Reply to Reviewers

We thank the two reviewers for their constructive comments and suggestions, which have further improved the quality of our manuscript considerably. Their comments are reproduced below with our responses in blue. The corresponding edits in the manuscript are highlighted with tracked changes. The line numbers are based on the revised manuscript.

Reviewer #1:

The manuscript titled "Cluster-based characterization of multi-dimensional tropospheric ozone variability in coastal regions: an analysis of lidar measurements and model results" by Claudia Bernier et al. developed a clustering analysis of multi- dimensional measurements of ozone in coastal regions. The lidar clusters provided a more comprehensive perspective to evaluate the performance of three-dimensional models. The manuscript provides valuable information for understanding ozone chemistry in complex coastal regions. I would recommend publication if my following comments are well addressed.

For the issue about the models' poor performance in simulating mid-level O3, one influential process is the transport of O3 in the free troposphere from the continent to coastal areas. I wonder whether your model can capture this process accurately.

This is a great idea. To test this we used flight data from OWLETS-2 to evaluate the model's ability to simulate CO in the free troposphere (Figure S8). We evaluated just 6 flights (morning and evening flights) all which profiles belong to HMO, MCO, and HLO clusters. From this analysis we find that CO is simulated well in the free troposphere at lower levels (100 -110 ppbv) which is indicative of background levels of CO. On the other hand, the model handles higher levels (130 – 140 ppbv) of CO worse, with consistent underestimations in the free troposphere. Since higher levels of CO in the free troposphere are indicative of long-range transport from outside the study region, we can conclude that the GEOS-Chem model struggles to simulate transport in the free troposphere which directly influences the mid-level performance.

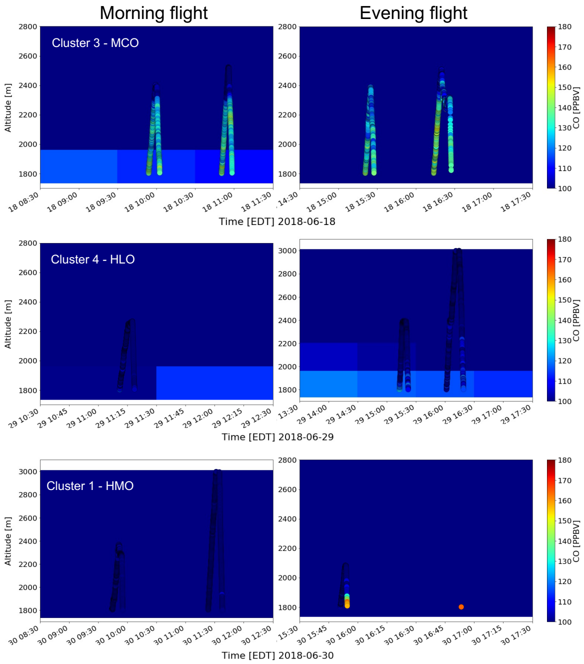


Figure S8. Aircraft measurements and GEOS-Chem simulated CO in the free troposphere during OWLETS-2. Measurements from the UMD Cessna 402B Research Aircraft.

Lines 517 – 523: "Additionally, transport of emissions in the free troposphere (FT) is another influential factor that could contribute to the misrepresentation of mid-level O_3. In Figure S8, aircraft measurements from OWLETS-2 are used to evaluate GEOS-Chem simulated CO in the FT (1800 – 2500 m AGL). The flight days evaluated are all curtain profiles that were assigned to the clusters with higher levels of O_3 in the mid-level (HMO, MCO, and HLO). It is evident that the model is able to capture lower levels of CO in the FT (100 – 110 ppbv) (e.g., background levels) but struggles to capture the higher levels (130 – 140 ppbv). Since increased levels of CO in the FT are indicative of possible long-range transport (Neuman et al., 2012), FT transport could be a factor contributing to the GEOS-Chem poor performance in the mid-level."

