Final Author Response

Title: High resolution CMAQ simulations of ozone exceedance events during the Lake Michigan Ozone Study
Author(s): Robert Bradley Pierce et al.
MS No.: egusphere-2023-152
MS type: Research article

We appreciate the constructive comments made by both reviewers and have revised the manuscript to address all the substantive comments and suggestions as well as the majority of the specific comments. This includes new sections that discuss comparisons of the simulations with both insitu and remotely sensed measurements at the LMOS Zion super site. We feel that this significantly strengthens the manuscript and also helps to clarify some of the main points of the manuscript.

Based on the reviewer comments from our companion paper (Oktin et al, 2023), which recommended not publishing the comparisons with the baseline WRF configuration that used observation nudging and indirect soil nudging, we have had to re-do the comparisons between the CMAQ model simulations and measurements using a new WRF baseline simulation without this nudging. This resulted in minor changes to the statistical comparisons that are reflected in the revised manuscript.

The following is the reviewer's comments and our response to each comments in boxes.

Reviewer 1

This manuscript is well written, and the Figures are well organized and clearly convey pertinent information relevant for this assessment. It is very useful to have more modeling assessments that show skill at capturing complex meteorology and ozone chemistry in the Lake Michigan area. Some substantive comments and suggestions are provided to strengthen the paper then smaller suggestions follow.

(1) First, some more clarity around the "EPA run" would be appreciated. EPA does not have a recommended set of options for the application of WRF or CMAQ but tends to use a certain configuration for both of these models that has traditionally performed well over the contiguous U.S. at 12 km grid resolution. The name is also a little confusing in that it might suggest that EPA provided the WRF simulation for this project. Perhaps that simulation could be called "ACM2/PX" or something along that line which identifies it by certain major PBL physics options.

Response: We agree that the use of "EPA run" could imply that EPA provided the WRF simulation for this project and needs clarification. We have changed the name of the baseline run to AP-XM to reflect the use of the ACM2 PBL "AP", the Pleim-Xiu LSM "X' and the Morrison Microphysics "M". Also, based on reviewer comments on our companion paper (Otkin, 2023 https://egusphere.copernicus.org/preprints/2023/egusphere-2023-153/), we are now using a baseline WRF simulation that does not include observation nudging or indirect soil moisture and temperature nudging. This results in slight changes in all the baseline results but does not change the overall conclusions in the manuscript.

(2) Second, it is difficult to fully review this manuscript due to key aspects referencing a paper (Otkin, 2022) that is not in the reference list. A quick Google Scholar search did not turn up this paper. If this has not been published perhaps it could be provided as part of this review.

Response: The reference Otkin, 2022 manuscript was unintentionally omitted from the reference list. While this companion paper was submitted at the same time as our manuscript it was not assigned an editor until after our manuscript became available for discussion. The preprint for this manuscript is now available at https://egusphere.copernicus.org/preprints/2023/egusphere-2023-153/.

(3) Third, some rationale would be appreciated about why the lake breeze timing and strength was only analyzed at Sheboygan and the vertical wind measurements made at Zion were not included as part of this assessment. Both sites are important for ozone impacts in southeast Wisconsin and since there are only 2 sites with vertical wind profiles it seems within scope for a paper focused on that particular process.

Response: We have added a discussion of the Zion wind and temperature profile measurements to Section 3.3 of the manuscript an additional figure (Figure 18 in the revised manuscript) showing comparisons between the observations and the YNT_SSNG and AP-XM simulations.

(4) Fourth, how did the modeling system predict isoprene and isoprene oxidation products? These model simulations did much better at capturing formaldehyde at Sheboygan compared to other assessments that used the same biogenic emissions model.

Response: We have added a section (3.1.3) that includes an evaluation of the YNT_SSNG and AP-XM simulations using the Zion NO2 and isoprene measurements. Figure 6 in the revised manuscript show that both the YNT_SSNG and AP-XM underestimate isoprene relative to the Zion measurements, with the YNT_SSNG isoprene showing larger underestimates than the AP-XM simulation. This is consistent with the use of more realistic, and lower (relative to climatology) Green Vegetation Fraction in the YNT_SSNG simulation (see Figure 2 in Otkin et al, 2023).

(5) Perhaps the authors could provide a little more information about the hemispheric scale RAQMS simulation that was used to provide lateral boundary inflow. That seems to be a potentially big difference from other simulations for this field study (e.g., Baker et al, 2023).

