Souri et al present a detailed study highlighting four major shortcomings associated with FNRs and their ability to categorize ozone sensitivity. The sections about column-to-PBL translation, spatial representation error, and retrieval error are all well-written. The manuscript as a whole has understandable writing style and clear, well-made figures. However, I do have a few major concerns, mostly surrounding the modeling section of this work. I recommend that the manuscript be sent to the authors for major revisions.

Answer
We thank the reviewer for taking their time to provide constructive comments. Our response follows:

VOC inputs for the box model: The modeled radical environment can be incredibly sensitive to changes in VOC inputs, especially in polluted urban areas. This manuscript is lacking detail about how VOC inputs were created, leaving readers to assume the authors used a simplistic approach that excludes many potentially important VOCs. As written, the authors' treatment of VOC inputs does not rise to the level established in previous modeling studies performed for the same field campaigns, leaving this reviewer wondering if the modeling presented in this study can represent the ambient radical environment.

The field campaigns modeled in this study have unique VOC measurement suites which require unique data engineering strategies to generate realistic VOC inputs. DISCOVER-AQ was served only by a quadrupole PTRMS, and features a very limited set of VOCs. The authors do not give adequate detail about how they generated VOC inputs based on these data. For example, previous studies (i.e. Schroeder et al 2017) generated speciated VOC box model inputs for DISCOVER-AQ using a fusion of VOC data from concurrent airborne campaigns (DISCOVER-AQ+SEAC4RS+FRAPPE). This enabled somewhat realistic estimation of VOCs that were not measured by the PTRMS during DISCOVER-AQ.

During KORUS-AQ, the whole air sampler was flown concurrently with a PTRMS, giving a richer suite of speciated VOCs. However, these two instruments had wildly different sampling cadences and integration times, with WAS measurements being incapable of resolving fine-structure details in pollutant gradients. As a result, previous studies (i.e. Schroeder et al 2020) fused the two datasets together to generate a pseudo-high-resolution set of VOC inputs for their box modeling work with KORUS-AQ.

As it is currently written, I have serious concerns about the VOCs used as model inputs, and thus have lower confidence in the results presented here. Can you show that the simplistic VOC inputs used in this study do not yield significantly different results from the two Schroeder papers?

Perhaps a more pointed observation: the box model inputs and outputs from the two Schroeder papers are publicly available online. What does this study gain by running its own model simulation – with questionable VOC representation – instead of using the freely-available Schroeder/Crawford data which has already been heavily vetted and used in multiple studies?

Answer
We fully understand the reviewer's concern regarding the gaps associated with the VOC measurements and that each field campaign provides a different set of measurements. When setting up the model, we factored in two criteria about how we should go about the VOC treatment (which includes both measurements and the model chemical mechanism). **Two major criteria:** **1. How much difference does a specific chemical mechanism make in terms of OH, HO_2, and HCHO, given a fixed number of measured VOCs?**

We have tested the GEOS-Chem v12 chemical mechanism with an update to aromatic VOCs based on Bates et al., 2021, MCM, and the well-established CB06 mechanism. Based on our observations, we realized that the CB06 mechanism could simulate results similar to the MCM but at a much cheaper computational cost. Unlike previous studies, we feel it is better to not have HCHO constrained so that we can truly understand how much variance (information) in the observations each model realization can replicate, given the chemical mechanism and measurements used. We concluded that the selected VOCs (Table 1) are sufficient to replicate observations with more than 70% variance in HCHO. We also noticed that the performance of radicals such as HO_2 and OH are highly similar compared to previous studies such as Souri et al. (2020), Schroeder et al. (2017, 2021), and Brune et al. (2021) (who also compared two different model realizations including the LaRC model used in Schroeder et al. (2021)). If the VOC treatment had been unsatisfactory, we might have observed an inferior performance in terms of HOx compared to Souri et al. (2020), Schroeder et al. (2017, 2021), and Brune et al. (2021), which is not the case (statistics had been provided in the text).