Lines 636 – 637: "Another factor inhibiting the poor simulation in the mid-level is the model failing to capture long-range transport of emissions in the FT."

Lines 513-530: You speculate that the overestimation of O3 in the morning is because of underestimation of NO titration at night in the cluster MCO. Some evidence should be provided. The verification of modeled NOX by observed values can help to understand this issue.

Model overestimation of O_3 at night and in early morning hours is a common problem for 3-D Eulerian CTMs. It is thought to be caused by the model's inability to resolve the shallow surface layer at night, which enhances nighttime NO titration and O_3 dry deposition. Lines 481 – 485 have been revised.

Lines 481 – 485: "Model overestimation of O_3 at night and in early morning hours is a common problem for 3-D Eulerian CTMs. Overnight, O_3 concentrations from the evening before can remain lingering in the residual layer. This residual layer sits at about 1000 m or higher depending on the conditions of the environment. O_3 trapped in this residual layer can directly correlate with the next day afternoon O_3 (e.g., Figure 3a; HLO cluster). Models struggle to resolve the shallow surface layer at night, which enhances nighttime NO titration and O_3 dry deposition. If this residual layer and the titration of O_3 overnight in the shallow surface layer is not resolved, next day simulated O_3 will most likely warrant even greater biases."

Also, in Fig. 6 it seems that GEOS-CHEM better captures the O3 levels in the morning than GEOS-CF for the clusters MCO and HLO. The reason for the discrepancy between the two models' performance should be clarified.
We have updated Lines 505 - 510 have been updated to emphasize this discrepancy.

Lines 505 - 510 : " GEOS-Chem does not have such an issue overestimating low-level O_3 in the afternoon. In the other clusters, GEOS-Chem actually underpredicts early morning low-level O_3 in the full vertical profile and does an overall better job than GEOS-CF simulating morning low-level O_3, such as in the HLO cluster. A better estimation of early morning O_3 does not warrant the same build-up of afternoon O_3. In these cases, GEOS-Chem handles the multi-day simulations better than GEOS-CF. This gives some explanation to why GEOS-Chem underpredicts the other clusters with higher O_3 concentrations in the low-level (HMO and HLO)."

Lines 531-532: The GEOS-CF model also overestimated the O3 in the afternoon even if it does not overestimate early-morning O3 (e.g., LLO and LMO). This means there exist processes other than nighttime NO titration causing the overestimation of O3 in the afternoon. I suggest to point out this issue and try to explain the potential causes.
This is an important distinction between the models and has been added to the manuscript (see Lines 528 – 532)

Lines 528 – 532: "GEOS-CF does best simulating morning low-level O_3 in cases of lower O_3 extent (LLO and LMO), but still overestimates the afternoon O_3. Since in these cases the afternoon does not seem to be related to morning estimations, other factors must be contributing. In the LLO cluster, the full curtain profile implies excessive mixing throughout the vertical profile is adding to afternoon O_3 overestimation. Similarly, for the LMO cluster, mid-level O_3 seems to be influencing low-level O_3 which could be adding to afternoon biases."

In Sec. 3.4.2: The specific effect of the models' performance simulating wind on O3 simulation is not clearly explained. For example, how the wind leads to the overestimation of O3 in MCO case by models, while leads to an underestimation of O3 in HLO case by GEOS-CHEM. From my understanding, the wind will at least influence the dilution rate and horizontal and vertical transport of O3.
Indeed, wind speeds are underestimated in all of the cases and by both models. This does help explain the overestimation of O_3 for both cases in GEOS-CF as well as the MCO case in GEOS-Chem but does not explain the underestimation in the HLO case for GEOS-Chem. The overestimation is easily explained as attributed by stagnant conditions in all the cases except for the HLO case by GEOS-Chem. The manuscript was updated to address and highlight the difference. One of the factors that could be contributing to the underestimation of O_3 in the HLO case could be the simulation of the boundary layer. We conclude the sea breeze mechanism is

not captured well by the GEOS-Chem model but better simulated by the GEOS-CF model likely due to resolution differences.

