Referee#1

Huang et al. present part 2 on proposing benchmarks for CTM applications in simulating ozone in China. The evaluation criteria is based off prior work by Emery et al. (2017) which may be tailored to the U.S. and Europe and not suitable for China, and the authors propose revised criteria and methodology for simulations focusing on China. The work is generally well written, though I have major concerns regarding some areas for the manuscript which need to be clarified prior to recommending this work for publication.

Response: We appreciate the constructive comments and suggestions. All comments have been carefully considered and addressed in the revised manuscript. The changes made to the main text and supplementary information file are highlighted in yellow. Below is our point-by-point response to each comment, with our responses marked in blue.

Major comments:

1. L59: "... which may not be suitable for China." Could the authors elaborate on why Emery et al.'s criteria are not suitable and the steps the authors propose for revising them? Is it the range of simulated/observed values in China different from other regions? Differences in the chemical regimes controlling ozone in China? Differences in the input data uncertainty? Differences in model tuning targeting different regions?

Response: The reviewer raises a good question regarding the motivation of the study. Several factors necessitate the establishment of a tailored benchmark for model applications specific to China.

First and foremost, ozone concentrations in China are considerably higher than those observed in the United States and have been on a consistent upward trend since 2013, as indicated in the "China Blue Book for the Prevention and Control of Atmospheric Ozone Pollution" (Figure 1, adopted from Zhang et al. 2020). The fourth highest maximum daily 8-hour average (4^{th} MDA8) ozone concentration across 74 major cities in China rose from 189 $\mu g/m^3$ (~95 ppb) in 2013 to 236 $\mu g/m^3$ (~118 ppb) in 2019. In contrast, the 4^{th} MDA8 levels in the United States were recorded at or below 150 $\mu g/m^3$ (~75 ppb) during 2013-2018 (Table 1). A comparative analysis of the 4^{th} MDA8 and the 90^{th} percentile maximum daily 8-hour average (MDA8) ozone concentrations between these 74 Chinese cities from 2013 to 2018 and the United States, which has maintained 1,151 operational ozone monitoring sites since 2010, reveals that both ozone pollution indicators in China are significantly elevated relative to those in the United States. Moreover, while the ozone pollution indicators in China

exhibit an annual increase, the United States has demonstrated overall stability in these metrics. The ozone pollution levels in the 74 cities of China from 2015 to 2019 were comparable to those in the U.S. during the late 1980s, when 196 ozone monitoring sites were in operation since 1980.

Table 1 The 4th MDA8 and 90th percentile MDA8 in 74 Cities of China and 1151 Sites across the United States (2013–2019) (unit: μg/m^3, adopted from China Blue Book for the Prevention and Control of Atmospheric Ozone Pollution 2020)

Statistical methods	Region	2013	2014	2015	2016	2017	2018	2019
Fourth-highest MDA8	74 cities in China	189	201	204	204	227	222	236
	US	142	141	143	144	143	145	—
MDA8-90	74 cities in China	139	145	150	154	167	168	181
	US	122	121	123	122	121	122	—

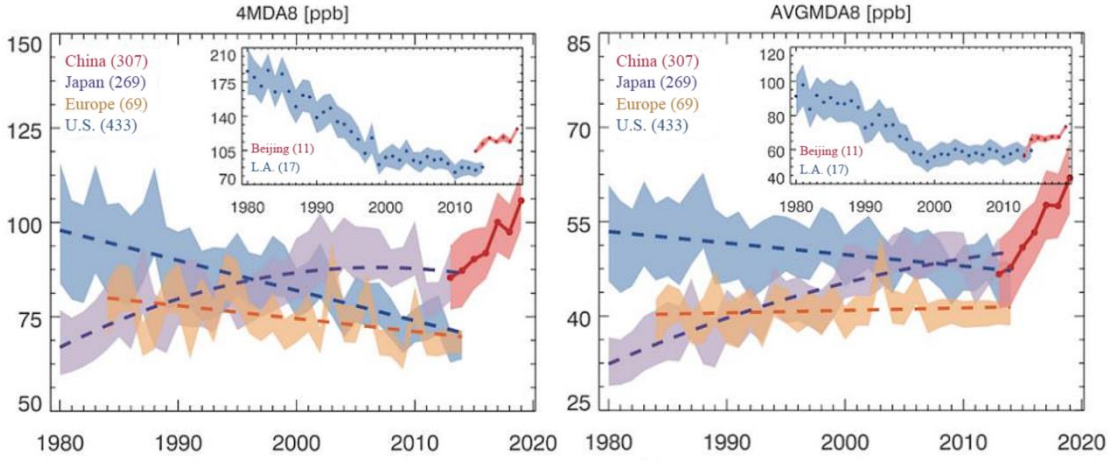

Figure 1 Evolution of urban surface ozone levels in China (red), Japan (purple), Europe (orange), and the United States (blue) from 1980 to 2019 (adopted from Zhang et al. 2020)

Secondly, the contribution of background ozone demonstrates different trends between China and other regions (China Blue Book for the Prevention and Control of Atmospheric Ozone Pollution 2020). According to atmospheric background monitoring data from the World Meteorological Organization, tropospheric ozone background concentrations in the Northern Hemisphere remained relatively stable from 2013 to 2019. Conversely, the background concentration of ozone in China has shown a year-on-year increase, particularly pronounced in urban areas.

