Review of Bernier et al., 2022

This paper provides a unique method for taking advantage of ozone lidar data for evaluating models. This could be a valuable contribution and appropriate for ACP, however the description of this method and the model evaluation need major revision. As presented, the model evaluation using lidar data does not suggest anything to improve about the models other than increased resolution. Overall, the paper is lacking in explanations for the model failures in representing the five ozone clusters. If model resolution is the only clear factor, the authors should apply their method to regional, high-resolution modeling done for LISTOS (see comments below) at least. The model evaluation also does not provide much insight into how the clustering method is superior to a simple average comparison of model vs. observations. This study would be significantly strengthened by clear examples of how their clustering example provides specific insights over a simple average comparison. The authors could also better describe the benefits of lidar data beyond other types of profiles (e.g. ozone sondes, aircraft profiles) and show specific examples of these benefits.

Major comments.

Clustering algorithm

The k-means clustering algorithm needs better explanation. Did the authors choose the 8 features and then confirm they best represented the data with k-means? Did they try different numbers of features? What was the rational for looking at the data in this way? What about day to day variability, driven by different synoptic? I do not understand how the clustering algorithm was applied to the eight features. The authors might consider a diagram that shows how the 8 features lead to the 5 clusters.

Analysis

In the discussion of modeled vs. observed meterology, key findings should be discussed, not 'slight differences'. The readers are interested in what your model/observational comparison reveals about missing model processes that would be important to simulating surface ozone, particularly exceedances. Focus on highlighting those results and remove discussion about minor features.

Model selection

As the authors conclusion is that neither model has sufficient resolution to capture the seabreeze, this study would greatly benefit from including the regional modeling done for LISTOS (and possibly OWLETS?). WRF-Chem modeling was done for LISTOS. I suggest contacting NESCAUM to ask for this output.

Minor comments.

Line 58 – You say "set out to address this issue" twice.

Line 165 – Could you please further explain this sentence "Input features (seed values) were rationally established.."

Line 223 – Can you describe whether using the 40 complete profiles before datafilling was performed would give similar results? It seems somewhat problematic to fill the data based on observed patterns and then cluster the results also on observed patterns. Giving a little more information on why this approach is valid would be useful.

Line 288 – By temporal variation, do you mean diurnal variation?

Line 299 – Discuss Fig. 3a first or switch the order of the panels in Fig. 3.

Fig. 3a – It would be more informative to separate the profiles with altitude into day and night, or 12pm vs. 6am. Fig. 4 shows how the observations and models are both better mixed between roughly 12-16 EDT than during other hours.

Line 310 – Can you examine each of the 5 curtains and tell us for sure whether this is the reason? Or you could provide a standard deviation version of Figure 4 that would help us understand the cluster variability.

Line 319 – Give the cluster definitions earlier in the text before the introduction of Table 1.

Line 329 – Are the clusters spread across the three campaigns? Describe how each campaign contributes to each cluster.

Table 1 – Consider including Tmin and Tmax, and WSmin and WSmax. They could just be in parenthesis instead of separate columns.

Line 330 – What do you mean by "…could demonstrate background O3 in the case studies"?

Line 373 – When you show comparisons to the lidar data for GEOS-CF, are you only including lidar data clusters from OWLETS 2 & LISTOS?

Section 3.3.1 - The authors should use surface ozone monitors to determine whether ozone exceedances of the NAAQS occurred during any of the clusters. This would provide greater weight to the analysis of poor model performance for a given cluster.

Line 381 – The statement "In Figure 6, we first evaluate the overall relationship and correlation between both models and the lidar data, disregarding the specific clusters" is confusing as Fig. 6 is split into the 5 specific clusters.

Figure 6 – By "Spatial O3 difference", are you just referring to the differences in the vertical and in the diurnal cycle?

Line 403 – Do the models simulate higher ozone due to insufficient vertical resolution and/or excess vertical mixing? Is there anything to be learned in the only large model underestimate at the surface (GEOS-Chem, HLO)?

Line 404 – Are the lidar observations averaged to the model vertical (and temporal) resolution?

Line 427 – Is this a typo, do you mean "positive percent biases at 13.9, 18.9, and 19.7 %" instead of "positive percent biases at 0.139, 0.189, and 0.197 %? The other bias % values also look like they might be decimal values that need a 100 multiplier.

Line 448 – You state "Using the clustering, we are able evaluate how the cluster specific differences reveal additional model performance insight that would be conceivably overlooked when evaluating overall performance." Please give actual examples of how clustering is better than just "to simply group data by altitude to achieve a summarized model evaluation." A clear description of the benefits of clustering over the approach would greatly improve this discussion.

Line 464 – Does the model overestimate ozone on the first day of these multi-day events?

Figure 7 – Please make the limits on the x and y axes the same and add a 1-1 line.

Line 487 – Models are not intended to simulate 'intricate details', but rather the patterns that lead to high/low ozone at the surface. Could you rephrase to discuss how lidar data can contribute to that effort? What can the lidar data provide that surface ozone and sonde data cannot, and give specific examples of why this matters.

Line 498 – What "additional model performance insight" have you given us? Be more specific.

Figure 8 – Why not show a difference plot similar to Fig. 6?

Line 621 – As these models were run with emissions not provided specifically for the years 2017 and 2018, it might be more informative to look at normalized bias patterns as opposed to absolute biases. As this study is attempting to use the lidar to uncover areas of poor model performance with a focus on coastal meteorology, this approach would remove the impact of emissions not suited to the simulation year.

Line 631 – Throughout the paper, ‘slightly’ better results are not worth describing. Please focus on the most important, high-level results. For example, the finding that the models perform most poorly against the most common cluster is useful. Why do the models do better in cases other than the MCO? Is it because the sea-breeze is the most common pattern and this is most difficult for the models to capture? If so, this is a useful finding and should be more clearly stated.

Line 659 – You state, “Using the cluster assignments, we are able evaluate how the cluster specific differences reveal additional model performance insight that could be conceivably overlooked when evaluating overall performance.” Be specific about the insights you have revealed. The current manuscript is not clear about what the major findings are from the manuscript, nor what the most relevant conclusions are for air quality models.

Data availability – The authors need to provide the data links for the observational data used in this study. The authors could also consider providing their clusters as a data product for model evaluation.