Lines 622 – 627: "Since the results reveal O_3 is underestimated, this suggests that there are more factors affecting O_3 results in this specific case. One of these factors can be the simulation of the boundary layer as the sea/bay breeze develops. If the boundary layer is simulated to be larger in depth, the ability for the model to simulate higher O_3 concentrations may be hindered such as found in Dacic et al. (2017). Since the HLO case indicates a common sea breeze event based on the timing and shift, it appears that GEOS-Chem really struggles capturing this intricate process while GEOS-CF does a better job."

Lines 623-635: I think this paragraph should focus on how the sea/bay breeze events cause a difference in O3 profiles between and MCO and HLO cases and how they influence the model simulations. The reason why the two curtains are not in the same cluster is not important for your research objective as it is just an observed phenomenon.
We agree with the reviewer and have added how the sea/bay breeze event can cause differences in the curtain profiles and how they influence the simulations (see Lines updated sect. 3.4.2 and Lines 628 – 637). But we have kept the final statement of why the curtains are not in the same cluster. We believe this is important since the purpose of categorizing the curtain profiles could suggest that two possible sea/bay breeze events should be in the same cluster. In highlighting the differences between the two cases, we reassure that the developed cluster method is still useful and actually proves the method's ability. We have updated Lines 633 – 634 to reflect this.

Lines 633 – 634: "Analyzing their full curtain profiles, it is easy to conclude why these events were not assigned to the same cluster and the differences are also apparent in the individual model performance."

Lines 628 – 637: "It is evident from these cases that differences in sea/bay breeze events can lead to diverse O_3 profiles. The HLO case has high O_3 levels that reach down to the surface, with peaks > 75 ppb at both 12:00 and again at 16:00 EDT. Just above this extreme O_3 plume at 2000 m, there is an O_3 deficit of almost 50 ppb. The MCO case differs in that the highest O_3 concentrations do not reach the surface. Also, O_3 is more distributed and mixed throughout the curtain profile and the vertical gradient, although present, is not as stark as the HLO case. The HLO cases also has higher O_3 captured aloft above 2500 m which is not captured in the MCO case. Analyzing their full curtain profiles, it is easy to conclude why these events were not assigned to the same cluster and the differences are also apparent in the individual model performance. For both cases, the models generally seem to underestimate wind speed and overestimate O_3 (to different extents) but the GEOS-Chem performance in the HLO case is different. The uniqueness of this case implies that GEOS-Chem struggles to simulate this sea/bay breeze based on factors other than wind speed and direction."

The manuscript is too long and many sentences are redundant. There are a lot of repeated description in the main text, such as the description about the advantage of the lidar measurement and cluster approach, and the models' performance in low- and mid-levels. This will reduce the readability of the paper. I suggest to remove some redundant sentences. In addition, the

conclusion is also too long. I suggest to simplify the conclusion and only convey the key information.
We agree with the reviewer's comment and have removed redundant sentences throughout the manuscript and simplified the conclusion (please see Section 4 for updates).