Thirdly, the mechanisms underlying ozone formation may differ between China and the United States. However, a direct comparison of these formation regimes proves

challenging, as both countries encompass vast regions with distinct ozone dynamics. Research conducted by Jung et al. (2022) identified notable shifts in the western United States from a NOx-saturated regime to a transition regime (or from a transition regime to a NOx-limited regime), while rural areas, especially in the eastern and southeastern United States, have become increasingly sensitive to VOC emissions. In China, VOC-limited regimes were predominantly observed in the Beijing-Tianjin-Hebei (BTH), Yangtze River Delta (YRD), and Guangdong (GD) regions in 2013 (Zhang et al., 2024). By 2019, a significant transition was noted in the BTH areas from VOC-limited to transition regimes, which was accompanied by a reduction in VOC-limited areas within the YRD and GD.

In summary, the disparities in ozone concentrations, background contributions, and formation mechanisms underscore the necessity for a customized benchmark for model applications in China. Such a benchmark is essential for appropriately addressing the unique challenges posed by ozone pollution within the country.

We have added above descriptions in "1. Introduction" of the revised manuscript (L58-L78):

"Several key factors necessitate the establishment of a tailored benchmark for model applications specific to China. Firstly, ozone concentrations in China have been significantly higher than those in the U.S. and have shown a consistent upward trend (Zhang et al. 2020). For instance, the fourth highest maximum daily 8-hour average (4MDA8) ozone concentration across 74 major cities in China increased from 189 $\mu g/m^3$ (~ 95 ppb) in 2013 to 236 $\mu g/m^3$ in 2019 (~118 ppb), compared to levels at or below 150 $\mu g/m^3$ (~ 75 ppb) in the U.S. during the same period (Table S1). Secondly, background ozone contributions exhibit different trends between China and other regions, with China experiencing a year-on-year increase, especially in urban areas (Zhang et al. 2020). Thirdly, the mechanisms of ozone formation may differ between China and the U.S. However, a direct comparison of these formation regimes proves challenging, as both countries encompass vast regions with distinct ozone dynamics. Jung et al. (2022) identified notable shifts in the western U.S. from a NOx-saturated regime to a transition regime (or from a transition regime to a NOx-limited regime), while rural areas, especially in the eastern and southeastern U.S., have become increasingly sensitive to VOC emissions. In China, VOC-limited regimes were predominantly observed in the Beijing-Tianjin-Hebei (BTH), Yangtze River Delta (YRD), and Guangdong (GD) regions in 2013 (Zhang et al., 2024) whereas in 2019 a significant transition was noted in the BTH areas from VOC-limited to transition

regimes, which was accompanied by a reduction in VOC-limited areas within the YRD and GD. These disparities in ozone concentrations, background contributions, and formation mechanisms underscore the necessity for a customized benchmark for model applications in China, which is essential for appropriately addressing the unique challenges posed by ozone pollution within the country. Therefore, the increasing prevalence of CTM applications in China necessitates specific CTM benchmarks tailored to this region."

2. L79 and Figure 1 pose "WRF-Chem" as a single model, which is not very accurate. WRF-Chem provides an extremely large amount of chemical schemes available (e.g., refer to User's Guide https://repository.library.noaa.gov/view/noaa/14945 Page 14-) ranging from simple RADM2 without aerosols with a dozen species to the MOZART chemical mechanism with hundreds of species, not to mention the different configurations of aerosols, photolysis, and underlying meteorology simulated by WRF. Different papers using different schemes of WRF-Chem are not comparable to each other. Fortunately, the authors do separate the studies by chemical mechanism later in the text (in "Choice of gas-phase chemical mechanism") - I would suggest that this separation is done earlier in the text and in Figure 1 to make it clear that individual chemical mechanisms available in WRF-Chem are evaluated separately and not grouped together. I would request that the supplement data in Table S1 be updated similarly to reflect the chemical mechanism in the WRF-Chem studies.

Response: We acknowledge the reviewer's observation regarding the diverse chemical mechanisms offered by WRF-Chem. Figure 1 and Table S1 (now Table S2) have been modified accordingly to include the information on chemical mechanism utilized by each model application. We have also included clarifications in Section 2.1 regarding the various chemical mechanisms utilized by WRF-Chem to avoid any potential confusion (L106-L108):

"Different configurations could be used even within the same model. For example, WRF-Chem provides different chemical mechanisms, ranging from simple RADM2 without aerosols to the MOZART chemical mechanism with hundreds of species."