Response: It was actually a global RAQMS chemical analysis that was used for lateral boundary conditions, which has been clarified in the revised manuscript. We have also added two additional references (Song et al, 2008 and Lee et al, 2012). These show that utilizing RAQMS lateral boundary conditions (LBC) for CMAQ continental scale simulations significantly increases upper tropospheric ozone and improves daily maximum surface O3 concentrations (Song et al, 2008) and improves agreement with OMI tropospheric ozone column (Lee et al, 2012) relative to fixed LBC.

(6) Fifth, please provide some more details in the methods section about the version of CMAQ applied for this assessment and which version of the Carbon Bond chemical mechanism (CB6r3, etc.) was used for these model simulations.

Response: We have revised the manuscript to indicate the version of CMAQ used (v5.2.1) including new references (Appel et al, 2017 and US EPA, 2018) and clarified that the Carbon Bond 6 mechanism used was revision 3 (CB6r3)

(7) Was the GLSEA product used as part of the EPA simulation? Better resolved lake temperatures seem important for capturing thermal gradients at the lake/land interface and if only one of the simulations used that product that could be as much or more important than the different physics options selected.

Response: No, the GLSEA product was not used in the EPA simulation. Surprisingly, our companion paper (Figure 3, Otkin et al, 2023) shows that using high resolution GLSEA SST does not have a significant impact on 2-m Temperature, water vapor mixing ratio, or 10-m windspeed bias or RMSE relative to the YNT physics without the GLSEA SST.

More specific comments follow.

(8) Lines 62-63. Is there a reference for 1-10 km or is this the opinion of the authors? This paper does not conclusively show that 12 km can not capture high O3 in this region that would be relevant for SIP demonstrations.

Response: We have clarified that this is our opinion.

(9) Line 94. Please define RMSE or provide a reference

Response: RMSE has been defined as root-mean-square-error

(10) Figure 2. Please comment on the over-prediction (orange) in upper right panel in Figure 2 that looks to be in Chicago.

Response: We have added “with the exception of two AQS sites in North Chicago were the YNT_SSNG simulations shows overestimates of 4-8 ppbv in 8-h maximum ozone” in the revised manuscript

(11) Figure 3. It would be helpful to replace the Julian day scale with the calendar dates (same comment for Figures 4, 5, and 6).

Response: Done for Figures 3, 4, 5, 6, 7, and 8 in revised manuscript

(12) Figure 3. YNT_SSNG NO2 is greater than EPA. Does this increase in NO2 translate to more O3 in the YNT_SSNG model? If so, how does that impact contemporaneous O3 performance? Is there a better way to show this in figures or explain in text?

Response: Since the baseline simulation used in the revised manuscript uses a different AP-XM configuration (no obs nudging or indirect soil nudging) we now find that both the AP-XM and YNT_SSNG simulations overestimate NO2 at the EPA Trailer, although the YNT_SSNG has larger overestimates (0.68 verses 0.2 ppbv). These differences are in reductions in ozone in the YNT_SSNG (0.07 verses 1.76 ppbv, not shown) which are often associated with lower predicted ozone at night. The following discussion has been added to the revised manuscript: "The larger high biases in NO2 and reduced low biases in HCHO in the YNT_SSNG simulation leads to significant reductions in high biases in ozone in the YNT_SSNG compared to the AP-XM simulation (0.07 verses 1.76 ppbv, not shown) and may be due to more nighttime ozone titration in the YNT_SSNG simulation.

(13) Fig 5 and 6 not so easy to see, consider shortening the x-axis to include only the A B and C time periods the authors wanted to highlight. For Figures 5 and 6 there does not seem to be a notable difference in performance. Some days one configuration is better and then the other is better on other days. Some days both are poor. I think the aggregated metrics might be over-emphasized in the text. Do they include nighttime minimums that are not SIP relevant?

Response: Figures 5 and 6 (7 and 8 in the revised manuscript) are meant to show the entire simulation time period and serve as a transition into the more focused discussion of the composite results during ozone events A, B, and C. We feel that the aggregated metrics (which have been updated to reflect the results of the new baseline simulation) provide a valuable summary of the overall performance of the simulations at these key monitoring sites. The statistics do include nighttime minimums, which we agree are not relevant for 8-hr maximum ozone calculations.

(14) Line 211. "While the overall hourly ozone statistics at Sheboygan KA and Chiwaukee Prairie are relatively similar between the EPA and YNT_SSNG simulations at these sites, the simulations during high ozone events are quite different. This is illustrated by looking at composite statistics during events A, B, and C." How are they different? The subsequent figures and paragraphs generally seem to say the two sims are more similar than different.

Response: The updated (with the new baseline AP-XM) composite wind rose statistics show that the AP-XM simulation does a better job of representing the percentage of high ozone (above 60ppbv) than the YNT_SSNG at both Sheboygan KA and Chiwaukee Prairie. These differences are larger than found with the original baseline, particulary at Chiwaukee Prairie. The discussion has been updated in the revised manuscript.

