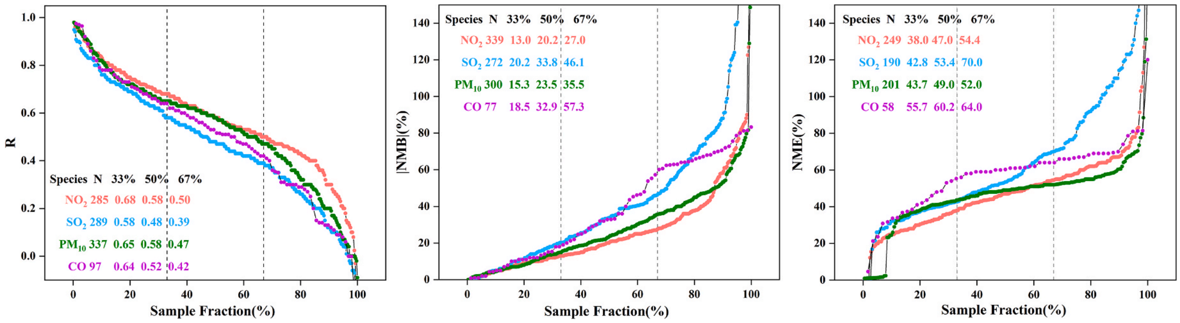

Fig. 6. Rank-ordered distributions of R, NMB, and NME for NO_2, SO_2, PM_{10}, and CO. The number of data points and the 33rd, 50th, and 67th percentile values are also listed.

Table 1
Recommended benchmarks for evaluating AQM applications in China for NO_2, SO_2, PM_{10}, and CO.

Metrics	Benchmark level	NO_2	SO_2	PM_{10}	CO
R	Goal	>0.60	>0.55	>0.60	>0.60
	Criteria	>0.50	>0.35	>0.45	>0.40
NMB	Goal	<±20.0%	<±25.0%	<±20.0%	<±25.0%
	Criteria	<±30.0%	<±50.0%	<±40.0%	<±60.0%
NME	Goal	<40.0%	<45.0%	<45.0%	<60.0%
	Criteria	<55.0%	<70.0%	<55.0%	<65.0%

which most studies can meet. For example, to meet the goal and criteria benchmarks for NMB of NO_2, the values should be within 20.0% and 30.0%, respectively. The corresponding values for NME are 40.0% (goal) and 55.0% (criteria). In terms of R, the recommended benchmark value for NO_2 is 0.60 for “goal” and 0.50 for “criteria”. Due to the relatively small amount of data points for CO, caution should be taken when applying these benchmarks. Table S5 compares the proposed benchmarks for criteria pollutants with existing benchmarks for $PM_{2.5}$ from our previous study (Huang et al., 2021) and benchmarks for $PM_{2.5}$ and O_3 from Emery et al. (2017), which were derived from simulation results conducted for North America. Compared to the benchmarks for $PM_{2.5}$ and O_3, the benchmarks for the four criteria pollutants proposed in this study are much less stringent (except for $PM_{2.5}$ in Emery’s study), especially for NMB and NME.

4. Conclusions

This study presents a comprehensive analysis of the application of photochemical air quality models in China, with a focus on four primary air pollutants, namely NO_2, SO_2, PM_{10}, and CO. A total of 164 published papers that reported extractable model performance results were compiled. Key model configurations, such as model types, study areas, emission inventories, and results of performance evaluation metrics, are analysed and discussed. The impact of different model configurations on the model performance is examined, including the study area, study period, spatial resolution, and choice of emission inventory. This study proposes a set of benchmarks for three widely used metrics, namely R, NMB, and NME, for each of the four pollutants, based on the principles of "goals" and "criteria". To meet the "criteria" benchmark, the recommended R values for NO_2, SO_2, PM_{10}, and CO are above 0.50, 0.35, 0.45, and 0.40, respectively. If the "goal" benchmark is to be achieved, the corresponding R values are 0.60, 0.55, 0.60, and 0.60. The "goal" benchmarks of NMB for NO_2, SO_2, PM_{10}, and CO are <±20%, <±25%, <±20% and <±25%, respectively. The "goal" benchmarks for NME are <40%, <45%, <45%, and <60% for the four pollutants. The "criteria" benchmarks of NMB for NO_2, SO_2, PM_{10}, and CO are <±30%, <±50%, <±40% and <±60%. The "criteria" benchmarks of NME for the four pollutants are <55%, <70%, <55% and <65%, respectively. The findings of this study are part of a broader effort to establish quantitative and objective performance benchmarks for air quality modelling in China. By focusing on primary pollutants and introducing additional benchmarks for key indicators, this study contributes to the development of a more comprehensive evaluation system. It fills the gap in China’s AQMs benchmark for primary criteria pollutants. These performance benchmarks serve as a valuable tool for researchers to identify any discrepancies between their simulation results and expected outcomes from other studies. In cases where a significant gap exists, researchers can then investigate the reasons behind the disparities, such as the accuracy of emission inventory. Moreover, these benchmarks play a vital role in building confidence in the base case simulation results. These results are used to evaluate the effectiveness of emission reduction policy for the purpose of air quality management and planning, of which the ultimate goal is to protect public health and the environment.