In our previous study, we compared a very similar setup with the NASA LaRC box model over a highly complex environment (Seoul, Korea) and observed a strong agreement between our model and NASA LaRC. Please see https://www.sciencedirect.com/science/article/pii/S1352231020305276

We realized that a HCHO-unconstrained version of the model output exists to test against our setup (https://www-air.larc.nasa.gov/cgi-bin/ArcView/korusaq?MODEL=1#CRAWFORD.JAMES/). We were hoping to conduct the same analysis with DISCOVER-Colorado AQ but we were unable to find the LaRC unconstrained simulation results on the DISCOVER-AQ archive. We synched the timetag between _input and _unconstrained files and averaged 1 Hz data to 10 sec. We conducted several sensitivity tests, including running the model with a fixed dilution factor (the original setup), without considering a dilution factor (=0), unconstraining H2O2 and HNO3, and compared the simulated HCHO with the observed ones:

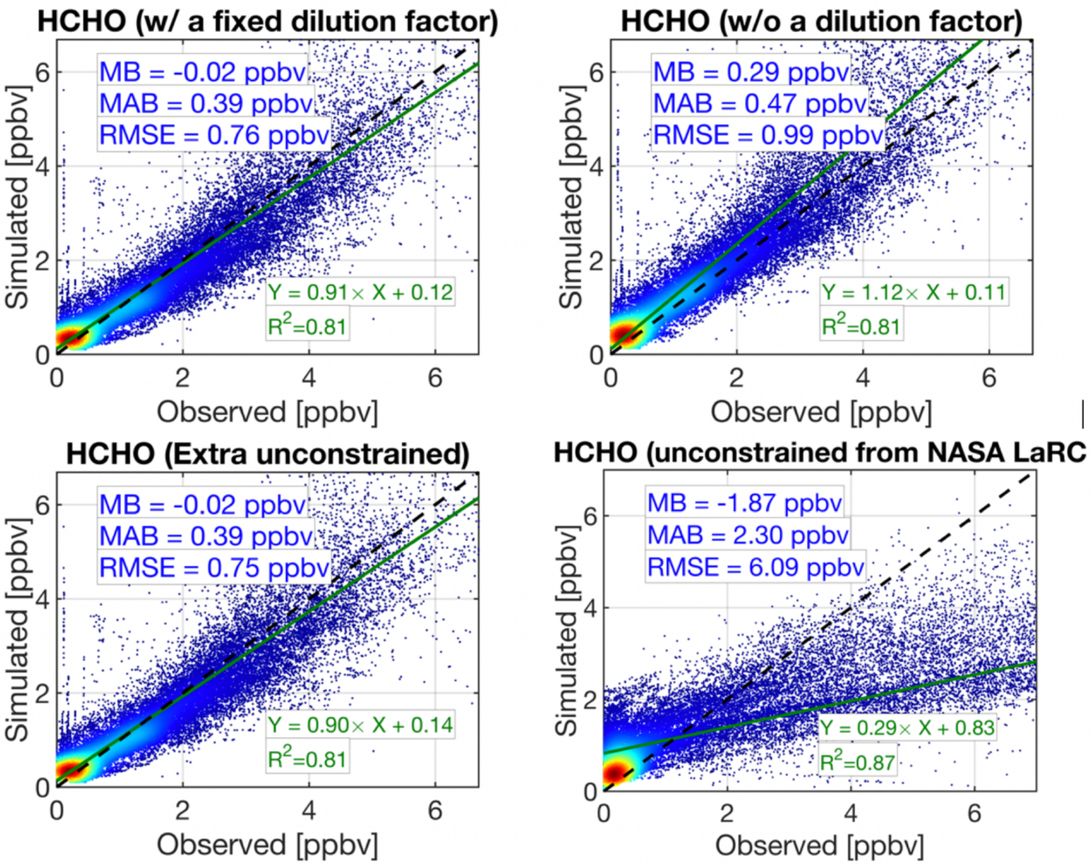

Figure S2. The comparison of simulated HCHO mixing ratios compared to observations for (top left) our F0AM setup with the dilution process on, (top right) same model but without the dilution process, (bottom left) our F0AM setup with dilution process on and without constraining HNO3 and H2O2, and (bottom right) NASA LaRC unconstrained model based on Schroeder et al., 2021. All points are based on 10-sec sampling size.