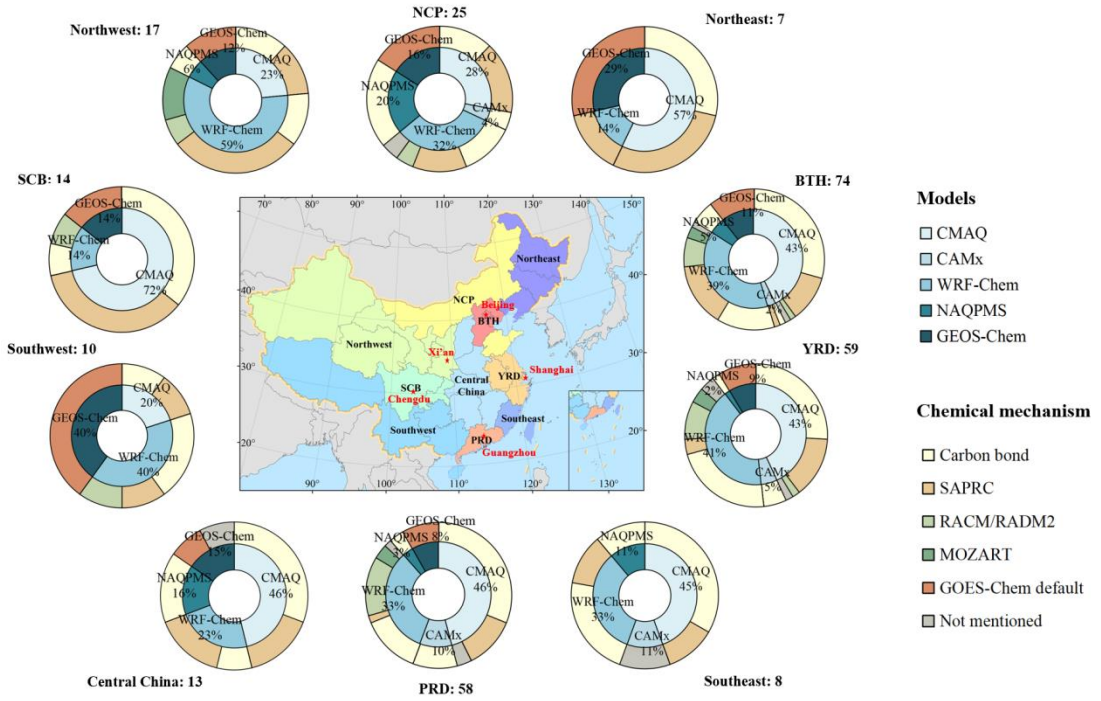

Figure 1 CMAQ modeling domain with definitions of regions used in this study. The surrounding pie charts display the total number of studies for each region (excluding studies for the entire China) and the percentage of different CTMs used. Red stars represent the five cities selected in uncertainty analysis.

3. P92 - the authors convert mixing ratios to µg/m^3 in the analysis. I understand this may be for consistency with the Chinese MEE observational data which is reported in µg/m^3. I recall that there may be a temperature / pressure condition used by China MEE for use in the unit conversion to/from ug/m^3 - can the authors confirm that 273.15K at 101.325 kPa is the one used (and possibly provide a reference)? This would affect the model to obs. comparisons and should be clarified.

Response: Thanks for pointing this out. The conversion factor of 2.14 between ppb and µg/m^3 is derived with an ambient temperature of 273K and a pressure of 101.325kPa, which is referred as the "standard state" in the Chinese Ambient Air Quality Standards (GB3095-2012, https://www.mee.gov.cn/ywgz/fgbz/bz/bzwb/dqhjbh/dqhjzlbz/201203/t20120302_224165.htm, accessed on Dec. 8, 2024). It should be noted that we utilized this conversion factor only for the model performance metrics expressed in absolute concentrations, for example, mean bias (MB), to ensure consistent comparisons across different studies. The decision regarding whether to present the comparison between model outputs and observational data in ppb or µg/m^3, as well as the choice of conversion factor, rests with the authors of the respective studies. We have added clarifications in Section 2.1 of the revised manuscript (L110-L113).

"For consistency, we converted O_3 concentrations (for example, mean bias, root

mean square error) expressed in parts per billion by volume (ppbv) to μg/m³ using a factor of 2.14. This factor of 2.14 refers to the "standard state", i.e., an ambient air temperature of 273.15 K at 101.325 kPa, defined by the Chinese Ambient air quality standards (GB 3095—2012, MEE, 2016)."

4. L122 - "a uniform O_3 concentration of 29 ppb was used as the initial and boundary conditions (BCs)". I have three questions here:

4.1. I assume 29 ppb is at the surface and there is a vertical profile applied to this? What does the vertical profile look like?

Response: The ozone vertical profile is constant, i.e. 29 ppb. This is the default boundary condition file that EPA provides for CMAQ (https://github.com/USEPA/CMAQ/tree/main/PREP/bcon/src/profile, accessed on Dec.8, 2024).

4.2. A 10-day spin-up from uniform initial conditions (and not previously spun-up distributions) of 29 ppb for simulating ozone seems very short. How was this chosen? That is shorter than the mean tropospheric lifetime of ozone (although it may be fine for the PBL) but I have concerns about the effects this may have for free tropospheric ozone and influences from that which may be important for East Asia.