(15) Lines 311. "Given the relatively high precision of GeoTASO NO2 compared to the column amounts observed during high ozone events A, B and C, we conclude that the high bias in NO2 columns in the EPA simulation is meaningful." Meaningful how? Does this translate to differences in O3 production and therefore performance in the model?

Response: We've changed "meaningful" to "significant" since we are referring to the magnitude of the bias relative to the precision of the GeoTASO NO2 measurements.

(16) Figure 12. How does the penetration compare for B and C?

Response: To limit the length of the manuscript we have limited section 3.3 to just the June 2nd ozone event since this was considered a "classic" lake breeze event.

(17) Line 389. "The YNT_SSNG simulation captures the thermal structure of the nocturnal boundary layer and timing of the maritime boundary layer but underestimates the surface temperatures (by roughly how much? Please quantify or provide model perf statistics here) within the convective boundary layer. The YNT_SSNG simulation captures the vertical structure of the lake breeze wind speed and direction, but the timing of the switch in wind direction is about 3 hours too early. In contrast, the EPA simulation shows no thermodynamic signature of a nocturnal or marine boundary layer and underestimates (by how much? quantify or provide model perf statistics here) the sharp change in the observed windspeed and direction."

Response: This section has been revised to quantify the differences between the observed and simulated temperatures, wind direction, and wind speed. The revised discussion is: "The YNT_SSNG simulation captures the thermal structure of the nocturnal boundary layer (temperature differences are less than 2^{o}C below 100m) and timing of the arrival of the maritime boundary layer but underestimates the near surface (below 200m) convective boundary layer temperatures by up to 10^{o}C at 15 UTC. The AP-XM simulation shows significant (temperature differences are greater than 5$_{o}$C below 100m) overestimates of the nocturnal boundary layer temperatures and shows a gradual warming of temperatures below 200m after 15Z, resulting in large (greater than 7^{o}C) overestimates in temperatures and no evidence of the cooler lake breeze. Both simulations underestimate the observed increase in wind speed prior to the arrival of the lake breeze by ~2m/s. The YNT_SSNG simulation shows a more rapid shift in wind direction associated with the arrival of the lake breeze than the AP-XM simulation, but the timing of the switch in wind direction is about 3 hours too early in the YNT_SSNG simulation. This results in errors in wind speeds of up to 5m/s near 200m in the YNT_SSNG simulation prior to the observed reduction in wind speed at 15 UTC. The observed depth of the wind shift is underestimated in both simulations, but the YNT_SSN simulation does a better job of capturing the vertical extent of the wind shift and reduction in wind speed above 200m. This is most evident above 400m where the AP-XM wind speeds are underestimated by up to 5m/s." We have also added a more detailed discussion of the comparison between the simulations and the Microwave temperatures and Sodar wind measurements at Zion IL (new figure 18).

(18) REFERENCES

Baker, Kirk R; Liljegren, Jennifer; Valin, Lukas; Judd, Laura; Szykman, Jim; Millet, Dylan B; Czarnetzki, Alan; Whitehill, Andrew; Murphy, Ben; Stanier, Charles; Photochemical model representation of ozone and precursors during the 2017 Lake Michigan ozone study (LMOS), Atmospheric Environment, 293,,119465,2023

Response: This reference has been added and referred to in the conclusions

Reviewer 2

I would rate this paper as “minor revisions” – I have a lot of comments, but reading them through, they are along the lines of clarifications needed in the text, some additional references worth quoting since they are recent papers on the same topic in the same region that the authors have missed, the need for some caveats on some of the conclusions, etc., as opposed to serious methodological issues. I also made some suggestions for additional possible causes for model differences that might be worth investigating (a check on the inputs for ozone deposition velocity and whether the model changes between the two versions might have affected the simulated deposition velocities, for example). Definitely worth publishing, subject to the clarifications, etc., below in my specific comments.

Specific Comments:

Introduction:

(1) Line 53, list of references: A few on the Canadian side of the border worth referencing, since they are on the same topic, and give insight into model processes (including two studies at 2.5km resolution):

Stroud et al., 2020: Chemical analysis of surface-level ozone exceedances during the 2015 Pan American games, Atmosphere, 11(6), 572; https://doi.org/10.3390/atmos11060572 , identifies the updraft region of lake breeze fronts as the region where ozone production is occurring, and a transition from VOC to NOx-sensitive O3 formation within that region. Lake Ontario water temperature and strength of the large-scale flow shown to be critical in timing of the observed and modelled LBFs.