Several tendencies can be observed:

- **All F0AM configurations capture 81% of the variance in observed HCHO with relatively low bias**
- **The dilution factor used in our study does not change the correlation coefficient or bias between simulated and observed HCHO.**
- **NASA LaRC box model captures the variance in HCHO slightly (6%) better than F0AM/CB06 but with a significant bias.**

A valid criticism is how we can get HCHO right while leaving PAN unconstrained. When we compare the simulated PAN mixing ratios with the observations w/ and w/o dilution factor (see figure below), the results significantly changed, which could adversely impact the performance of HCHO. But when we super-imposed the aircraft altitude, we realized that those highly overestimated PAN mixing ratios mostly occurred in very high altitudes where PAN is thermally stable. It indicates that our model is skillful at reproducing PAN in lower altitudes, even without considering a dilution factor. To sum up, without considering the PAN constraint and a dilution factor, our model can still capture the magnitude and variability of HCHO which suggests that the box model results in this study are an accurate representation of ambient chemical conditions.

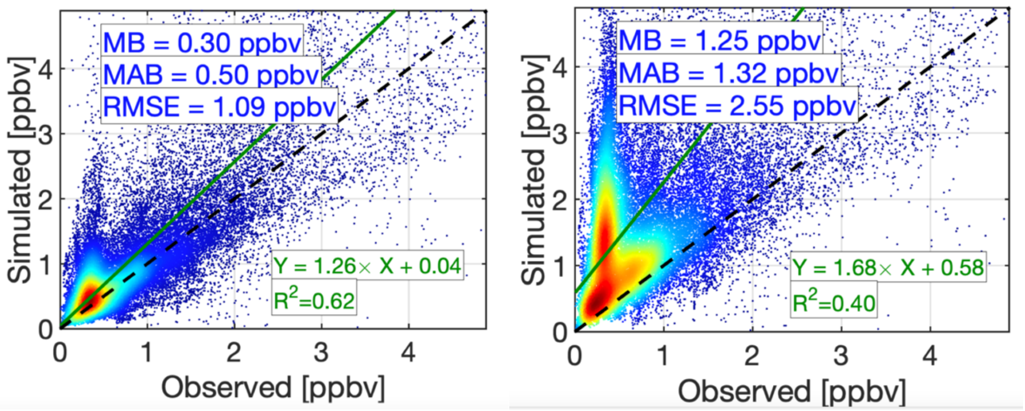

Figure S3. (left) The comparison of PAN mixing ratios w/ a fixed dilution factor and (right) w/o a dilution factor during KORUS-AQ campaign.

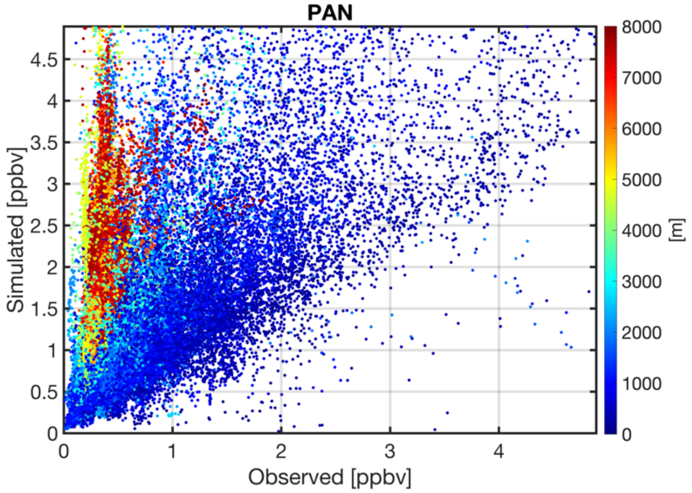

Figure S4. Same as the above right figure but with aircraft altitude superimposed.