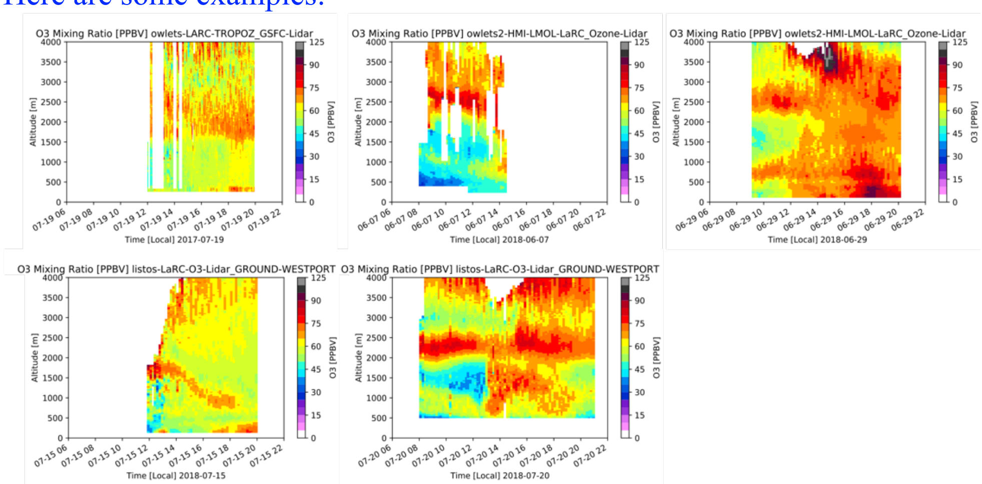
Other comments:

Line 335-336: "the HLO cluster reveals the specific case in which higher O3 is captured early in the temporal profile in the low-level and translates to the higher O3 captured in the low-level as well". This sentence is ambiguous. Please rephrase it.
This sentence has been updated for clarification.

Lines 335-226: "For example, the HLO cluster reveals a unique low-level case in which elevated (~ 1000 m) high O_3 concentrations are captured early in the temporal profile that translate to the higher O_3 concentrations at the surface later in the evening."

Lines 339-340: how do you infer that the cluster HMO indicates concentrated residual layers in the mid-level. Can you provide any evidence?
Evaluating individual profiles from HMO, there are some cases that show a residual layer aloft. These cases reveal layers of higher O_3 concentrations above 2000 m earlier in the temporal profile.

Here are some examples:



Due to lack of vertical velocity and vertical velocity variance data, we are not able to say whether or not these layers contribute to higher concentrations later in the profile near the surface as we state in Lines 342 – 344.

Lines 369-370: I suppose the low vertical mixing may reduce the descending O3 from above level, leading to lower low-level O3 concentrations.
Yes, this is true. We have added this to provide more clarification to the statement.
Lines 369-370: "…which can reduce any possible descending O_3 from aloft."

Lines 370-371: "Relatively calm wind speeds and lower temperatures indicate other possible meteorological factors such as high cloud cover that could have contributed to the lower O3

concentrations in LLO". This sentence is unclear because lower temperature will also lead to lower O3 production.
This sentence was reworded.
Lines 370-371: "Relatively calm wind speeds, lower temperatures, and other possible meteorological factors such as high cloud cover could have contributed to the lower O_3 concentrations in LLO."

The title of Sec. 3.3.3 is not appropriate. You mainly discuss the potential causes that influence the model performance capturing the clusters' O3 levels. A better title should be considered.
Section 3.3.3. Advantages of cluster approach and derived model conclusions

Line 511: Do you mean "despite having a low correlation in other cases"?
The overall correlation between the observations and the GEOS-CF for the mid-level is 0.22 (Figure S7) which is a poor correlation. When we calculate individual cluster correlations, the LMO cluster is revealed to have a good correlation. So, although the average mid-level correlation for GEOS-CF was low, there can still be some cases in which the GEOS-CF can reproduce O_3 pattern in the mid-level. This is an emphasis on how the use of a clustering approach to evaluate model performance rather than overall, summarized approach is beneficial in this case.