Brook et al, 2013: https://acp.copernicus.org/articles/13/10461/2013/ Overview of the main findings of the BAQSMet2007 study, including interactions of LBFs and ozone formation.

Makar et al, 2010 : Mass tracking for chemical analysis: the causes of ozone formation in southern Ontario during BAQS-Met 2007, https://acp.copernicus.org/articles/10/11151/2010/ 2.5km resolution analysis, lake breeze, recirculation and other effects studied.

He et al., 2011 : https://acp.copernicus.org/articles/11/2569/2011/ observation-based analysis of O3.

Response: References added to introduction

Methods:

(2) Line 80: please mention the vertical resolution of the model for the first few layers up from the surface, as part of this discussion. Later, in the conclusions, mention is made of a disconnect between the ACM2 physics and the modelling setup for YNT_SSNG: please provide some background information, either in the Introduction or Methods sections, on this disconnect and why it might impact model results.

Response: We have added the following sentence to the revised manuscript: "Both simulations have the same vertical resolution throughout, with 6 model layers below 200m, 4 model layers below 100m, and the lowest three layers at ~9, 27, and 55m above ground level." We have also revised the discussion of the AP-XM (previously EPA) configuration to: "The AP-XM simulation employs the Morrison microphysics (Morrison et al. 2005), ACM2 PBL (Pleim 2007), and the Pleim-Xu LSM (Gilliam and Pleim 2010; Xiu and Pleim, 2001) parameterization schemes, **which are the same schemes used within CMAQ and is considered our baseline meteorological simulation."** to provide some more background on the differences between the simulations.

(3) Line 83: mention is made of the Morrison, Thompson microphysics as well as the Kain-Fritsch cumulus parameterization. Explicit microphysics schemes are usually thought of as applying only at high resolution, while cumulus parameterizations such as KF are usually used at lower resolutions. Which type of cloud approach was applied at which resolution, in the authors' setup of the nested model, and why? Is this the disconnect mentioned by the authors I mention above with reference to line 80 and later in the text?

Response: We have clarified that the Kain-Fritsch cumulus scheme is only used on the outer two domains, and explicit convection on the innermost domain for both simulations. The KF cumulus parameterization is for representing sub-gridscale convection while the Morrison (AP-XM) and Thompson (YNT_SSNG) microphysics schemes are for representing resolved cloud processes. These microphysics schemes may perform better at high resolution, but they are also used at coarser (12km and 4km) resolution simulations. This is not the disconnect between the PBL and surface parameterizations used in the meteorological (WRF) and chemical (CMAQ) simulations that the reviewer mentioned above.

(4) Line 93: The authors have discussed the model setup, but not the input analyses used to drive it, or how often these were updated. That is, they mention the EPA meteorology, and the optimized WRF configuration: were both of these meteorological models driven from the same meteorological analysis or from different analyses? Depending on the source of driving met information used for those two sources of information, different results for the meteorology might occur, and may explain some of the differences between the two CMAQ simulations (if they were different – e.g. different analysis hours, different analyses from different sources). Please give some more information on the source of analysis information used to generate the meteorology for the two simulations – was it the same information or different – and how often the meteorological models were updated using a new analysis. This is a key part of the meteorological setup for this sort of experiment; needs to be included in Methods.

Response: A detailed description of the WRF meteorological modeling set up is provided in our companion paper (Otkin, et al, 2023, https://egusphere.copernicus.org/preprints/2023/egusphere-2023-153/)) which was unintentionally omitted from the reference list. While this companion paper was submitted at the same time as our manuscript, it was not assigned an editor until after our manuscript became available for discussion. Both the AP-XM and YNT_SSNG WRF runs were constrained with 6-hourly, 0.25-degree GFS Final reanalyses. We have added this to the revised manuscript

(5) Line 95: I didn't see a description on whether either the CMAQ or meteorological models include a parameterization for horizontal diffusion when operating at high resolution. Is this process included (please mention if or if not, in the text)?

Response: Horizontal diffusion is included at all resolutions.

(6) Line 108: the paper would benefit from a figure showing the nested domains used in the simulations, please add.

Response: The nested domain figure is included in our companion paper (Otkin et al, 2023). This is noted in the revised manuscript.

