Please see below our point-by-point response (in blue) to both reviewers' comments (in black). Quoted text from the revised manuscript is *in italic*.

The most updated (20 August 2024) US EPA AQS surface measurements are included in this revision. Color-blind friendly color schemes are applied to figures, referring to Crameri et al. (2020, https://doi.org/10.1038/s41467-020-19160-7) and other references in ACP's submission guidelines.

Report 1 (Referee #3)

Figure 6: I found this figure difficult to follow in relation to the text. Are NH3 (anth+ fire) emissions considered part of the Nr emissions? Additionally, what is the difference in emissions between water and land? Does 'anth NOx emission (water)' refer to shipping? Better notations on the figure would greatly enhance both the figure and the associated discussion.

Thanks for the questions. The definitions of "reactive nitrogen (Nr)" in literature vary. In some cases, it includes NH_x, and in the others, it does not. In this article, while oxidized nitrogen is more emphasized, NH_x is also included in reactive nitrogen. See definition of Nr in Section 1. Labels like ".. emission (water)" refer to emissions from the model grids that are classified as water. These are not necessarily the same as shipping emissions, because some of the anth emissions from the shipping sector are assigned to grids overland (e.g., ports and surrounding areas). We added "*water and land model grids are defined in Fig. 1b*" to Fig. 6 caption.

Lines 425-430 and Figure 11: I'm unclear on what is meant by a higher correlation between O3 and the NO2 column compared to the HCHO column. This information doesn't seem directly applicable to identifying NOx-sensitive chemical regimes, which are typically determined by the relationship between P(O3) (O3 production rate) and NOx levels. What are the key takeaways from these correlation results?

Good point. Fig. 11 has been updated, which now indicates the NO_2 columns-daytime surface O_3 relationship as well as its dependency on column $HCHO/NO_2$ ratio. Relevant sentences in this paragraph and elsewhere are modified to explicitly draw implications from this plot (along with other results discussed in this paragraph) regarding the effects of NO_x changes on O_3 in this area, as well as the utility of remote sensing NO_2 and HCHO column data in inferring surface O_3 variability across the area. Please also see the previous paragraph on satellite $HCHO/NO_2$ as an indicator of chemical regimes.

Report 2 (Referee #1)

Lines 68-79: The studies reported here about ability of regional models to model surface ozone are outdated, and reference DA attempts to improve results dating back to 2007. Thus, I think stating biases of surface O3 up to 20 ppbv is truly not representative of the state-of-science regional CTMs/AQMs that are available today for the U.S. CMAQv5+ for instance can well simulate surface O3 with biases much less than 20 ppbv in the U.S. without significant DA (e.g., CMAQ). Some Examples. 1) Offline CMAQv5.3.1, Appel et al. (2021) https://gmd.copernicus.org/articles/14/2867/2021/, reports surface ozone well within +/- 5 ppb.). 2) Two-Way Coupled WRF-CMAQ with Noah LSM updates: Campbell et al. (2019), https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2018MS001422 show improvements in ozone without DA. This should be at least discussed here for state-of-science CTMs in context to fully coupled ESM configurations that also try to model surface ozone (maybe less successfully). Furthermore, its worth noting that when using well constrained, bottom=up emissions for the simulation period (e.g. NEI) and robust chemical mechanisms (and even empirical approaches to dry deposition) offline or online CTMs like CMAQv5+ are shown to recently perform very well for surface ozone in the U.S. (<< +/-5 ppb).

Changed to: "*…large model-observation mismatches in surface O_3 of up to tens of ppbv were not well explained or attributed mainly to the models' uncertain/outdated anth emission inputs…*"

We do not agree with this reviewer that ESMs (which also evolve quickly through time) perform less successfully. We point out that the overall model biases reported in some of these suggested papers (and other modeling works) are results of positive and negative biases in different US regions being cancelled out, and Appel et al. (2021, suggested by this reviewer) showed that CMAQ updates degraded the model performance for some US regions/seasons. Additionally, none of these studies include source attribution analysis which is a discussion point in this paragraph. As this reviewer noted, model performance is highly dependent on their inputs and parameterizations, and the same models' performance (including CMAQv5.3.1+ runs at regulatory and operational agencies and in academia) can vary substantially in different applications. This is also supported by Figure S1 of Hogrefe et al. (2023, process level study) that showed big differences in Appel et al. (2021) and AQMEII CMAQv5.3.1 O_3 and aerosol fields. Model simulations with the bottom-up NEI (not available for every year) emissions for their base years over the US are supposed to lead to better model performance than for non-NEI years. Such findings are in some of the papers we already cited, which also demonstrate that chemical data assimilation is an effective approach to improve the NEI, especially for non-NEI years. The choice of chemical mechanism may also impact O_3 in regional models by several ppbv according to numerous existing sensitivity studies.