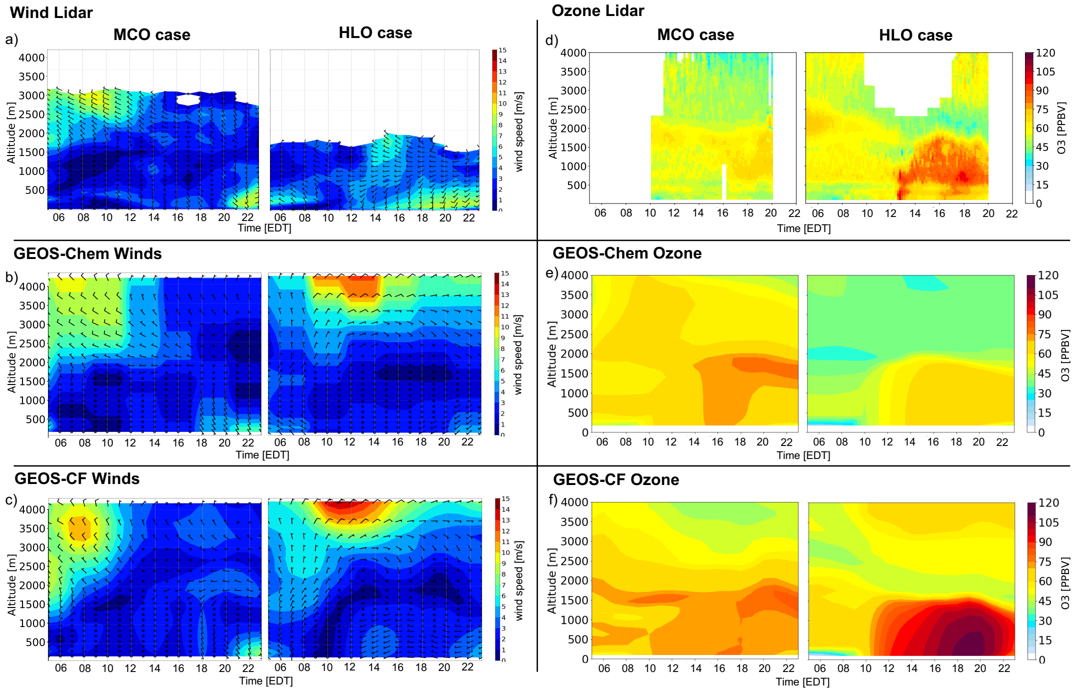
Line 511: "An even more profound case is exposed in which GEOS-CF has a strong correlation with mid-level O_3 in the LMO cases despite having a low correlation overall and in the other clusters."

Line 526-527: "In HLO alone, there were 4 (out 527 of 18) of the profiles that were consecutive while in MCO there were 8 (out of 28)". This sentence is unclear. What do you mean by the word "consecutive".
We used the consecutive to describe curtain profiles that were following one after the other in order, e.g., multi-day events. This was better clarified with a change to Line 527: "Given multiple cases of multi-day or consecutive high O_3 events from the lidar measurements…"

Lines 536-547: The structure of this paragraph is weird. Since you mainly discuss the situation in low level, it is not appropriate to discuss the mid-level situation. It is better to move these sentences to the next paragraph where you mainly focus on the mid-level O3.
We agree with the reviewer and have split the paragraph up. The first half of the original paragraph is now included in the paragraph before (Lines 527 – 533) and the second half of the original paragraph is now included in the paragraph after (Lines 534 – 544).

Fig. 8. The symbol of wind direction is weird. Ordinary arrows are better to indicate wind direction. The arrows in panel (a) are too dense and the color is unclear. In addition, I didn't see the shift of wind direction from westerly to easterly winds in the MCO case. I suggest to define the meaning of arrow direction in the Figure legend.
The wind symbols are not arrows but wind barbs. I have fixed the density of the barbs to make them clearer to see. The shift in the MCO cluster is slight and is earlier in the morning at about 06:00 EDT.

New Figure 8:



Line 597: “led” should be “led to”.
Done.

Lines 599-600: It seems that easterly winds prevail in the early morning and shift to northerly winds in the afternoon at the low level.
The wind symbols are wind barbs which always depict where the winds are coming from. Therefore, for the HLO cluster the winds begin as northwesterly – westerly winds that briefly shift to south – southeasterly winds at about 11:00 EDT before shifting to south – southwesterly winds.