(7) Line 114: the information that the 4 km emissions were interpolated and downscaled by 1/9 (I assume that this was actually $1.3^2/4^2$, or about 1/9.47?), rather than generated by SMOKE using area source polygons gridded directly to higher resolution spatial allocation was a bit surprising. The use of 4km data at 1.3 km will negate some of the advantages of going to 1.3km resolution. Why was this approach taken? Had 1.3 km^2 resolution emissions been generated directly but resulted in poor performance, or is this a stage still to be carried out? The text needs to note at this point that the linear interpolation and scaling of 4km emissions to 1.3 km resolution will prevent the resolution from being able to capture, e.g., area sources of smaller than a 4x4 km grid cell, effectively diluting the 1.3km emitted pollutant mass on input to the 1.3km model. This could have a substantial impact on model performance, and needs to acknowledgement in the Methods and Conclusions sections (ideally, rerunning the model with emissions at the higher resolution would be a better approach). Were the biogenic area emissions also downscaled in this fashion?

Response: The "1.3km" run is actually "1.3333km" so the 1/9 downscaling is appropriate. This is clarified in the revised manuscript. We don't have access to SMOKE and were not able to generate the 1.3km emissions from the area source polygons so we had to compromise and downscale the 4km gridded emissions. We have acknowledged this in the revised manuscript: "We acknowledge that the use of downscaled 4km emissions will degrade the performance of the 1.3km simulations but generating 1.3km area emissions from the Sparse Matrix Operator Kernel Emissions (SMOKE) programs was beyond the scope of this project."

1. Results

(8) Line 155: The authors have stated that "the biases and RMSE in 8-h maximum ozone are generally smaller by 2ppbv in the YNT_SSNG simulation". This needs to be corrected - I think what's happening is more nuanced and should be qualified a bit by the authors. What I see in Figure 1's mean bias panel is that the concentration biases in the YNT_SSNG parameterization are uniformly more positive than the EPA simulation, aside from the 80—90 ppbv values where the EPA simulation has the least negative bias. That is, the YNT_SSNG concentrations are likely in general higher, across the entire < 80 ppbv concentration range. For example, relative to the zero bias line, the YNT_SSNG results for 20 – 30 ppbv and for 30 – 40 ppbv are actually worse (higher positive bias) than for the EPA model. Similarly, the EPA model has the lower magnitude bias for the upper end concentrations between 80 and 90 ppbv. The general upward shift in bias towards more positive numbers in the YNT_SSNG model versus the EPA model made me wonder if this offset is associated with a different deposition algorithm being used between the two models, or if the same deposition algorithm is used, that the differences in GVF, soil temperature and moisture may have resulted in different deposition velocities for O3. That kind of difference in bias may be an indication of a different loss flux for O3 between the two models.

Response: Based on reviewer comments on our companion paper (Otkin, et al, 2023) we are now using the AP-XM WRF simulation without observation nudging. This results in slight changes to the comparisons presented in the manuscript, which has been revised to reflect these differences. For figure 1, the nuances raised by the reviewer are not longer apparent since the two simulations perform very similar between 40-80 ppbv for 8-hour averaged ozone. The discussion in the revised manuscript has been revised to reflect this. The statement that "the biases and RMSE in 8-h maximum ozone are generally smaller by 2ppbv in the YNT_SSNG simulation" refers to Figure 2 (spatial distribution of 8-h maximum ozone) and still holds true. Generally, the YNT_SSNG simulation is one color scale closer to zero than the AP-XM, except for northern Chicago, which reviewer one noted and has been acknowledged in the revised manuscript: "When compared on a site-by-site basis, the biases and RMSE in 8-h maximum ozone are generally smaller by more than 2 ppbv in the YNT_SSNG simulation with the exception of two AQS sites in North Chicago were the YNT_SSNG simulations shows overestimates of 4-8 ppbv in 8-h maximum ozone."

(9) Similarly, the RMSE would be better described as "better for the YNT_SSNG model *between 40 and 80 ppbv*", since between 20 and 40 ppbv and between 80 and 90 ppbv, the EPA model has a lower RMSE than the YNT_SSNG model.

So – mention the range of concentrations where each model is outperforming the other, rather than the blanket statement currently in use.

Response: This discussion has been updated in the revised manuscript to account for the new AP-XM baseline CMAQ experiment. We have tried to be more explicit in indicating the range of concentrations in the revised discussion.

(10) Figure 2: agree, YNT_SSNG definitely looks better on average for both mean bias and RMSE in these maps (which will give the average performance across all concentration ranges).

Response: This is the figure where we refer to the YNT_SSNG generally being smaller by 2ppbv (see above)

(11) Line 188: maybe that should be "*hence* better representation of biogenic VOCs". Another possibility here: were the same land use category maps used for deposition in each of the models, or were they also updated / different between the two simulations? Land use type, Leaf-area-index also have a key impact on deposition velocity – if those have changed, they could well account for the offset in ozone bias noted above. Its worth adding a few lines to the methodology regarding whether any of the inputs to the model's gas-phase dry deposition code have changed as a result of the different configurations.