Campbell et al. (2019) demonstrate that tuning several sets of static, hard-coded parameters can improve the modeled (with Noah LSM and empirical dry deposition methods) air pollution fields for slightly over 50% of their grids and the model performance in other grids was worsened. Certainly, tuning static parameters is one way to improve models. However, in general, less complex modeling systems with empirical approaches for processes like dry deposition are less suited to evaluate the sensitivities of air pollution states and processes to various climatic factors (e.g., Niyogi and Raman, 1997, and many later studies, on assessing stomatal resistance by different schemes). Evaluating air pollution responses to climate change has become increasingly important to help better understand the Earth systems and their interconnectivity as well as assisting in developing emission control strategies. Less complex modeling systems may be more computationally efficient, have fewer and more easily identifiable sources of uncertainty, and their pollution fields can be less responsive to the applications of data assimilation that adjust environmental and biophysical conditions. Therefore, models and their configurations should be chosen based on the objectives of research and applications and carefully evaluated.

Hogrefe, C., Bash, J. O., Pleim, J. E., Schwede, D. B., Gilliam, R. C., Foley, K. M., Appel, K. W., and Mathur, R.: An analysis of CMAQ gas-phase dry deposition over North America through grid-scale and land-use-specific diagnostics in the context of AQMEII4, Atmos. Chem. Phys., 23, 8119–8147, https://doi.org/10.5194/acp-23-8119-2023, 2023.

Niyogi, D. S. and Raman, S.: Comparison of Four Different Stomatal Resistance Schemes Using FIFE Observations, J. Appl. Meteorol. Climatol., 36, 903–917, https://doi.org/10.1175/1520-0450(1997)036<0903:COFDSR>2.0.CO;2, 1997.

Lines 167-168: There are many studies on the impacts of background NOx sources when inferring emissions from satellite sources in the literature. This sentence should be revised with adequate citations, e.g., Silvern et al. (2019): https://doi.org/10.5194/acp-19-8863-2019 Qu et al. (2021) https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2021GL092783 East et al. (2022) https://acp.copernicus.org/articles/22/15981/2022/acp-22-15981-2022.pdf

As noted in our initial response, the related point of view in some of these suggested papers has been covered in Elguindi et al. (2020) and we think this citation is sufficient for this context. The lightning NO_x biases shown in East et al. (2022) were also recognized by Elguindi et al. (2020) qualitatively and quantified in much earlier studies, such as:

Jourdain, L., Kulawik, S. S., Worden, H. M., Pickering, K. E., Worden, J., and Thompson, A. M.: Lightning NO_x emissions over the USA constrained by TES ozone observations and the GEOS-Chem model, Atmos. Chem. Phys., 10, 107–119, https://doi.org/10.5194/acp-10-107-2010, 2010.

Miyazaki, K., Eskes, H. J., Sudo, K., and Zhang, C.: Global lightning NO_x production estimated by an assimilation of multiple satellite data sets, Atmos. Chem. Phys., 14, 3277–3305, https://doi.org/10.5194/acp-14-3277-2014, 2014.

Lines 181-182: Since the driving meteorological reanalysis dataset and corresponding WRF land/meteorological simulations can strongly control the chemistry and surface-atmosphere exchange processes, I feel that a base evaluation of the WRF output across eastern U.S. domain (against 2D/3D Met observations like METARs/RAOBs/BSRN/PRISM) is lacking here. Particularly when downscaling from relatively coarse 32 km ICs/BCs. I suggest that some basic meteorological evaluation are included at least in the supporting information and discussed in later results section to help qualify the overall Met results and discuss implications for Met biases on AQ and sfc-atm-x processes. This could also prove very useful to discuss impacts of the land/SM DA on the Met performance, not just quantifying the potential influences of land->weather changes on AQ. For example, does the WRF Met performance also improve when including the land/SM DA? Why or why not?

Thanks for the suggestion. The multi-year model precipitation and temperature performance is already indicated in several figures. Our previous SM DA case studies include detailed met evaluation. In this paper's case study, evident improvements in the modeled air temperature due to SM DA are now presented (Section 3.3.1/Figs. 14 and S18), which contributed to the improved surface O_3 performance.

Lines 185-186: This sentence is unclear. Are the authors saying that other chemical reanalysis products are more accurate than WACCM, or vice versa? Also, references/citations are needed here to support this statement.

Changed to: "*are likely to be more accurate*".