2) **What are the overlaps between the CB06 mechanism and the measured VOCs?**
The CB06 mechanism does not include several measured VOCs such as butane, hexane, and styrene. The reviewer can see the list of compounds this mechanism covers Table 5.1 in https://github.com/AirChem/F0AM/blob/master/Chem/CB6r2/CAMxUsersGuide_v6-30.pdf). We did our best to find the overlaps listed in Table1.

Modifications

We first added this part in section 2: "Regarding the KORUS-AQ campaign where HOx observations were available, we only ran the model for data points with HOx measurements. Similar to Souri et al. (2020), we filled gaps in VOC observations with a bilinear interpolation method with no extrapolation allowed. In complex polluted atmospheric conditions such as that over Seoul, South Korea, Souri et al. (2020) observed that this simplistic treatment yielded comparable results with respect to the NASA LaRC model (Schroeder et al. 2020) which incorporated a more comprehensive data harmonization."

We removed these two parts from section 3.1 and the conclusion, because the dilution factor did not allow the total bias to go above 5%:
Concerning HCHO, our model does have considerable skill at reproducing the variability of observed HCHO (R^2=0.73)~~ with a low bias of -4.9% (-0.09 ppbv).~~

Our box model showed a reasonable performance at recreating some of unconstrained key compounds such as OH (R^2=0.64, bias=17%), HO_2 (R^2=0.66, bias<1%), and HCHO (R^2=0.73~~, bias=5%~~).

We added the comparison of HCHO w/ and w/o dilution factor and with NASA LaRC:

"Concerning HCHO, our model does have considerable skill at reproducing the variability of observed HCHO (R^2=0.73). To evaluate if this agreement is accidentally caused by the choice of the dilution factor and to identify if our VOC treatment is inferior compared to the one adopted in the NASA LaRC (Schroeder et al., 2021), we conducted three sets of sensitivity tests for the KORUS-AQ campaign, including ones with and without considering a dilution factor and another without HNO_3 and H_2O_2 constraints. When not considering a dilution factor results in no difference in the variance in HCHO captured by our model (R^2=0.81). Our model without the dilution factor is still skillful at replicating the magnitude of HCHO with less than 12% bias. It is because of this reason that the optimal dilution factor for each camping is within 12 hr to 24 hr which is not different than other box modeling studies (e.g., Brune et al., 2022; Miller and Brune, 2022). We observed no difference in the simulated HCHO when HNO_3 and H_2O_2 values were not constrained. The unconstrained NASA LaRC setup oversampled at 10-sec frequency captures 86% variance in the measurements, only slightly (6%) outperforming our result. However, the unconstrained NASA LaRC setup greatly underestimates the magnitude of HCHO compared to our model results."

We also added the PAN comparison w/ and w/o dilution factor for the KORUS-AQ campaign:
"Moreover, we should not rule out the impact of the first-order dilution factor which was only empirically set in this study. For instance, if we ignore the dilution process for the KORUS-AQ campaign, the bias of the model in terms of PAN will increase by 33% resulting in a poor performance (R^2=0.40) (Figure S3). We notice that this poor performance primarily occurs for high altitude measurements where PAN is thermally stable (Figure S4); therefore, this does not impact the majority of rapid atmospheric chemistry occurring in the lower troposphere such the formation of HCHO."

Model Setup: I have a few concerns with model setup:

Why use an arbitrary model run time of 5 days? Ideally, the model should be run indefinitely until it converges on a solution for key species, but I understand the desire to set a lower limit for the sake of computation. Do your outputs change if you use 4 days? Or 6 days? Or 20 days? Can you include a sensitivity analysis to back up your work – that is, show that your arbitrary choice of 5 days does not impact results?