Response: The revised manuscript includes a comment: "Differences in near surface wind speed and GVF will also impact deposition velocities in the CMAQ simulations." in the methods section and notes that the improved ozone bias in the YNT_SSNG simulation "may be due to the use a more realistic, and lower (relative to climatology) Green Vegetation Fraction (see Figure 2 in Otkin et al, 2023) in the YNT_SSNG simulation which would tend to reduce ozone deposition velocities and increase ozone concentrations (Ran et al, 2016).

(12) Figure 4: An aside: I actually thought that *both* of these are pretty good, for HCHO, since it is so dependent on the secondary chemistry as well as the primary VOC emissions.

(13) Line 208: Discussion on Figures 5 and 6: its rather hard to tell the two panels in each of these figures apart at a glance; more useful information is provided by the summary values in r, bias and rmse on the panels. Aside from establishing that the two simulations do have differing performance at two different sites, I'm not sure if they add to the paper – either remove them, put them in an SI, or try zooming in on the period encompassing start of A through end of C to better show the differences in the region of interest (start at Julian day 152, end at 169).

Response: We feel that these figures add value to the manuscript since they show comparisons between the simulations and observations for the whole modeling period. They also serve as a transition towards the discussion of the model results during the three high ozone periods.

(14) Also, Figure 5, 6 discussion. Can the authors make some statement on whether the model results are different from each other in a statistically significant way (e.g. 90% confidence limits calculated for the two model runs and compared, etc.). Can something quantitative be stated regarding the model differences not be due to chance, etc.?

Response: We conducted Student T-tests between the AP-XM and YNT_SSNG simulations and found statistically significant differences in mean ozone concentration at Sheboygan KA but not at Chiwaukee Prairie. The following has been added to the revised manuscript: "Student T-Tests between the AP-XM and YNT_SSNG simulations at each site show that the simulations have statistically significant differences (99% confidence level) in mean ozone concentration at Sheboygan KA but not at Chiwaukee Prairie."

(15) Line 233: differences in the meteorology had me wondering whether the two sets of driving meteorology were generated from different analyses, see comment above.

Response: Both simulations used the same GFS analyses for lateral boundary conditions and analysis nudging.

(16) Line 258: "provide an overall comparison of the ". They provide a comparison at two locations. Are there other meteorological stations in the region which can be used to provide an overall evaluation table of meteorological performance? That would be better than a two station comparison…. Figure 1 provides part of this, I think and is a better indication of model performance than Figure 7.

Response: We meant at these two stations. The revised text clarifies this. Our companion paper, Otkin et al, 2023 provides overall evaluation for all met stations within the 1.3km domain.

(17) Figure 9: There needs to be some explanation or at least speculation of why the models both are showing much greater variability in the simulated wind direction relative to the observed wind direction, for wind directions < 200 degrees. I'm wondering for example what the wind magnitudes were here; the Figure 7 wind roses give frequency but not wind speed… for example, are the winds "light and variable" when the winds are coming from the < 200 degree directions, hence possibly explaining why the direction has such higher variability? i.e. are the models having issues resolving wind direction at low wind speeds?

Response: We conducted additional analysis using 1m/s and 3m/s thresholds for the observed wind speeds to see whether the median westerly bias and large variability in wind direction when the observed winds had an easterly component at Chiwaukee Prairie was due to the model's inability to predict wind direction at low wind-speed. Results for no filter and a 1 m/s filter showed similar results. The 3 m/s filter resulted in loss of up to 75% of the data within each observed wind direction bin and the statistics showed more variation between wind direction bins. We've added a speculative comment about errors in the timing of the arrival of the lake breeze at Chiwaukee Prairie as a possible reason for the large variability and westerly bias. Specifically, we've added: "This could be associated with errors in the timing of the arrival of the lake breeze." Also, to partially address the reviewer's concerns we have used the 1 m/s threshold filter in the revised manuscript. The discussion of these figures has been updated and now includes the following statement: "The 1 m/s threshold was included to reduce the impact of light and variable winds at these sites."

(18) Lines 313 to 315: Looking at Figure 11, the YNT_SSNG has a larger number of points clustering off of the 1:1 line than the EPA simulation, in the upper end of the range. The figures need some addition to show the range of variability that might be expected in the observations, to better make the point regarding HCHO results for the two models being a bit less hard to

quantify given the uncertainties in the obs. Plotting the obs expected range of variability (e.g. factor of 2 lines if the obs are as much as a factor of 2 off) on the figures would make a better case, and I suggest the authors do this – I'm not convinced by the text and looking at the figures that the HCHO values are sufficiently uncertain that both EPA and YNT_SSNG are doing equally well. I suspect that Figure 11 has a large number of points in the red zone outside of the expected error of the instrument for both simulations.