Response: The spin-up period needed for a limited-area photochemical model is not so much dependent on atmospheric lifetime, but rather the time required for atmospheric transport to flush the initial air mass fully out of the domain. This could range from a single day for small urban-centered domains, to a month for continental domains. Among the CMAQ-related articles we reviewed, 58 out of 90 specified their spin-up periods, among which 95% (55 studies) applied a spin-up period less or equal to 10 days (19 studies ≤ 5 days, 36 studies between 5-10 days). Only 3 studies applied a spin-up period more than 10 days. Thus a spin-up of 10 days seems to be a common practice and was therefore adopted in our study. We agree with the reviewer that a longer spin-up period would help reduce the impact of uncertainties associated with the initial conditions, especially if a uniform ozone distribution is specified. We added this point in one of our recommendations in a new Section "3.5 Recommendations for Future Modeling Practices" (see also responses to other comments) (L441-L445):

"4. The majority of model applications reviewed in this study applies a spin-up period of less than or equal to 10 days. However, studies (Hogrefe et al. 2017; Karamachandani et al. 2017) have shown that a commonly used spin-up period of ten days (or a week) might not be sufficient to reduce the effects of initial conditions to less than 1%. Thus, a longer spin-up period, preferably 20 days depending on domain

size, is recommended to mitigate the influence of initial conditions."

4.3. Can the authors confirm that a uniform 29 ppb is used as the boundary conditions? For regional CTMs the transport from outside the domain, which ventilates the simulated region from the boundary conditions, can be quite important for the ozone distribution inside the simulated domain. Why were "realistic" boundary conditions from a global model not used here?

Response: A spatially and temporally uniform ozone concentration of 29 ppb was used to define the initial and boundary conditions in the CMAQ sensitivity simulations conducted in this study. We agree with the reviewer that this is a simplistic assumption and the impact of boundary conditions within the domain can be substantial for ozone. Among the CMAQ application studies collected, 54 of 90 describe the configuration of the initial and boundary conditions while the remaining studies provide no information. Of the 54 papers that do, 65% (35 papers) applied the CMAQ default (i.e. 29 ppb) values. Thus, we decided to apply CMAQ with the default configuration. Our purpose for the ozone uncertainty analysis was to quantify how variability in boundary conditions affect simulated ozone concentrations, and so our approach to mirror how many other studies in China have applied CMAQ is logical given that context. Our results show that boundary condition uncertainty is not especially important for the highest ozone levels that occur throughout the majority of heavily urbanized areas of eastern China. We have added clarifications in the revised manuscript in Section 2.3 (L143-L149):

"The use of a spatially and temporally uniform ozone concentration is a rather simplistic assumption and as we illustrate later the impact of boundary conditions within the domain can range from substantial to minimally impactful. Among the CMAQ application studies collected, 54 of 90 describe the configuration of the initial and boundary conditions and 35 of those applied the CMAQ default profile. Since our purpose for the ozone uncertainty analysis was to quantify how variability in boundary conditions affect simulated ozone concentrations throughout China, we elected to mirror how many of the studies have applied CMAQ."

5. L209... Impact of grid spacing. I would suggest "horizontal resolution" here. The authors claim in L219 that "no clear trend was evident to indicate better model performances as grid spacing decreases." I understand there's further discussion later in this section but this statement is potentially misleading when unqualified without mentioning that it is not controlled for the same model, the same emissions, input data, etc... The authors state at the end of the section that "reducing grid spacing does

not necessarily lead to improved model performance if the input data resolution (i.e., spatial resolution of the emissions) is not correspondingly high or well-matched." In my opinion, such an argument is better phrased as a caution to model configuration instead of a conclusion - if flawed model configurations where the input data resolution is insufficient for the model resolution are analyzed, I would argue it is evident that improved model resolution may not provide the benefits modelers are looking for. At first glance the authors are close to presenting a "dangerous" argument that model resolution provides no benefits then later saying only if the model is configured incorrectly!

Response: Thanks for the nice comment. We have changed "grid spacing" to "horizontal resolution" throughout the manuscript.

Regarding the comment to present as a caution instead of a conclusion, we revised the previous statement as follows:

L247-L248:

"Figure 6 shows the distribution of eight statistical indicators by different horizontal resolutions while ignoring the differences in other model configurations."

L257-L260:

"Therefore, modelers should exercise caution and avoid optimism when configuring their model at finer resolutions as reducing horizontal resolution does not necessarily lead to improved model performance if the input data resolution (i.e., horizontal resolution of the emissions) is insufficient for the model's resolution."

Specific comments:

- L78: GEOS-Chem is not an acronym - see https://geoschem.github.io/narrative.html.

Response: Corrected in the revised manuscript.

-L115: delete "grid". What is the model top height?

Response: Deleted in the revised manuscript. The top of the model goes to 10 hPa. This point is added in the revised manuscript.

- L116: What are the other configuration parameters of the WRF simulation providing the meteorology? e.g., PBL scheme, ...

Response: Configurations of the WRF model are added in Table S6.

- L119: Link for EDGAR is wrong, www.meicmodel.org is written here.

Response: Corrected in the revised manuscript.

- L192: Would be helpful to define BTH, YRD, and PRD here for readers unfamiliar with the region terminology.

Response: Added in the revised manuscript.

- L212 "i.e. GEOS-Chem" - GEOS-Chem can be used regionally. Many studies use

GEOS-Chem nested for China dating back to Y.X. Wang et al. (2004).