CRediT authorship contribution statement

Hehe Zhai: performed the data analysis and prepared the manuscript with contributions from all co-authors, collected data and conducted data analysis. **Ling Huang:** performed the data analysis and prepared the manuscript with contributions from all co-authors. **Chris Emery:** contributed to academic discussions and review. **Xinxin Zhang:** collected data and conducted data analysis. **Yangjun Wang:** contributed to academic discussions and review. **Greg Yarwood:** contributed to academic discussions and review. **Joshua S. Fu:** contributed to academic discussions and review. **Li Li:** formulated the research goals and edited the manuscript.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Data availability

Data will be made available on request.

Acknowledgement

This study was financially sponsored by the National Natural Science Foundation of China (No. 42005112, 42075144, 42375102).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2023.120290.

References

Bao, C., Yong, M., Bi, L., Gao, H., Li, J., Bao, Y., Gomboludev, P., 2021. Impacts of underlying surface on the dusty weather in central inner Mongolian steppe, China. Earth Space Sci. 8, e2021EA001672 https://doi.org/10.1029/2021EA001672.

Biswas, K., Chatterjee, A., Chakraborty, J., 2020. Comparison of air pollutants between Kolkata and siliguri, India, and its relationship to temperature change. J. Geovis. Spat. Anal. 4, 25. https://doi.org/10.1007/s41651-020-00065-4.

Bouarar, I., Brasseur, G., Petersen, K., Granier, C., Fan, Q., Wang, X., Wang, L., Ji, D., Liu, Z., Xie, Y., Gao, W., Elguindi, N., 2019. Influence of anthropogenic emission inventories on simulations of air quality in China during winter and summer 2010. Atmos. Environ. 198, 236–256. https://doi.org/10.1016/j.atmosenv.2018.10.043.

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, Special issue on Model Evaluation: Evaluation of Urban and Regional Eulerian Air Quality Models 40, 4946–4959. https://doi.org/10.1016/j.atmosenv.2005.09.087.

Byun, D.W., Ching, J.K.S., 1999. SCIENCE ALGORITHMS OF THE EPA MODELS-3 COMMUNITY MULTISCALE AIR QUALITY (CMAQ) MODELING SYSTEM.

Campbell, P., Zhang, Y., Wang, K., Leung, R., Fan, J., Zheng, B., Zhang, Q., He, K., 2017. Evaluation of a multi-scale WRF-CAM5 simulation during the 2010 east asian summer monsoon. Atmos. Environ. 169, 204–217. https://doi.org/10.1016/j.atmosenv.2017.09.008.

Cheng, J., Su, J., Cui, T., Li, X., Dong, X., Sun, F., Yang, Y., Tong, D., Zheng, Y., Li, Y., Li, J., Zhang, Q., He, K., 2019. Dominant role of emission reduction in PM2.5 air quality improvement in Beijing during 2013-2017: a model-based decomposition analysis. Atmos. Chem. Phys. 19, 6125–6146. https://doi.org/10.5194/acp-19-6125-2019.

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. J. Air Waste Manag. Assoc. 67, 582–598. https://doi.org/10.1080/10962247.2016.1265027.

Fan, M., He, G., Zhou, M., 2020. The winter choke: coal-Fired heating, air pollution, and mortality in China. J. Health Econ. 71, 102316.

