General comments:

1. Since the chemical and dynamical processes controlling ozone concentrations in the stratosphere are quite different from those controlling near-surface ozone, what is the rationale for performing the GMM classification on the entire ozone profile? Could you instead cluster different vertical regions, such as stratosphere or troposphere, separately? It seems like the results might be easier to interpret and the clusters more applicable to model comparisons of specific features like surface concentration if signals from near-surface processes weren't mixed together with signals from stratospheric circulation in the creation of the clusters.

Thank you for this suggestion. In our revised approach, we have omitted the near-surface values before carrying out the classification step, which has indeed given us a set of classes that are somewhat easier to interpret. The highest pressure in our analysis is now 850 hPa. Our approach does still combine some tropospheric values with some stratospheric values, but the influence of the near-surface has been greatly reduced. As an extra advantage, our new classification approach is able to cover a larger fraction of the global model domain, as we have not had to discard profiles with surface pressures less than 1000 dbar.

Although it could be insightful to carry out completely separate stratospheric and tropospheric clustering analysis, we are interested in characterizing the entire ozone profile in the mid-to-upper troposphere and the entire stratosphere, because such an approach makes use of more of the data and will therefore be more general. Analyzing the entire profile also gets around the problem of having to decide exactly where to separate the troposphere from the stratosphere; given that the tropopause is gentler in some regions than others, the boundary between the two is not always easy to unambiguously determine. In addition, there are not very many pressure levels in the stratosphere, which would complicate any attempts to cluster there. We hope that you find our solution of discarding the near-surface pressure levels to be acceptable.

2. What is the advantage of using the GMM clustering method over just grouping profiles by e.g. tropopause height or altitude of peak ozone, since these seem to be prominent features distinguishing the derived classifications? It is encouraging to see that the GMM analysis leads to results that are consistent with known sources of variability, but to justify the complexity of this GMM approach, it would be helpful to also highlight specific cases where the GMM creates a more meaningful classification than could be obtained with a single variable such as tropopause height.

GMM takes the entire structure, within the selected pressure range, of all ozone profiles in the training dataset into account, which makes it more general than sorting or grouping the profiles based on a single value (e.g. tropopause height). Single-value classification schemes would have to rely on ad-hoc decisions about “cutoff” values between one group and another (e.g. low tropopause vs high tropopause). Although statistical distributions may offer some guidance there, generally speaking, we can’t expect such ad-hoc, specific cutoff values to be applicable across many different numerical model experiments, especially when considering different future climate scenarios, over which the cutoff values separating one class from another may shift. In contrast, by taking the entire structure of ozone profiles into account, GMM offers an approach that is insensitive to specific choices about ad-hoc cutoff values. In addition, the fact that, in our revised analysis, the classes vary with season makes them more useful and informative than a simple latitude-based classification scheme.

As you have said, the fact that GMM returns an intuitive, reasonably interpretable set of classes across several different numerical experiments indicates that it is generally applicable; we don’t have to select cutoff values. In addition, although we have not made much use of this fact in this manuscript, the GMM approach is probabilistic; it returns a set of probabilities across classes that could be used to examine boundaries between classes in a probabilistic fashion, which has been done in the Southern Ocean (e.g. Thomas et al., 2021, https://doi.org/10.5194/os-17-1545-2021). This could be an avenue for future exploration.

Specific Comments:

1.Lines 46-49: Please provide a reference.

Thank you for the suggestion.

We have added several references here:

However, an acceleration of the Brewer-Dobson circulation (BDC) associated with increasing greenhouse gas concentrations may lead to reductions in lower tropical stratospheric ozone mixing ratios (Eyring et al., 2013; Meul et al., 2016; Keeble et al., 2017) while increasing transport of ozone into the mid-latitudes troposphere (Banerjee et al., 2016; Meul et al., 2018).

2. The discussion of previous work on ozone clustering could be expanded.

We have expanded the discussion. Now the discussion reads:

Clustering techniques have already been used in ozone concentration studies for understanding long-term variability. Boleti et al. (2020) have applied a multidimensional clustering technique to understand the long-term trend of ozone. Diab et al. (2004) used a six clusters analysis which resulted in distinct clusters of "background" and "polluted" with below and above ozone mixing ratios from over 100 ozonesonde profiles launched from a subtropical Southern Hemisphere Additional Ozonesondes (SHADOZ) (Thompson et al., 2003) site, Irene, South Africa. Jensen et al. (2012) performed a cluster analysis named self-organizing maps (SOM) (Kohonen, 2012) on over 900 tropical ozonesonde profiles. Their findings with four-cluster results were similar to Diab et al. (2004). Both studies showed that the seasonal influences of biomass burning and convection dominate ozone variability. Stauffer et al. (2016) documented the influence of meteorological conditions on the shape of the ozone profile from the troposphere to the lower stratosphere by applying the SOM clustering technique to ozonesonde data from specific northern hemisphere midlatitude geographical regions. Later they expanded the study for global ozonesonde sites to show the variation of ozone profiles cluster for various regions and how they vary based on meteorology and chemistry depending on latitudes (Stauffer et al., 2018).