Response: Thanks for pointing this out. We changed "i.e. GEOS-Chem" to "i.e. GEOS-Chem applications at global scale)" to avoid confusion (L238-L240):

"Generally, a coarse horizontal resolution (> 50 km) is utilized for global simulations (i.e. GEOS-Chem applications at global scale), while a finer horizontal resolution (< 4km) with nested grids is preferred for regional or city-scale modelling."

References

Jung, J., Choi, Y., Mousavinezhad, S., Kang, D., Park, J., Pouyaei, A., ... & Kim, H. (2022). Changes in the ozone chemical regime over the contiguous United States inferred by the inversion of NOx and VOC emissions using satellite observation. *Atmospheric research*, *270*, 106076.

Zhang, J., Shen, A., Jin, Y., Cui, Y., Xu, Y., Lu, X., ... & Fan, Q. (2024). Evolution of ozone formation regimes during different periods in representative regions of China. *Atmospheric Environment*, *338*, 120830.

Zhang, Y., Zheng, J., Chen, C., et al.: China Blue Book for the Prevention and Control of Atmospheric Ozone Pollution, Science Press, China, ISBN 9787030716644, 2020

Referee#2

Given China's unique pollution characteristics, this study develops benchmarks to assess the accuracy of chemical transport models (CTMs) in simulating ground-level ozone (O_3) pollution in China. A systematic literature review was conducted on 216 studies from 2006 to 2021, covering five widely used CTMs (CMAQ, CAMx, GEOS-Chem, WRF-Chem, and NAQPMS) to establish region-specific benchmarks for O_3. The benchmarks are divided into "goal" values (optimal performance) and "criteria" values (achievable performance) for commonly used model performance evaluation (MPE) metrics, including mean bias (MB), normalized mean bias (NMB), root mean square error (RMSE), normalized mean error (NME), correlation coefficient (R), and index of agreement (IOA).

The study also conducts an uncertainty analysis using the decoupled direct method (DDM) with the CMAQ model, identifying key sources of uncertainty in O_3 predictions. Significant contributors to uncertainty include anthropogenic VOC emissions in urban regions and boundary conditions in rural areas. Spatial and seasonal patterns are noted, with regional differences in uncertainty and model accuracy. These benchmarks and uncertainty insights are intended to guide modelers in China, help standardize CTM applications for ozone and improve the reliability of model-based air quality assessments.

Response: We are grateful for the constructive feedback. All comments have been carefully considered and addressed in the revised manuscript. The changes made to the main text and supplementary information file are highlighted in yellow. Below is our point-by-point response to each comment, with our responses marked in blue.

Specific Comments:

While the study explains the need for China-specific benchmarks, a more detailed comparison with international standards (e.g., differences in precursor emissions and climatic impacts) could reinforce why global benchmarks are unsuitable. The authors may consider adding a more thorough analysis of how China's unique pollution sources and climate contribute to differences in ozone formation and modeling challenges compared to North American or European settings. This could strengthen the rationale for the proposed region-specific benchmarks.

Response: Thanks for the comment. The other reviewer proposed a similar comment. The same response applies here.

First and foremost, ozone concentrations in China are considerably higher than those

observed in the United States and have been on a consistent upward trajectory since 2013, as indicated in the "China Blue Book for the Prevention and Control of Atmospheric Ozone Pollution" (Figure 1, adopted from Zhang et al. 2020). The fourth highest maximum daily 8-hour average (4th MDA8) ozone concentration across 74 major cities in China rose from 189 μg/m^3 (~95 ppb) in 2013 to 236 μg/m^3 (~118 ppb) in 2019. In contrast, the 4th MDA8 levels in the United States were recorded at or below 150 μg/m^3 (~75 ppb) during 2013-2018 (Table 1). A comparative analysis of the 4MDA8 and the 90th percentile maximum daily 8-hour average (MDA8) ozone concentrations between these 74 Chinese cities from 2013 to 2018 and the United States, which has maintained 1,151 operational ozone monitoring sites since 2010, reveals that both ozone pollution indicators in China are significantly elevated relative to those in the United States. Moreover, while the ozone pollution indicators in China exhibit an annual increase, the United States has demonstrated overall stability in these metrics. The ozone pollution levels in the 74 cities of China from 2015 to 2019 were comparable to those in the U.S. during the late 1980s, when 196 ozone monitoring sites were in operation since 1980.

Table 1 4th MDA8 and MDA8-90 in 74 Cities of China and 1151 Sites across the United States (2013–2019) (unit: μg/m^3, adopted from China Blue Book for the Prevention and Control of Atmospheric Ozone Pollution 2020)

Statistical methods	Region	2013	2014	2015	2016	2017	2018	2019
Fourth-highest MDA8	74 cities in China	189	201	204	204	227	222	236
	US	142	141	143	144	143	145	—
MDA8-90	74 cities in China	139	145	150	154	167	168	181
	US	122	121	123	122	121	122	—