Response: The observed HCHO values are generally less than the precision of the measurement ($10x10^{15}$ mol/cm^2) given the 1km co-adding of the radiances. Since this is the case, the differences in the biases between the two simulations are much smaller than the precision of the measurement and we can't use the GeoTASO HCHO measurements to draw conclusions about the simulations. We have added +/- precision estimates to the figures to make this point more clearly and added the following discussion to the revised manuscript: "Given the relatively high precision of GeoTASO NO2 compared to the column amounts observed during high ozone events A, B and C, we conclude that the high bias in NO2 columns in the AP-XM simulation is significant, with more AP-XM NO2 columns found outside the +/- 0.5 x 1015 mol/cm2 precision range then found in the YNT_SSNG simulation. We have less confidence in the significance of the differences between the YNT_SSNG and AP-XM HCHO columns relative to the GeoTASO retrievals since the observed HCHO columns are on the order of the precision of the instrument (10 x 1015 mol/cm2) and the biases in column HCHO simulations are both mostly less than the GeoTASO precision during high ozone events A, B, and C. "

(19) Figures 10 and 11 lack colour bar scales for the number of points in a given cell, and the dividing line between points clustered by cell and individual points obs/model pairs as dots is not clear. Colour bar scales need to be added to the two figures.

Response: Each obs/model pair is indicated by a dot. The "clustered by cell" appearance is because there are a large number of points in the bins used to count the pairs. Color bars have been added to the figures in the revised manuscript

Section 3.3:

(20) General comment: this section was more convincing – for this particular case, the YNT_SSNG model was doing better than the EPA model.

(21) Line 330: Why was this particular time chosen? One thing I've seen from high resolution AQ model runs in the great lakes area in past research (see above references) is that the timing of the front arrival can be off. In that respect, with regards to the EPA simulation – did the EPA front ever reach as far inland as the YNT front during the event – or was it always close to the shoreline?

Response: June 2, 2017 was chosen since it was considered a classic lake breeze driven event. The AP-XM front remained near the shoreline during this event.

(22) Line 334, left panel in Figure 12: do the authors have access to the original imagery so they could alter the grey scale to a colour scale making the differences a bit easier to see in the image?

Response: The grey scale has been enhanced to make the differences easier to see in the image.

(23) Line 339: However, it is also clear from Figure 12 that the ozone at several of the stations is more accurately captured by the EPA model setup (e.g. the northernmost station, two stations near the "-68" and "43" on the figure, the two stations near Chiwaukee); worth adding a caveat to that effect. YNT_SSNG gets the southern group of three stations better, however. A scatterplot of O3 for across all stations at this time might help to make a better case for YNT_SSNG here, too. I agree that the EPA simulation is not getting the inland penetration right and YNT_SSNG is doing a better job, however.

Response: The following discussion has been added to the revised manuscript "While the YNT_SSNG simulation shows deeper penetration of the lake breeze circulation it also leads to somewhat lower surface ozone concentrations near the shoreline leading to underestimates in the observed ozone concentrations at this time."

(24) Line 358: agree, good case that YNT_SSNG is doing better for column NO2. Figure 13: both simulations appear to be missing some of the column NO2 over the lake on the eastern side of the observations. EPA could arguably be doing better there, though it shows an erroneous peak to the SE not in the obs. Any thoughts on what's happening further out over the lake and why the models are missing it? Upper atmosphere NO2 coming from somewhere else?

Response: The GeoTASO NO2 and HCHO both appear to have a solar zenith angle dependence with higher values on the eastern edge of the raster pattern that was sampled later than the western edge but we don't know whether this is realistic or not. We have added the following sentence to the revised manuscript: "The observed NO2 columns also show enhancements over the lake on the eastern part of the GeoTASO raster pattern that are best captured by the AP-XM simulation."

(25) Lines 367 to 370: I don't follow the explanation here - a few more sentences of description of how/why the drift relative to the baseline happens are needed. EPA simulation seems to be closer to obs in this eastern part of the domain, but authors are saying we don't trust the data there: they need to explain reasoning why the observation results are less trustworthy here better, preferably with some references. Note that the NO2 columns also show higher values on the east side, which would argue a real effect as opposed to artifact.

Response: This statement was included based on discussions with the GeoTASO instrument PI. We have removed the reference to the drift in the baseline. The increased NO2 and HCHO on the eastern side of the raster has a very banded structure that is most likely an artifact.