Filonchyk, M., Yan, H., Li, X., 2018. Temporal and spatial variation of particulate matter and its correlation with other criteria of air pollutants in Lanzhou, China, in spring-summer periods. Atmos. Pollut. Res. 9, 1100–1110. https://doi.org/10.1016/j.apr.2018.04.011.

Foley, K.M., Roselle, S.J., Appel, K.W., Bhave, P.V., Pleim, J.E., Otte, T.L., Mathur, R., Sarwar, G., Young, J.O., Gilliam, R.C., Nolte, C.G., Kelly, J.T., Gilliland, A.B., Bash, J. O., 2010. Incremental testing of the community Multiscale air quality (CMAQ) modeling system version 4.7. Geosci. Model Dev. (GMD) 3, 205–226. https://doi.org/10.5194/gmd-3-205-2010.

Gaffney, J.S., Bornick, R.M., Chen, Y.H., Marley, N.A., 1998. Capillary gas chromatographic analysis of nitrogen dioxide and PANs with luminol chemiluminescent detection. Atmos. Environ. 32, 1445–1454. https://doi.org/10.1016/S1352-2310(97)00098-8.

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. Atmos. Environ. 39, 6957–6975. https://doi.org/10.1016/j.atmosenv.2005.04.027.

Hong, C., Zhang, Q., He, K., Guan, D., Li, M., Liu, F., Zheng, B., 2017. Variations of China’s emission estimates: response to uncertainties in energy statistics. Atmos. Chem. Phys. 17, 1227–1239. https://doi.org/10.5194/acp-17-1227-2017.

Huang, L., An, J., Koo, B., Yarwood, G., Yan, R., Wang, Y., Huang, C., Li, L., 2019. Sulfate formation during heavy winter haze events and the potential contribution from heterogeneous SO2 + NO2 reactions in the Yangtze River Delta region, China. Atmos. Chem. Phys. 19, 14311–14328. https://doi.org/10.5194/acp-19-14311-2019.

Huang, L., Zhu, Y., Zhai, H., Xue, S., Zhu, T., Shao, Y., Liu, Z., Emery, C., Yarwood, G., Wang, Y., Fu, J., Zhang, K., Li, L., 2021. Recommendations on benchmarks for numerical air quality model applications in China - Part 1: PM2.5 and chemical species. Atmos. Chem. Phys. 21, 2725–2743. https://doi.org/10.5194/acp-21-2725-2021.

Janssen, S., Thunis, P., Adani, M., Piersanti, A., Carnevale, C., Cuvelier, C., Durka, P., Georgieva, E., Guerreiro, C., Malherbe, L., Maiheu, B., Meleux, F., Monteiro, A., Miranda, A., Olesen, H., Pfäfflin, F., Stocker, J., Sousa Santos, G., Stidworthy, A., Stortini, M., Trimpeneers, E., Viaene, P., Vitali, L., Vincent, K., Wesseling, J., European Commission. Joint Research Centre., 2022. FAIRMODE Guidance Document on Modelling Quality Objectives and Benchmarking: Version 3.3. Publications Office, LU.

Koo, Y., Choi, D., Kwon, H., Jang, Y., Han, J., 2015. Improvement of PM10 prediction in East Asia using inverse modeling. Atmos. Environ. 106, 318–328. https://doi.org/10.1016/j.atmosenv.2015.02.004.

Kumar, A., Dhakhwa, S., Dikshit, A.K., 2022. Comparative evaluation of fitness of interpolation techniques of ArcGIS using leave-one-out scheme for air quality mapping. J. Geovis. Spat. Anal. 6, 9. https://doi.org/10.1007/s41651-022-00102-4.

Lang, J., Zhou, Y., Chen, D., Xing, X., Wei, L., Wang, X., Zhao, N., Zhang, Y., Guo, X., Han, L., Cheng, S., 2017. Investigating the contribution of shipping emissions to atmospheric PM2.5 using a combined source apportionment approach. Environ. Pollut. 229, 557–566. https://doi.org/10.1016/j.envpol.2017.06.087.

Langner, M., Kull, M., Endlicher, W.R., 2011. Determination of PM10 deposition based on antimony flux to selected urban surfaces. Environmental Pollution, Selected papers from the conference Urban Environmental Pollution: Overcoming Obstacles to Sustainability and Quality of Life (UEP2010) 159, 2028–2034. https://doi.org/10.1016/j.envpol.2011.01.017, 20-23 June 2010, Boston, USA.