Figure 8. Profile curtains of wind speed/direction (a-c) and ozone (d-f) from the lidar (top panel), GEOS-Chem (middle panel), and GEOS-CF (bottom panel). Results from OWLETS-2 at HMI. Wind direction is depicted by wind barbs.

Line 630: I do not see that the MCO case has higher afternoon O3 concentrations captured above 2000 m than the HLO case in Fig. 8d. They seem similar. In Fig. 4, the MCO cluster has lower afternoon O3 concentrations captured above 2000 m than the HLO cluster.
This line has been removed. What we meant to highlight was the lower O_3 concentrations just at/above 2000 – 2500 m in the HLO cluster, not necessarily that the entire mid-level had lower O_3. As it is explained more thoroughly in the line after (Line 631) we have removed the Line 630 to avoid the confusion.

In addition, does the white color in Fig.8d represent 0 ppb of O3 or missing data?
The white in the Fig.8d is indeed missing data in the measured curtain profiles. The description for Fig.8 was edited to address this.

Figure 8. Profile curtains of wind speed/direction (a-c) and ozone (d-f) from the lidar (top panel), GEOS-Chem (middle panel), and GEOS-CF (bottom panel). Results from OWLETS-2 at HMI. Wind direction is depicted by wind barbs. The white spaces indicate missing data for both the a) wind and d) lidar curtain profiles.

Line 581: "The white spaces in both the wind and O_3 lidar indicate missing data."

Line 660: GEOS-CF has an overall lower unsystematic bias range.
Done.

Reviewer #2:
Major comment:
The authors have made revisions that have improved the manuscript in response to the two reviewer comments. In addition to the specific comments below, the authors still need to clarify how the lidar data is better than ground, aircraft, or sonde data. There is no discussion in the abstract of why lidar data is needed, and in some of the discussion, it appears that surface data would suffice. I would support publication after the authors improve their discussion of how the ozone vertical structure information is beneficial over other data, and make that clear in the abstract and conclusions, and in their example discussions.

Specific comments:

Page 12, line 336 – The word 'background' in the ozone literature often refers to what ozone would be without anthropogenic influences. Consider using a different word to avoid confusion. The HLO curtain has lower free tropospheric ozone than the HMO profiles but higher surface ozone. So, it is likely that surface production is the most important factor here exacerbated by entrainment.
The line has been changed to avoid confusion.
Line 336: The mean profile curtain indicates these cases did not have "clean air" to begin with which can allow a greater accumulation in the low-level in the afternoon.

Also, the following sentence is confusing. Did you mean to say "low-level" twice? "For example, the HLO cluster reveals the specific case in which higher O3 is captured early in the temporal profile in the low-level and translates to the higher O3 captured in the low-level as well."
This sentence has been updated for clarification.

Lines 335-226: "For example, the HLO cluster reveals a unique low-level case in which elevated (~ 1000 m) high O_3 concentrations are captured early in the temporal profile that translate to the higher O_3 concentrations at the surface later in the evening."

Line 354 – Are the cluster average temperatures only at the LIDAR locations? Or across the whole domain?

There were not stations exactly where the LIDARs were located so we used the stations nearest to the locations of the LIDAR placements to represent the meteorology at the surface during the campaign dates (shown in Figure 5). Line 351 is updated to specify this.

Line 351: "To support the lidar clustering results, daily averaged meteorological surface observations from AQS stations nearest to the lidar locations pertaining to the campaign period and GEOS-Chem surface model output were evaluated in regard to the five clusters."

Line 366 – Please remove 'slightly', if it is not a real difference don't discuss.
Done.

Line 380 – There is higher ozone at the surface in the MCO and HLO curtains in Figure 4. Why is MDA8 ozone highest in HMO? If you calculate MDA8 ozone from Figure 4, do your results agree? Or are the regulatory sites missing these elevated ozone concentrations observed by the lidar?
This is a very good observation. Indeed, the AQS stations seem to not capture the same elevated ozone that is observed by the lidar. Since the lidars were in some cases placed to capture "over water" concentrations, many (about half) of the higher O_3 curtain profiles assigned to the HLO cluster were from the lidar stationed "over water". These measurements were not replicated by the nearest AQS station.