(26) Figure 15: YNT_SSNG definitely better than EPA for temperature and windspeed here. Both seem to be off for the wind direction (centre panels). I note that this corresponds to an underprediction by both models of the wind speed for the central region at about 10 GMT; see my earlier note regarding Figure 9 and wind direction uncertainties being exacerbated at low wind speeds.

Response: The easterly bias in wind direction during this event is consistent with the overall statistics during ozone events A, B, and C during westerly winds as shown in Figure 11 in the revised manuscript. This is now commented on in the revised manuscript.

Conclusions:

(27) Some further summary comments, based on the Figures (authors should note if these take-away messages differ from their intent and modify the text if so):

Response: Thanks for these comments (28-33) on how the reviewer considers each of the figures in terms of the two simulations. We have taken them into consideration in the revised conclusions.

(28) Figures and text of section 3.3 make the best case that YNT_SSNG is doing a better job.

(29) Figure 1 suggests that YNT_SSNG biases have been shifted more positive uniformly across the concentration bins, RMSE are better for YNG_SSNG for 40-80 ppbv range, worse for smaller and higher concentration ranges, than EPA. Figure 2: YNT_SSNG doing a better job.

Figure 3: toss-up.

Figure 4: YNT_SSNG better.

(30) Figure 5 & 6: together, a toss-up (and two stations are not necessarily meaningful - I wonder what a time series of the average of all stations for the obs and the two models would look like?).

Figure 7: hard to say which is better.

Figure 8: lots of variability, but EPA seems to be doing better for the highest O3 events.

(31) Figure 9: models have so much variability in the wind direction at low angles that its very hard to see which is better. This needs to be combined with wind speeds somehow.

Response: See response to comment 17

(32) Figures 10 and 11 could use lines showing the upper and lower range associated with the satellite observations. I'm not convinced by the authors's statement that the HCHO values are sufficiently suspect that both models performance relative to those observations is similar. Show it on the figures.

Response: See response to comment 18

(33) Figure 12 - 15: EPA doing better for some stations - but YNT_SSNG doing better for getting the lake breeze penetration distance and the shape of the front, etc.

Other comments on the Conclusions:

(34) Lines 417-418: Figure 11 seems to indicate that column HCHO biases are in places more negative in YNT_SSNG than EPA? See earlier comment about adding some lines indicating the observation variability on that Figure.

Response: Lines showing precision have been added to the figure

(35) Line 422: biogenic emissions differ, don't they? Add a qualifier here, emissions are not necessarily identical. Re: differences being linked to differences in horizontal and vertical transport: it might be better to day that they are linked to differences in meteorology, but not necessarily just transport. For example, are the cloud fields the same for the two simulations (aside from aqueous chemistry effects, whether or not convection has kicked off may also influence photolysis rates).

Response: We have added a discussion of the impacts of the use of more realistic GVF in the YNG_SSNG simulation and it's impact on biogenic isoprene emissions in the revised manuscript.

(36) Line 437: mismatch between ACM2 vertical diffusion and YNT_SSNG mentioned here for the first time – see my note above – some prior discussion of this issue and what the mismatch consists of needs to appear earlier in the text.

Response: We have added a discussion of why the AP-XM is chosen as the baseline simulation (because it uses the same ACM2 and P-X LSM as CMAQ) in the methods section.

(37) Line 443-444: The authors statement here is too broad – the YNT_SSNG shows improved timing and extent of the lake breeze front for one high ozone episode; that of section 3.3. The statement should be rewritten as, e.g., "in the YNT_SSNG simulation *for the June 2, 2017 ozone episode*". The abstract had it right, in that respect; this line in the conclusions needs to match the abstract.

Response: This statement has been revised to restrict it to the June 2, ozone episode.

(38) The authors mention that going to two-way coupling is worth looking at for this problem. I would agree, with some caveats and advice (since our group has gone there). One of the first thing that happens with full coupling is that all of the cloud fields move – since the activation of clouds is linked to the locations where hygroscopic aerosols are located. However, these movements may not be *significant*, since individual convective cells kicking off or not in adjacent grid cells give a local large difference in water and chemistry… but don't necessarily affect things like the timing of frontal passage: you need to look at multiday averages to see

whether the feedback effects have a net effect – single case studies are much more difficult to show a significant effect. Use of confidence ratios (cf Figures 8, 12, 13 of Makar et al, 2021 https://acp.copernicus.org/articles/21/12291/2021/) are one way to estimate the significance of changes like these. Calculating similar confidence ratios might be another way to see the extent to which the YNT_SSNG and EPA simulations are significantly different, in the current study.

Response: We have replaced "two-way coupled" to "more tightly coupled" WRF-CMAQ model, since WRF-CMAQ does not actually have feedbacks. Instead, it allows for higher frequency of input meteorology.