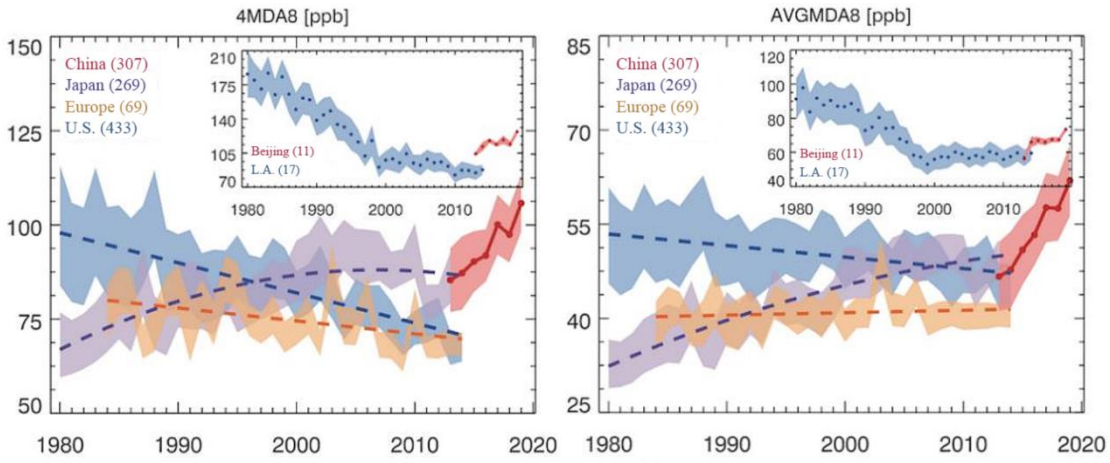

Figure 1 Evolution of urban surface ozone levels in China (red), Japan (purple), Europe (orange), and the United States (blue) from 1980 to 2019 (adopted from

Zhang et al. 2020)

Secondly, the contribution of background ozone demonstrates different trends between China and other regions (China Blue Book for the Prevention and Control of Atmospheric Ozone Pollution 2020). According to atmospheric background monitoring data from the World Meteorological Organization, tropospheric ozone background concentrations in the Northern Hemisphere remained relatively stable from 2013 to 2019. Conversely, the background concentration of ozone in China has shown a year-on-year increase, particularly pronounced in urban areas.

Thirdly, the mechanisms underlying ozone formation may differ between China and the United States. However, a direct comparison of these formation regimes proves challenging, as both countries encompass vast regions with distinct ozone dynamics. Research conducted by Jung et al. (2022) identified notable shifts in the western United States from a NOx-saturated regime to a transition regime (or from a transition regime to a NOx-limited regime), while rural areas, especially in the eastern and southeastern United States, have become increasingly sensitive to VOC emissions. In China, VOC-limited regimes were predominantly observed in the Beijing-Tianjin-Hebei (BTH), Yangtze River Delta (YRD), and Guangdong (GD) regions in 2013 (Zhang et al., 2024). By 2019, a significant transition was noted in the BTH areas from VOC-limited to transition regimes, which was accompanied by a reduction in VOC-limited areas within the YRD and GD.

In summary, the disparities in ozone concentrations, background contributions, formation mechanisms, and emission inventory details underscore the necessity for a customized benchmark for model applications in China. Such a benchmark is essential for appropriately addressing the unique challenges posed by ozone pollution within the country.

We have added above descriptions in “1. Introduction” of the revised manuscript (L58-L78):

“Several key factors necessitate the establishment of a tailored benchmark for model applications specific to China. Firstly, ozone concentrations in China have been significantly higher than those in the U.S. and have shown a consistent upward trend (Zhang et al. 2020). For instance, the fourth highest maximum daily 8-hour average (4^{th} MDA8) ozone concentration across 74 major cities in China increased from 189 $\mu g/m^3$ (~ 95 ppb) in 2013 to 236 $\mu g/m^3$ in 2019 (~118 ppb), compared to levels at or

below 150 μg/m^3 (~ 75 ppb) in the U.S. during the same period (Table S1). Secondly, background ozone contributions exhibit different trends between China and other regions, with China experiencing a year-on-year increase, especially in urban areas (Zhang et al. 2020). Thirdly, the mechanisms of ozone formation may differ between China and the U.S. However, a direct comparison of these formation regimes proves challenging, as both countries encompass vast regions with distinct ozone dynamics. Jung et al. (2022) identified notable shifts in the western U.S. from a NOx-saturated regime to a transition regime (or from a transition regime to a NOx-limited regime), while rural areas, especially in the eastern and southeastern U.S., have become increasingly sensitive to VOC emissions. In China, VOC-limited regimes were predominantly observed in the Beijing-Tianjin-Hebei (BTH), Yangtze River Delta (YRD), and Guangdong (GD) regions in 2013 (Zhang et al., 2024) whereas in 2019 a significant transition was noted in the BTH areas from VOC-limited to transition regimes, which was accompanied by a reduction in VOC-limited areas within the YRD and GD. These disparities in ozone concentrations, background contributions, and formation mechanisms underscore the necessity for a customized benchmark for model applications in China, which is essential for appropriately addressing the unique challenges posed by ozone pollution within the country. Therefore, the increasing prevalence of CTM applications in China necessitates specific CTM benchmarks tailored to this region."

This study could be strengthened further by considering higher-order sensitivities or additional metrics in future analyses, mainly to capture uncertainties in meteorological inputs and chemistry beyond first-order impacts.

