We thank the reviewer for their time reviewing the submitted manuscript and for their insightful comments. Please find our responses to each comment below - original reviewer comments in bold, author responses beneath. In addition to addressing the comments from the reviewer, it has also been possible to include in the analysis presented in the revised manuscript 8 additional CMIP6 models that have made available diagnostics since the original submission. Ozone and water vapour data is now available from the AWI-ESM-1-1-LR, CESM2-FV2, CESM2-WACCM-FV2, E3SM-1-1, MPI-ESM-1-2-HAM, MPI-ESM1-2-HR, MPI-ESM1-2-LR and NorESM2-MM models. We have added co-authors to the manuscript from research groups that have prepared and made available the data from these additional models. The inclusion of these models does not change the conclusions of the papers, nor does it significantly change the CMIP6 multi-model mean for the historical period or projections under different SSPs. However, including these models in the revised manuscript gives a more complete evaluation of available CMIP6 models.

Reviewer 3:

General Comment

This paper evaluated CMIP6 models in terms of their ability to simulate past to future variation in the stratospheric ozone and water vapor. The authors examined how well the CMIP6 models represent stratospheric ozone profile and past change in total column ozone (TCO) by comparing with the observation in global and regional perspective. They showed the considerably diversified future projections of stratospheric ozone with different SSP scenarios. The stratospheric water vapor in CMIP6 models was revealed to have a couple of problems for adequately simulating its observed features partly because some models ignore methane oxidation in the stratosphere. I fully acknowledge the importance of this work and it will provide a useful information to the climate science community. This paper is rightfully within the scope of ACP, however, I noticed several issues in this paper which cannot be passed over to be published. I suggested that the authors should consider the following comments: two major and several specific comments.

Major Comment 1:

The authors relatively well described the difference among CMIP6 models, pointing out some models showing over/under-estimation. However, they only provided a limited discussion and description about possible reasons for such a spread among models. As a result, the current manuscript ended up being a superficial model intercomparison. I recommend the authors to spend more words to discuss why some models differ significantly from the other models. For example, could you discuss more about why UKESM1-0-LL model greatly overestimate the TCO, GFDL models underestimate, MRI-ESM2-0 showed quite a small temporal change in TCO from 1950 onward, and the like?

We have added a new section (section 3.4) to the manuscript comparing models with interactive and prescribed ozone fields, which further details the differences between the models. However, the largest ozone differences are seen between models with interactive chemistry and identifying the causes of these biases is a significant challenge, requiring access to the models' source code and preforming a number of tests. Further, significant differences exist between models with prescribed ozone fields, which is likely related to how those models process and implement the CMIP6 ozone dataset. As such, identifying why a model has a significant high or low bias with respect to the CMIP6 multimodel mean is a significant challenge and beyond the scope of this study. Instead, we aim to evaluate how TCO and stratospheric water vapour compare to observations in these CMIP6 models, and how they are projected to change from the pre-industrial to the present day, and into the future, so that this information may be used when comparing, for example, radiative forcing and regional climate change across the CMIP6 models evaluated here.

Major Comment 2:

There is several important information which were not properly provided in the manuscript.

[1] Description of CMIP6 models in chapter 2.1 should provide more unified information on each model. The current descriptions differed considerably between models. At least, 1) model

resolution (horizontal and vertical), 2) treatment of ozone-related chemical process both in stratosphere and troposphere and 3) CH4 oxidation in the stratosphere should be provided for all models. Or it's better to include that information in Table 1.
There is no set definition for what makes a model, for example, an Earth system model and so it hard to standardise the text across the 22 models evaluated here. Some models, for example, include some Earth system components (e.g., atmospheric chemistry, ocean biogeochemical cycles, carbon and nitrogen cycles, interactive fire schemes), but exclude others. Similarly, some models use fairly advanced chemistry schemes with 10s of chemical species and 100s of reactions, while others use simplified schemes. As such, we have worked with the individual modelling groups to write the text for the model descriptions so that the modelling centres are happy with the description of their model and point towards the full documentation for each model so that the readers can further investigate which components are included in each model and the complexity of these components. We have expanded table 1 to include more information on each of the models evaluated here to cover the reviewers 3 points.

[2] Many statistics used in the manuscript are not well defined. The authors should carefully describe the statistics in the manuscript and figure captions. Sometimes I could not understand what kind of temporal and/or special means were used in some figures, since the descriptions in the manuscript or figure captions are so rough and blur. The lack of carefulness like this largely deteriorate readability and value of the manuscript. The author should carefully revise the manuscript and figure captions to provide sufficient description about the statistics used. I noticed several points in the following specific comments section.
Greater care has been taken to describe better the statistics and averaging used in analysis throughout the manuscript and in the figure captions.

Specific Comments:

- L80-81: Heterogeneous chemistry is also important.
This information is included in the last line of this paragraph, which states 'Heterogeneous processes play a major role in determining ozone 85 abundances in the polar lower stratosphere (e.g. Solomon, 1999) and following large volcanic eruptions (e.g. Solomon et al., 1996; Telford et al., 2009)'

- L99-100: Could you briefly describe how BDC control the oxidation of Cly, NOy and HOx species.
The BDC controls the oxidation of these species by transporting source gases (such as CFC-11, CFC-12, N2O and CH4) from the lower atmosphere, where they are chemically inert/have long lifetimes, to the middle and upper stratosphere, where they are more rapidly oxidised. The faster the BDC the greater the mass flux through the region of oxidation, and so the shorter the chemical lifetime of the species. We have modified the sentence to read '…and by influencing the chemical lifetimes of Cl_y, NO_y and HO_x, source gases (e.g. Revell et al., 2012; Meul et al., 2014).'