3. Lines 97-98: The requirement of surface pressure reaching 1000 hPa seems like a significant limitation. Would the results be much different (and the coverage increase) if you used something like 900 hPa instead?

Yes, thank you for this excellent suggestion. We have discarded pressure levels above 850 hpa from our data set. Now the results are based on pressure levels in the range 1-850 hpa pressure level. The resulting clusters cover more area overall and are easier to interpret.

4. Line 130: How does the pressure level standardization affect the relative importance of the stratospheric versus the tropospheric portions of the profile in determining the clusters?

Ozone concentration measured from different pressure levels does not contribute equally to the analysis because of large differences in tropospheric and stratospheric ozone values, and this might end up creating a bias in the algorithm. The primary purpose of standardization is to put ozone concentrations from different pressure levels on the same scale.

In standardizing ozone on each pressure level separately, we are allowing variations in the stratosphere to have the same impact as variations in the troposphere relative to the usual variability found on each pressure level. If we didn't do this, then our classes would simply be determined by the pressure levels on which the variability is highest, in absolute terms.

5. Line 149: Define BIC and refer the reader to the description in the appendix.

We changed the statement.

6. Lines 194-201: Do the higher tropopause and higher surface values both contribute to the definition of this cluster, or is it just that the clusters vary strongly with latitude (as shown in Fig. 4) and many other features also co-vary with latitude?

It is important to note that the algorithm does not have any information about the latitude of the profiles. Our implementation of GMM only uses the ozone values themselves. The classification will indeed be influenced by the entire structure of the ozone profile, especially since the ozone profiles have been standardized on each pressure level. In any case, now that we have excluded the near-surface values, the near-surface no longer has an influence on the classification.

7. Line 216: is mPa the right unit here?

We used mPa everywhere in this study to keep consistency.

8.Line 219: Replace “reasonable” with something more quantitative

Changed to “known”. This isn’t more quantitative, but we hope that you find it suitable.

9. Fig 4 (and 5) and Fig 4 caption: Does “median” make sense with respect to classifications here? Are the classes quantitatively ordered such that class 3 is in between classes 2 and 4? Also, is there much temporal variability (within the decade) in what class a particular grid box falls in? If so, it would be nice to show that since it could help clarify how the GMM classification differs from a purely latitude-based classification.

The label maps have been changed; we now examine seasonal variability.

10. Line 249: Does this mean the fact that class 1 has the lowest ozone or the fact that the class 1 ozone is lower in the historic run is consistent with the reduction with precursors?

It is consistent with the reduction of precursors. We have changed the statement to:

As with the historical experiment, class 1 has the lowest surface ozone (Table 1), which is consistent with the reduction in surface ozone precursors in this experiment. The maximum value of stratospheric ozone increases under this scenario, which is a signature of the recovery of the ozone hole (Keeble et al., 2021).

We hope you find it suitable.

11. Lines 272-275: Is this explanation proven by your analysis or just consistent with your results?

It is consistent with our results. We used references from studies that have shown these.

12. Lines 284-285: Please explain how this conclusion is reached from Figs 4-5

We have changed the statement since we are now doing the seasonal averages.

13. Lines 295-297: Is it possible to relate this quantitatively to the extent of the model’s Hadley cell?

Possibly, although an in-depth analysis of the circulation of the model is beyond the scope of this technical note. We only aim to highlight and demonstrate this ozone classification approach.

14. Lines 298-304: Are these results different from what would be inferred with latitudinal averages?

These results should be consistent with latitudinal averages. What is different here is we did not use the latitudinal information for our algorithm, still our result is consistent with what we expect from latitudinal averages.

15. Line 321: This statement needs more support. Relate to Table 4?

We agree that this statement was vague, and we have removed it from the paper.

We have made the technical changes to the manuscript. We wish to thank the reviewers for helpful insights. Your comment helped us to improve our manuscript.