Response: Thanks for the insightful suggestion and we agree with the comment. We acknowledge the limitations present in the current study, notably the lack of uncertainties associated with the meteorological inputs and ozone chemistry. As we state in Section 1, page 2, these processes play crucial roles. Among the 216 studies we analyzed, not all provided a performance evaluation of the meteorological model, which we find to be typical in most related literature globally. We have addressed the meteorological issue in our newly added section titled "3.5 Recommendations for Future Modeling Practices" (see also responses to other comments).

As for the higher-order sensitivities, first order sensitivities are effective at describing the predominantly linear ozone response from minor (e.g., 20-30%) changes to influencing variables, which is consistent with the uncertainties addressed here. For such changes, higher-order sensitivities add minor positive or negative adjustments, but for larger changes in highly non-linear environments, higher-order effects become

dominant. However, first and second order sensitivity algorithms incorporated into widely-used models reviewed and applied here address only emissions and chemistry and are not implemented to meteorological variable sensitivity. Several existing studies indicate that the differences in predicted changes in ozone concentration are minimal when employing both first- and second-order sensitivities in comparison to utilizing only first-order sensitivity (Yaluk et al., 2023; Arter and Arunachalam, 2021). For instance, Yaluk et al. (2023) examined the discrepancies between employing first-order sensitivity alone versus using a combination of first- and second-order sensitivities to predict changes in ozone concentration over the Yangtze River Delta. Their findings revealed that, with a 30% variation in NOx levels, the difference in ozone concentration between the two methodologies was a mere 1.4 $\mu g/m^3$ (< 1 ppb).

Additionally, the uncertainties related to chemical processes, for example as quantified by Dunker et al. (2020) using Chemical Process Analysis, were not included in this study due to constraints in time and resources. A comprehensive follow-up study is essential to adequately address this issue, as we state in Section 3.4, page 16.

While the study provides benchmarks, it offers limited guidance on how modelers or policymakers might implement these benchmarks practically. The authors should consider a short section on practical applications of the benchmarks and uncertainty findings. For example, recommend how modelers could adjust their configurations to meet "goal" benchmarks and suggest ways policymakers might use these insights to set air quality standards or prioritize emissions reduction strategies.

Response: Thanks for the insightful comment and suggestions. The purpose of this and our previous related papers is to establish, based on historical published modeling results, a guideline of sorts to provide modelers in China with a contextual reference about how their statistical performance metrics compare against the historical body of work. All models are subject to a certain degree of inherent and irreducible error related to discretization, approximations, parameterizations, etc. Stating that a model exhibits a 20% bias has no real value without context as to whether that is consistent with, or poorer or better than, what is commonly achieved or the best that can be expected. Statistical results outside the proposed benchmarks indicate poor model performance that should be improved in any number of ways so that more reliance can be placed on the model to properly characterize air quality and predict responses to changes in model inputs. There are multiple ways that modeling might be improved, specific to each individual application that involves various contributions

from meteorology, emissions, chemistry, deposition, and boundary conditions. It is not the intention nor capability of the benchmarks to inform about a particular extenuating factor, but rather to simply raise awareness of performance issues. Conversely, a modeler also needs to be careful that meeting a "goal" benchmark does not preclude the model from looking correct as a result of compensating errors. For these reasons, the proposed benchmarks provide contextual references limited to informing the overall model performance evaluation, and as such they alone cannot provide higher-level guidance or information to policy makers on setting air quality standards or emission reduction strategies.

To address the comment in a more general way, we have added a new section "3.5 Recommendations for Future Modeling Practices" in the revised manuscript (see also responses to other comments):

3.5. Recommendations for future modeling practices

The purpose of this and our previous related papers is to establish a guideline that offers modelers in China a contextual reference for evaluating their statistical performance metrics against a historical framework of published modeling results. It is crucial to recognize that all models inherently possess a certain degree of error, which may arise from factors such as discretization, approximations, parameterizations, and etc. Merely stating that a model exhibits a 20% bias has no real value unless contextualized within the framework of historical performance. Without such context, it remains unclear whether this bias aligns with, surpasses, or falls short of commonly achieved standards or optimal expectations. Statistical results outside the proposed benchmarks indicate poor model performance that should be improved so that more reliance can be placed on the model to properly characterize air quality and predict responses to changes in model inputs. Based on the above analysis, we list several recommendations for future modeling practices that might help improve model performance.

1. Meteorology is an essential input to CTMs. Many studies have highlighted the strong responses of ozone to various meteorological variables (Coleman et al., 2013; Lu et al., 2019), including transport patterns, temperature, planetary boundary height, relative humidity, etc. Consequently, it is imperative to conduct a thorough validation of meteorological simulations prior to initiating ozone simulations. The influence of uncertainties associated with simulated meteorological variables on ozone predictions necessitates further exploration.

2. Modelers are encouraged to select the highest feasible spatial resolution that matches the available emission data resolution. Our analysis illustrates that finer resolutions do not invariably lead to enhanced model performance, particularly when

the input data do not possess a correspondingly high resolution. Thus, it is advisable for modelers to engage in sensitivity testing to ascertain the optimal equilibrium between resolution and data quality.