Lines 381 – 382: "Since the AQS stations applied here were just the nearest station to the lidar placements, the MDA8 O_3 captured by the AQS stations do not necessarily reflect the high O_3 concentrations capture by the lidars."

Figure 7 – Some of the mid-level correlations don't look statistically significant at all.
We agree with this statement, hence why we emphasize how the models struggle to capture mid-level O_3 pattern. The cluster approach allows us to catch cases in which the models poorly simulate with extremely low correlations e.g., not statistically significant, but also cases where the model performance have statistical significance even though they may be low correlations.

Line 439 – Are these numbers percentages? Is +0.30 actually +30%? Clarify that these are mean normalized biases.
No, they are not percentages – they are calculated mean normalized biases. This is clarified throughout the manuscript and in the Supplementary Material.

Line 408 – 409: "The mean normalized biases for the five clusters displayed in Table S1 (in Supplementary Material)…"
Line 435 – 436: "…(subsequent cluster calculated normalized biases and correlation can be found in Table S1)."
Table S1. Calculated mean normalized bias and…

Line 447 – You just told us that GEOS-Chem underpredicts high concentrations, but then say this challenges the assumption that models struggle to capture extreme cases. This seems to be a contradiction. It could help to quantify whether GEOS-Chem underpredicts at a certain

percentage, say the 90th, or 95th etc. percentile. You could also consider adding a probability distribution function to clarify your points.
We stated in Line 447 that the GEOS-Chem model performs well in capturing extreme events in the low-level with "slight tendencies" to underpredicts high concentrations. The model even has the lowest biases in the cases compared to the rest of the clusters. Table S1 shows the mean normalized biases for these cluster range from R = 0.07 – 0.09. This are very low mean normalized biases. The model is not perfect but in comparison to the other clusters and the GEOS-CF model, the GEOS-Chem model is able to capture these extreme cases. This challenges the assumption that models struggle to capture extreme events.

Line 447: "These results suggest GEOS-Chem actually performs well in cases of high O_3 as well as cases of low O_3 with only a slight tendency to overpredict lower O_3 concentrations and underpredict higher O_3 concentrations."

Line 504 – This statement again would definitely benefit from a probability distribution plot.
We thank the reviewer for their suggestion, but we believe that the statistics provided in Table S1 support the statements made in this manuscript well. Also, the visual representation of the model – observation spatial O_3 difference in Figure 7 provides additional support for the conclusions derived.

Line 530 – The conclusion about multi-day events is a great one to consider including in the abstract as an example of how lidar data can help models. However, it needs an explanation about why surface ozone data would not be sufficient. The discussion seems to be mainly about the surface, so it is unclear why the vertical structure is needed here.
We have emphasized the advantage of using lidar curtain profiles is evaluating multi-day O_3 events.

Lines 513 – 516: "Overnight, O_3 concentrations from the evening before can remain lingering in the residual layer. This residual layer sits at about 1000 m or higher depending on the conditions of the environment. O_3 trapped in this residual layer can directly correlate with the next day afternoon O_3 (e.g., Figure 3a; HLO cluster). If this residual layer and the titration of O_3 overnight is not successfully simulated, next day simulated O_3 will most likely warrant even greater biases."

Lines 524 – 527: "Full vertical and temporal curtains provided by lidar instruments are essential in fully understanding the development and depletion of O_3 in these cases. The mean curtain profiles in Figure 3a indicate that what is captured at the surface (below 500 m) in the early morning does not represent what is captured in the residual layer (1000 m) by the lidar. Therefore, surface data would not be sufficient in evaluating a multi-day event."