Answer

Thanks for the comment. Our previous study (Souri et al., 2020) used three solar cycles which resulted in almost net zero HCHO production rates. For the current study, to be conservative, we decided to increase it to five days and phrase this as "approaching" instead of "reaching" steady state. To alleviate the reviewer's concern, we ran 10 solar cycles and compared HCHO concentrations to the 5 solar cycles used in our study. The results remained identical meaning we have already reached to the steady state.

Modifications

We added to Sect 2: "We test the number of solar cycles used in this study (5) against ten days on the KORUS-AQ field campaign model setup, and observe no noticeable difference in simulated OH and HCHO (Figure S1), indicating that the choice of five solar cycles suffices."

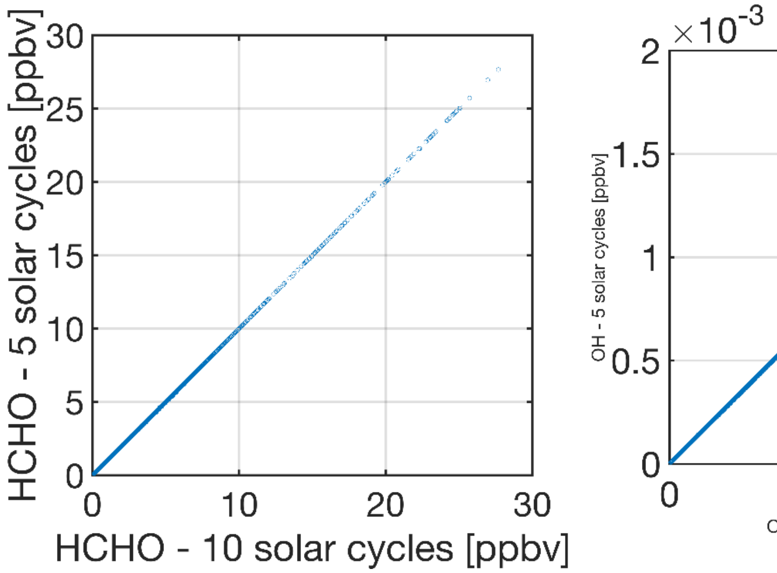

Figure S1. The comparison of simulated HCHO (left) and OH (right) with 5 (y-axis) and 10 (x-axis) solar cycles.

If I understand this correctly, you calculate a unique dilution factor for each field campaign, deriving it empirically to yield the best agreement between measured and modeled HCHO. What is the physical basis for why one field campaign would have different dilution rates than another? Without further explanation, this feels like an arbitrary "correction factor" to game the model for better agreement with observations – which does nothing to tell you how well the model represents the underlying chemistry. Can you explain?

Answer

The dilution factor is a highly oversimplified parameter that represents generic physical loss. It is typical, for example, to impart a 24-h lifetime to all model species to approximate physical losses due to transport, deposition, etc. Because those physical parameters vary greatly by time and space, it is logically flawed to consider a constant dilution factor for all locations/times. We chose a constant dilution factor for each campaign in order to avoid over-constraining the model by setting a dilution factor to a value for each point measurements to artificially match observations. We want to keep the box model as simple (but not overly simple) as possible to allow clear intuition about PO3 tendencies. We would like to clarify that we do not claim that using the mean bias of simulated HCHO is the best approach to understand the underlying processes of the dilution factor, as understanding many of those parameters requires the precise knowledge of land surface processes, momentum fluxes, etc. which will turn the box model framework to a full chemical transport model.