While there are many references on chemical reanalysis products, the conclusions therein cannot be directly applied to this study. We did not compare the chemical boundary models used and chemical reanalysis products for our study period. However, model sensitivities to chemical boundary conditions have already been presented in a case study (Section 3.3.3).

Line 188: Awkward wording to use "hiked by". Please revise.

Changed to "rose".

Lines 272-273: This relates to my earlier comment on adding Met evaluation with independent observations, not just quantify impacts.

Please see our response to your earlier comment on met evaluation. "*Impacts*" refer to the resulting changes in model fields and their accuracy.

Line 353: This sentence does not make sense. Do you mean "greatly resemble one another"?

Changed as suggested.

Line 418: "Manifests" is a very awkward writing in this sentence.

Changed to "indicates".

Line 426: Sentence is incomplete, please revise to "resemble one another".

Changed as suggested.

Line 427: This part of the sentence is also poor grammar/sentence structure,. please revise.

This sentence has been broken down into two sentences.

Lines 432-434: This is again poor sentence structure, run-on, missing appropriate commas.

Comma added.

Lines 438-444: This effect is spatiotemporally variable in the U.S., with some increases in daytime ozone due to COVID-19 induced emissions changes. See Figures 6-7 in: https://doi.org/10.1016/j.atmosenv.2021.118713.

These are domain-wide results. Spatial patterns of daytime surface O_3 fields are shown in Fig. 10 of this article.

Line 448: Please see earlier comment. I don't think that such large bias/error of ~ 20 ppb is common amongst state-of-science regional AQMs/CTMs. Please revise here and above.

Changed to "*tens of ppbv*", also accounting for RMSEs reported in Appel et al. (2021) for this region/warm seasons.

Lines 470-471: I don't agree with this argument based on Figure 12b. This simply shows the magnitude of Nr deposition to different vegetation and water. However, the relative ecosystem impacts plotted in some way (as a function of impact on herbaceous plants, lichen species, algae blooms/acidification/anoxic conditions, etc.) may be a whole different thing in regards to impacts. Suggest revising.

Fig. 12b was cited at the correct location, following "*The potential impacts of Nr deposition are strongest and weakest on croplands and water, respectively*". Figs. 8b-c and S15 are now cited immediately following the previous sentence, which together indicate the year-to-year changes in the speciated deposition fluxes and their respective contributions to the total fluxes.

Lines 488-509: If the observed large surface ozone of about 30 ppb was due largely due to the impacts of frontal passage, precipitation and soil moisture changes, then the improvement due to SSM DA of ~ 2 ppb is only about a 5% of this change. Likely the largest primary contribution to this drop is the direct weather effects resulting cleaner airmass and lower temperatures behind the frontal passage, not necessarily SSM responses and secondary weather feedbacks. I think this limitation should be better discussed.

A ~2 ppbv change in ambient O_3 concentration is non-trivial considering the economic cost of air pollution reduction. A ~2 ppbv reduction in the modeled O_3 bias is not small considering the many factors that can impact the O_3 performance. This is in fact better than/comparable to the improvements in O_3 for the New England region due to updating Noah LSM parameters of a WRF/CMAQ system shown in Campbell et al. (2019, suggested earlier by this reviewer). Please note that all the impacts discussed in these lines (not the 30 ppbv drop) must be attributed to SM DA, as they were determined from the no-DA and DA cases.

We agree that investigating SM DA impacts on O_3 and other variables across three dimensions under various weather conditions is an interesting direction - please see our previous studies during other field campaigns conducted in the US and Asia, several of which have been cited in this paper. Also, in the following paragraphs and Figs. 14 and S18, we highlight larger SM DA impacts (>4 ppbv) on the model's surface O_3 performance in other eastern US regions on interannual timescale.

Lines 591-592: Also see https://www.sciencedirect.com/science/article/pii/S0048969722032272 and https://library.wmo.int/records/item/62090-no-3-september-2023 (pp 7-8).

The approach in Campbell et al. (2022) is similar to that in an earlier paper we already cited. One of its NOAA authors' affiliation is written wrong. As this suggested paper appears that it was not carefully proofread and very likely not internally reviewed/approved by all relevant NOAA offices prior to its submission (a NOAA requirement), we do not cite it.

Line 627: This sentence is incomplete ..."caused biomass/crop yield losses by a few percent". In which direction?

Loss means reduction.

643-644: I suggest revising this based on my earlier comments, as such large ozone biases/error quoted in this paper is not reflective of state-of-the-science AQMs.

See our responses to your earlier comments. Corresponding sentences in other places describing the varying model performance for this region/warm seasons have been changed to "*tens of ppbv*".