Lines 664 – 666: "This feature is attributed to multi-day O_3 events where O_3 left in the residual layer can contribute to higher O_3 in the afternoon the next and proves to be a challenge for CTMs. Lidar curtain profiles prove to be essential in evaluating these multi-day cases as they can capture the full development and deposition of O_3 in the residual layer that is not provided in surface evaluations."

Line 572 – It seems that the best use for lidar would be in simulating elevated surface ozone that appears to be from transport and entrainment. I am surprised there isn't more discussion of this application.

Indeed, this is a great use of lidar measurements, and we have included a brief analysis on transport (Figure S8). We used flight data from OWLETS-2 to evaluate the model's ability to simulate CO in the free troposphere (Figure S8) to investigate transport. We evaluated just 6 flights (morning and evening flights) all which profiles belong to HMO, MCO, and HLO clusters. From this analysis we find that CO is simulated well in the free troposphere at lower levels (100 -110 ppbv) which is indicative of background levels of CO. On the other hand, the model handles higher levels (130 – 140 ppbv) of CO worse, with consistent underestimations in the free troposphere. Since higher levels of CO in the free troposphere are indicative of long-range transport from outside the study region, we can conclude that the GEOS-Chem model struggles to simulate transport in the free troposphere which directly influences the mid-level performance.

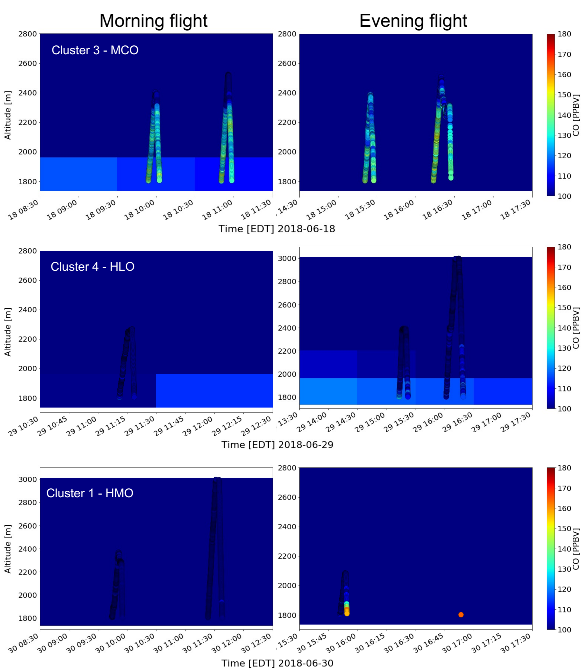


Figure S8. Aircraft measurements and GEOS-Chem simulated CO in the free troposphere during OWLETS-2. Measurements from the UMD Cessna 402B Research Aircraft.

Lines 517 – 523: "Additionally, transport of O_3 in the free troposphere (FT) is another influential factor that could contribute to the misrepresentation of mid-level O_3. In Figure S8, aircraft measurements from OWLETS-2 are used to evaluate GEOS-Chem simulated CO in the FT (1800 – 2500 m AGL). The flight days evaluated are all curtain profiles that were assigned to the clusters with higher levels of O_3 in the mid-level (HMO, MCO, and HLO). It is evident that the model is able to capture lower levels of CO in the FT (100 – 110 ppbv) but struggles to capture the higher levels (130 – 140 ppbv). Since increased levels of CO in the FT are indicative of possible long-range transport (Neuman et al., 2012), FT transport could be a factor contributing to the GEOS-Chem poor performance in the mid-level."

Lines 636 – 637: “Another factor inhibiting the poor simulation in the mid-level is the model failing to capture long-range transport of emissions in the FT.”
Entrainment is briefly mentioned in the manuscript as advantages of lidar measurements, but we do not go into details of cases found in this study since we would need more data such as vertical velocity data and velocity data variance data to support this.

Supplement.
Missing a ‘.’ In the paragraph before Table S1 between ‘needed’ and ‘These’.
Done.