Schroeder et al. 2021 mentioned in their paper: "For instance, Fried et al. (2011) demonstrated that model estimates of CH2O can be biased in pollution plumes when short-lived species dominate the model photochemistry, e.g., highly-reactive VOCs such as biogenic isoprene or alkenes from industrial sources. Under such conditions, using observations to constrain CH2O and other important radical reservoirs instead of calculating them will improve model calculations of the radical pool responsible for ozone production." **There is a trade-off between fully constraining HCHO (to get the radicals simulation straight) while masking the understanding of if the model can actually reproduce HCHO, and keeping the mean bias of HCHO low with an optimal dilution factor while allowing HCHO to fluctuate. We chose the latter because it enabled us to have a more stringent validation on the overall model performance.**

Given the differences in physical and meteorological conditions of the various regions of the field campaigns, one would expect that a different dilution factor for each campaign should be used. To alleviate the reviewer's concern, we performed a sensitivity test in which we entirely removed the dilution factor and reran the box model for the KORUS-AQ campaign. In this sensitivity simulation we observed negligible changes in HOx, NOx, and HCHO. PAN was obviously impacted the most; however, the vast majority of PAN overestimated values were at high altitudes where PAN is thermally stable (Figure S4).

Modifications

We have already included new figures regarding HCHO and PAN based on the first comment.

We added in Section 2 about the rationale behind using the dilution factor for HCHO:

"Because the model does not consider various physical loss pathways including deposition and transport, which vary by time and space, we oversimplify their physical loss through a first-order dilution rate set to 1/86400-1/43200 s^{-1} (i.e., 24- or 12-hr lifetime), which in turn prevents relatively long-lived species from accumulating over time. Our decision on unconstraining HCHO, a pivotal compound impacting the simulation of HOx, may introduce some systematic biases in the simulation of radicals determining ozone chemistry (Schroeder et al., 2020). Therefore, to mitigate the potential bias in HCHO, we set the dilution factor to maintain the campaign-averaged bias in the simulated HCHO with respect to observations of less than 5%. However, it is essential to recognize that HCHO can fluctuate freely for each point measurement because the dilution constraint is set to a fixed value for an individual campaign."

We also added the comparison of HOx w/ and w/o considering the dilution factor in the supplementary material.

"A sensitivity test involving removing the first-order dilution process demonstrates that the simulation of HOx is rather insensitive to this parameter (Figure S5). This might be caused by the fact that the simulated HCHO already agrees relatively well with the observations without the dilution factor."