Lin, C.-Y., Zhao, C., Liu, X., Lin, N.-H., Chen, W.-N., 2014. Modelling of long-range transport of Southeast Asia biomass-burning aerosols to Taiwan and their radiative forcings over East Asia. Tellus Ser. B Chem. Phys. Meteorol. 66, 23733 https://doi.org/10.3402/tellusb.v66.23733.

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

Liu, M., Lin, J., Wang, Y., Sun, Y., Zheng, B., Shao, J., Chen, L., Zheng, Y., Chen, J., Fu, T., Yan, Y., Zhang, Q., Wu, Z., 2018. Spatiotemporal variability of NO2 and PM2.5 over Eastern China: observational and model analyses with a novel statistical method. Atmos. Chem. Phys. 18, 12933–12952. https://doi.org/10.5194/acp-18-12933-2018.

Matsui, H., Koike, M., Kondo, Y., Takegawa, N., Kita, K., Miyazaki, Y., Hu, M., Chang, S., Blake, D., Fast, J., Zaveri, R., Streets, D., Zhang, Q., Zhu, T., 2009. Spatial and temporal variations of aerosols around Beijing in summer 2006: model evaluation and source apportionment. J. Geophys. Res. Atmos. 114 https://doi.org/10.1029/2008JD010906.

Sha, T., Ma, X., Jia, H., Tian, R., Chang, Y., Cao, F., Zhang, Y., 2019a. Aerosol chemical component: simulations with WRF-Chem and comparison with observations in Nanjing. Atmos. Environ. 218 https://doi.org/10.1016/j.atmosenv.2019.116982.

Shi, X., Zheng, Y., Lei, Y., Xue, W., Yan, G., Liu, X., Cai, B., Tong, D., Wang, J., 2021. Air quality benefits of achieving carbon neutrality in China. Sci. Total Environ. 795, 148784 https://doi.org/10.1016/j.scitotenv.2021.148784.

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. J. Atmos. Chem. 55, 253–272. https://doi.org/10.1007/s10874-006-9038-6.

Wild, O., Zhu, X., Prather, M.J., 2000. Fast-J: accurate simulation of in- and below-cloud photolysis in tropospheric chemical models. J. Atmos. Chem. 37, 245–282. https://doi.org/10.1023/A:1006415919030.

Yang, J., Zhao, Y., 2023. Performance and application of air quality models on ozone simulation in China - a review. Atmos. Environ. 293, 119446 https://doi.org/10.1016/j.atmosenv.2022.119446.

Yue, T., Zhang, X., Wang, C., Zuo, P., Tong, Y., Gao, J., Xue, Y., Tong, L., Wang, K., Gao, X., 2018. Environmental impacts of the revised emission standard for air pollutants for boilers during the heating season in Beijing, China. Aerosol Air Qual. Res. 18, 2853–2864. https://doi.org/10.4209/aaqr.2018.02.0046.

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. Atmos. Environ. 124, 285–300. https://doi.org/10.1016/j.atmosenv.2015.07.022.

Zhang, Z., Wang, W., Cheng, M., Liu, S., Xu, J., He, Y., Meng, F., 2017. The contribution of residential coal combustion to PM2.5 pollution over China’s Beijing-Tianjin-Hebei region in winter. Atmos. Environ. 159, 147–161. https://doi.org/10.1016/j.atmosenv.2017.03.054.

Zhang, S., Xing, J., Sarwar, G., Ge, Y., He, H., Duan, F., Zhao, Y., He, K., Zhu, L., Chu, B., 2019. Parameterization of heterogeneous reaction of SO2 to sulfate on dust with coexistence of NH3 and NO2 under different humidity conditions. Atmos. Environ. 208, 133–140. https://doi.org/10.1016/j.atmosenv.2019.04.004.

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. Atmos. Chem. Phys. 19, 3447–3462. https://doi.org/10.5194/acp-19-3447-2019.

Zheng, J., Zhang, L., Che, W., Zheng, Z., Yin, S., 2009. A highly resolved temporal and spatial air pollutant emission inventory for the Pearl River Delta region, China and its uncertainty assessment. Atmos. Environ. 43, 5112–5122. https://doi.org/10.1016/j.atmosenv.2009.04.060.

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. Atmos. Environ. 153, 94–108. https://doi.org/10.1016/j.atmosenv.2017.01.020.