3. The uncertainty analysis reveals substantial contribution of AVOC emissions throughout China. Therefore, it is essential to intensify efforts aimed at enhancing the accuracy of AVOC emissions, focusing on both magnitude and speciation profiles. Additionally, the chemical mechanisms within CTMs should be routinely updated to accommodate emerging species, such as volatile chemical products (VCPs, Yarwood and Tuite, 2024).

4. The majority of model applications reviewed in this study applies a spin-up period of less than or equal to 10 days. However, studies (Hogrefe et al. 2017; Karamachandani et al. 2017) have shown that a commonly used spin-up period of ten days (or a week) might not be sufficient to reduce the effects of initial conditions to less than 1%. Thus, a longer spin-up period, preferably 20 days depending on domain size, is recommended to mitigate the influence of initial conditions.

5. Given the considerable effect of boundary conditions on simulated ozone uncertainties—especially in areas characterized by low precursor emissions—modelers should carefully select and validate boundary conditions. This may involve using multiple global models or observational data to define more accurate initial and boundary conditions.

6. In the context of ozone attainment demonstrations, modelers should place a particular emphasis on the model's performance concerning high and peak ozone values. Merely achieving satisfactory average ozone concentrations may not suffice; it is essential to ensure robust performance in capturing peak ozone levels as well.

Editorial/minor Comments

Lines 17-18: please consider replacing "other factors result in a broad range of simulated O_3 concentration differences from observed values" with "other factors result in a broad range of differences between simulated and observed O_3 concentrations."

Response: Replaced in the revised manuscript.

Line 37: please delete "their" from "their CTM applications" for conciseness.

Response: Deleted in the revised manuscript.

Line 83: please replace "with a time range between 2006 and 2021" with "for studies published between 2006 and 2021."

Response: Replaced in the revised manuscript.

Line 109: please replace "uncertain analysis" with "uncertainty analysis."

Response: Corrected in the revised manuscript.

Line 151: replace "relatively less frequent" with "less frequent by comparison" for clarity.

Response: Replaced in the revised manuscript.

Line 187 (Figure 3): please use the same y-axis scale for most graphs except R and IOA for easy comparison.

Response: Figure modified as suggested.

Line 207 (Figure 4): see the above comment for Figure 3.

Response: Figure modified as suggested.

Line 208 (Figure 5): see the above comment. Please replace "quantile distribution of O_3 NMB values in different seasons" with "quantile distribution of O_3 R and O_3 NMB values in different seasons."

Response: Figure modified as suggested.

Line 231 (Figure 6): see the above comment for Figure 3.

Response: Figure modified as suggested.

Lines 235-236: please fix broken lines.

Response: Fixed in the revised manuscript.

Line 238: should insert a space for Chemistry Mechanism.

Response: Added in the revised manuscript.

Lines 245-246: there is no citation for "SAPRC22 mechanism"

Response: Added in the revised manuscript.

Line 276, more metrics are presented in Table S6 but not mentioned here.

Response: Deleted metrics (MB, ME, and RMSE) from Table S6 (now Table S8) to avoid confusion.

Line 346: not sure how the uncertainty factor of 1.68 is derived.

Response: There was a typo here. It should be "1.97" for anthropogenic VOCs emissions and we have corrected this in the revised manuscript. The uncertainty factor was derived from studies that report uncertainties associated with emission estimates. Take anthropogenic VOC emissions as an example. According to Cheng et al. (2019), the uncertainty associated with anthropogenic VOC emission estimate is +/-68% at 95% confidence, i.e., $2\sigma = 0.68$ where σ is the standard deviation. Same as Dunker et al. (2020), we assumed a lognormal distribution of the emissions. The uncertainty factor represents 2σ of the lognormal uncertainty distribution, which is $e^{2\sigma}$ (=1.97).

Lines 354-355: replace "with a more evenly distributed spatial impact" with "but has a more evenly distributed spatial impact."

Response: Replaced in the revised manuscript.

Line 407+ (references): A link is provided for every reference, but some are visible (in

blue and underlined) while some are not.

Response: Format unified in the revised manuscript.

References

Arter, C. A. and Arunachalam, S.: Assessing the importance of nonlinearity for aircraft emissions' impact on O_3 and $PM_{2.5}$, Sci. Total Environ., 777, 146121, https://doi.org/https://doi.org/10.1016/j.scitotenv.2021.146121, 2021.

Cheng, J., Su, J., Cui, T., Li, X., Dong, X., Sun, F., Yang, Y., Tong, D., Zheng, Y., Li, Y., Li, J., Zhang, Q., and He, K.: Dominant role of emission reduction in 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, 2019.

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

Yaluk, E. A., Wang, Y., Jiang, S., Huang, L., Lu, G., Zhu, A., Bian, J., Xue, J., Du, Y., Chen, N., Manomaiphiboon, K., Chen, H., Zhang, K., and Li, L.: Changes in first- and second-order sensitivities of ozone concentration to its precursors over the Yangtze River Delta region of China due to COVID-19 lockdown: Insights from CMAQ-HDDM modeling study, Atmospheric Environment, 309, 119931, https://doi.org/https://doi.org/10.1016/j.atmosenv.2023.119931, 2023.

Zhang, Y., Zheng, J., Chen, C., et al.: China Blue Book for the Prevention and Control of Atmospheric Ozone Pollution, Science Press, China, ISBN 9787030716644, 2020