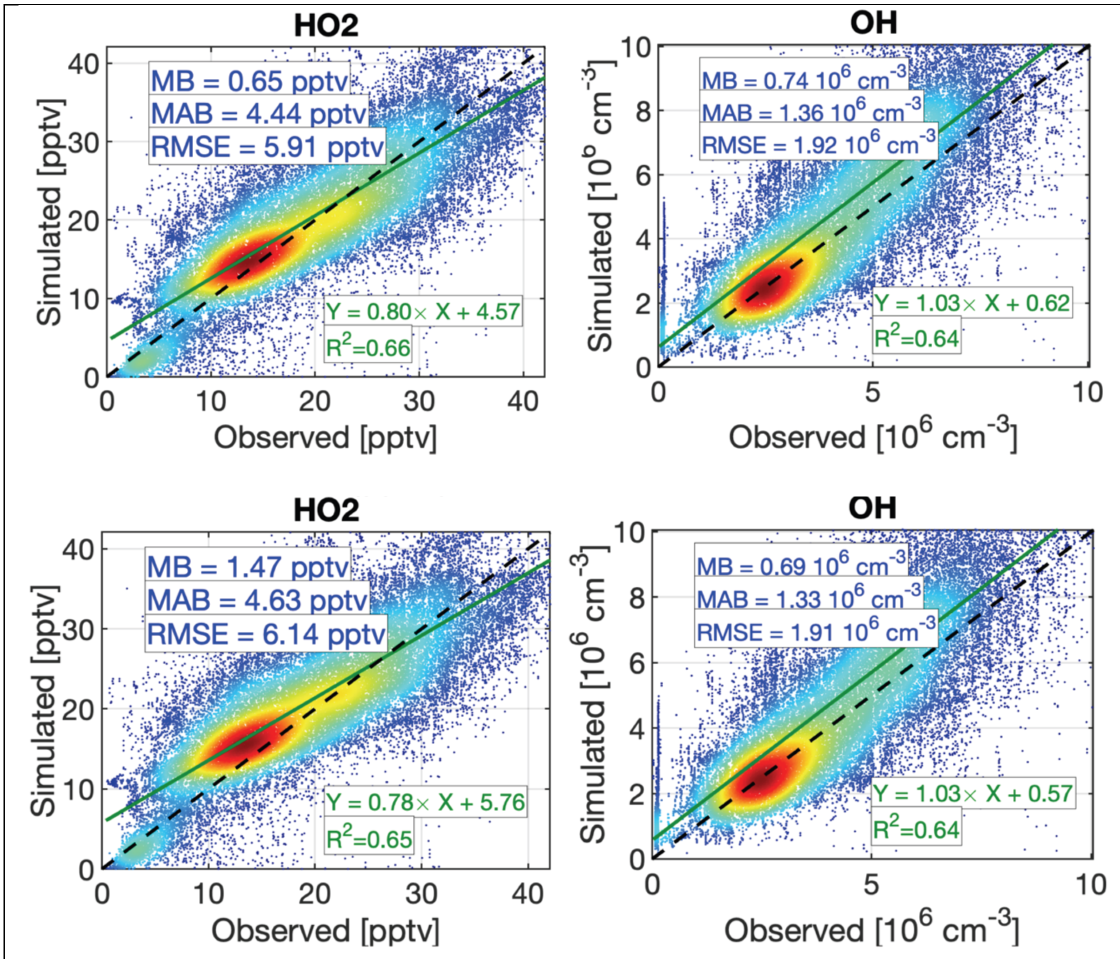

Figure S5. (top) The simulation of OH and HO_2 using a fixed dilution factor during the KORUS-AQ campaign. (bottom) without considering the dilution factor.

Model Validation: As written, the model validation section does not give me confidence in the model's ability to represent the ambient radical environment (especially given the simplistic treatment of VOCs).

If one of the model parameters (dilution rate) is based on empirical model/measurement agreement, then comparing simulated values to observations is cyclical. If I am understanding this correctly, then Section 3.1 is incredibly problematic. Based on the description given in line 220, campaign-average simulated HCHO is not allowed be >5% off from campaign-average observed HCHO – its part of the model setup with an empirically derived dilution factor. How can you evaluate the model's representation of the chemical environment with such a setup? For example, in line 254 you state that HCHO had a mean bias of less than 5% - which is meaningless because you've coded the model to do exactly that.

In practice, your dilution factor acts as a quasi-constraint on HCHO, which greatly influences calculated radical budgets. This eliminates your ability to truly test whether the model is capable of representing the radical budget from first principles. Furthermore, this does not allow you to test if your simplistic treatment of VOCs is adequate. You mention that the bias in PAN changes if you ignore dilution, but PAN can have a large impact on modeled radical and NO2 concentrations. What happens to other test-species if you ignore dilution?

I don't buy the idea that this model, as currently setup, has proven itself as sufficient for representing ozone chemistry.

Answer

We have addressed this valid concern in the previous comment. Our results are insensitive to the dilution factor.

In short, you have questionable VOC inputs and a questionable model setup which prevents you from truly testing model performance. I'd suggest re-running the model with a fixed dilution factor based on reasonable physics. This will enable "true" unconstrained model runs, providing a testbed for evaluating model performance (and your VOC inputs). Or, you could run another, established model in parallel on a subset of data and compare the two. Or, you could use the freely-available model inputs/outputs from published studies from the same field campaigns, rather than re-invent the wheel.

Answer

We showed that ignoring a dilution factor did not broadly impact the biases in HOx and HCHO compared to our baseline simulations in the study. We want to clarify that we did not apply a variable dilution factor for each point measurement to over-tune the model. A fixed dilution factor for each campaign did not impact the variability in simulated HOx and HCHO. The most significant impact caused by the dilution factor on our result was seen for PAN in high altitudes, whose effects on HOx and HCHO were inconsequential. Comparing the unconstrained NASA LaRC model output with our model during KORUS-AQ showed that both models were skillful at replicating HCHO (R^2=0.87 vs. R^2=0.81). Even so, our model shows less significant biases in HCHO than the NASA LaRC model.

Regarding the comment about "re-inventing the wheel" by not using the LaRC box model, we have strong reasons for not doing so:

1- **It is in our opinion that that while the NASA LaRC publicly available (not sure if it applies to DISCOVER-AQs) outputs are great, not having the capability to run, modify, update, perform sensitivity tests, receive technical and scientific helps from the steering committee and peers (such as for publicly available models such as F0AM (0-D), WRF, CMAQ, GEOS-Chem, CESM, and MUSICA) is a disadvantage for conducting research. For instance, we would not be able to conduct simple sensitivity tests requested by a reviewer, such as we have done here, when we cannot run the actual model. Furthermore, if a reviewer asked about chemical sinks or sources of HCHO, or individual terms of PO3 that are not saved in the outputs, it would be impossible to address such requests.**

2- **The F0AM model has been extensively used in more than 72 published studies with notable citations. This is a testament to open-access and transparent models. Plus, as this is NASA funded research it is important for the purposes of "open science" it is vital to use a publicly-available model to allow reproduction of our results.**

3- **A major concern of ours for running a box model was the choice of chemical mechanism. The F0AM model provides the flexibility to easily test and apply different well-established chemical mechanisms or update individual reactions. We did not include those sensitivity tests in the paper, because we do not want to overwhelm readers of this study.**

Purpose of the paper: Finally, I would challenge the authors to include paragraphs in the Introduction and Summary sections describing the motivation for doing this work. Why are incremental improvements in our understanding of FNRs and ozone chemistry necessary? Martin et al first published their paper about satellite FNRs more than twenty years ago – yet, to the best of my knowledge, no regulator or policymaker has ever used satellite FNRs in their ozone planning strategies. Clearly FNRs were first developed as a potential tool for policymakers to fine-tune ozone mitigation strategies, but if policymakers have shown no interest in using these tools, why continue refining them? Is there a pathway for FNRs to be used by anyone outside of academia? What does the author think is preventing policymakers from using this tool - or is this simply a tool for academics?

Answer
This study was designed specifically to tackle the limitations of using satellite-based FNRs and directly answers why the satellite-based FNRs (or, in general, satellite trace gases) have not been widely used by regulators. The purpose of this paper, as explicitly mentioned in the title, abstract, etc., is to quantify the errors associated with FNR rather than to refine the metric. To our knowledge nowhere in the manuscript do we suggest that satellite FNR values are sufficient for regulatory purposes. This study, for the first time, quantifies the main errors/uncertainties associated with satellite column trace gas retrievals and resulting FNRs. We argue these results could be used by regulatory agencies for a quantitative understanding of why satellite data is currently not sufficient for application in developing/testing ozone mitigation strategies. **While we think it is essential to share the limitations and challenges associated with using satellite products with air quality regulators, we believe it will be well out of the ACP focus, which is "*on studies with important implications for our understanding of the state and behaviour of the atmosphere*". Furthermore, the funding agency for this work, NASA, is a Research & Analysis (*R&A*) organization and is not focused on regulatory processes. Providing quantitative information about satellite capabilities for air quality research is the primary driver of this research and the results are directly applicable for regulators. Some limitations of satellites identified in this study have been overlooked for years and not mentioned in any studies back to Martin et al. Our study is an eye-opener for the retrieval community as our work demonstrates that retrieval errors must be reduced for satellites to be applicable for regulations.**