- L104: Why you didn't use the abbreviation "SWV" here? Please uniformly use the words which you dare to define throughout the manuscript.
We have revised the manuscript to use 'stratospheric water vapour' throughout.

- 2.1 Models:
[1] The models are described as "fully coupled", "online", or "interactive" chemistry. You should give precise description what these words mean. If there are no difference among them, you should use one specific word to describe it. I suppose all these words mean that calculated chemical species concentration is used in, so "coupled" or "interactive" with, the radiation calculation. Is it correct? Are there any model who calculate the chemical species "online" but they are not used in radiation calculation? Are there any models who only calculate stratospheric or tropospheric chemistry online

but used prescribed concentration in the other sphere? Could you clearly describe these details of each model in this chapter or summarize in Table 1 ?(Please also see Major comment 2 [1])
We have added more information on each model to Table 1. Please see also our response to your major comment above.

[2] As to prescribed chemistry models, it should be clearly stated how these models treat the chemical species concentration in the model. I could not understand why these models output different ozone concentration as depicted in the Figures even if they prescribed the same CMIP6 dataset, although I know CESM2 used their original ozone data and so could output different ozone fields. This is one of the key points for the readers to correctly understand this paper. Particularly, it should be described for each model whether the model prescribed concentrations entire model domain or only prescribed at the surface and allowed to calculate the atmospheric concentrations online.
This is a key finding for the manuscript – that models prescribing the CMIP6 ozone dataset do not agree in terms of zonal mean ozone mixing ratios and the total column. This is likely due to the numerous regridding phases the ozone fields go through. The ozone field is first regridded from the resolution of the CMIP6 ozone dataset provided by Input4MIPs (96 lats x 144 lons x 66 pressure levels) to the native model grid. The model out is then regridded to the 19 pressure levels used in the Amon output grid specified by CMIP6. Additionally, there are processing steps the models employ (redistributing the ozone dataset to match the model tropopause, prescribing only stratospheric ozone but modelling interactively tropospheric ozone) which also result in differences between the models prescribing the CMIP6 ozone dataset. However, while these processes clearly play an important role, it is beyond the scope of this study to identify where these differences arise from. Instead, we document these differences and argue that, when prescribing ozone fields, greater care should be taken to ensure that the total ozone column is conserved.

- L152: Appendix shows the relevant "difference" among models but do not provide the "details" of each model.
This has been changed to 'Relevant details of each model are provided below, and a summary is provided in Table 1.' This error came about as the model discriptions in section 2.1 were originally in the appendix, but were moved before submission. However, this sentence was not updated to reflect this.

- L273: How did each model force the historical changes in short-lived species (mainly air pollutants and its precursors) and long-lived GHG? Whether were they input as emission or surface concentration?
Many models which include interactive chemistry schemes use a mixed approach here, often using emissions for short lived species with high spatial variability (e.g. NOx, CO, VOCs) while prescribing surface concentrations of long lived (e.g. CO2, CFC-11, N2O). We do not aim here to give a detailed description of how each model is set up and the processes they include, simply because this would be a huge task and repeat information that is available elsewhere in the published literature. Instead, we sought to provide key information on how the models treat stratospheric ozone and water vapour and also provide references for each of the models evaluated in this study should more detailed information be required.

- L287: What are "low" SSPs?
For clarity, this has been changed to 'low numbered SSPs (i.e., SSP1 and SSP2) assume lower abundances of long-lived GHGs'

- L298-305: Any abbreviations should be spelled out at their first appearance, NIWABS, SWOOSH and satellite sensors names.

This has been done in the revised manuscript.

- L322: What is (10o)?

This should say '10° latitudinal resolution' and has been corrected in the revised manuscript.

- L343-345: CESM2 and FGOALS-g3 models showed larger overestimation than BCCEMS1. Why did you particularly pick up these two (BCC-ESM1 and SAM0-UNICON) models here? Also, SAM0-UNICOM does not have peaks in the mid-latitude.

These two models were singled out as having not just high biases in the upper stratosphere but also different spatial patterns, with significant biases particularly in the mid latitudes. To clarify, we have changed these sentences to 'Notable differences between the models occur in the uppermost stratosphere, and around the tropopause (Figure A1). In the upper stratosphere, the BCC-ESM1, CESM2, CESM2-FV2, FGOALS-g3, NorESM2-MM and SAM0-UNICON models all simulate much higher ozone mixing ratios than the CMIP6 MMM. Additionally, the BCC-ESM1 and SAM0-UNICON models also have a different spatial structure in the distribution of ozone at these levels, with maxima in the mid-latitudes at 1 hPa'

- Figure2: What is the shaded region?

The shaded region represents the standard error. As discussed for the major point above, we have gone through the manuscript and added more detail on the statistics used in the manuscript and added more detail to the figure captions.

- L358: It's hard to distinguish each model's line, so I'm not quite sure that I could tell BCC-ESM1 correctly, this model was not low-biased, but "negatively" high-biased. SAM0-UNICON model is also negatively high-biased.

Low biased is used here to indicate that the modelled values for the BCC-ESM1 and SAM0-UNICON models are lower than both the observations and CMIP6 MMM.

- L362-363: Differences of the CMIP6 MMM from the observation described here in the lower stratosphere and over 1 hPa in the mid-latitudes are not mutual among all CMIP6 models. From Figure A1 it is clear that these differences are mainly owing to only a few models. So the author should describe here more carefully. Putting a figure of standard deviation among models together in Figure 4 might be an option.

We have added more text to the discussion of Figure 4 stating that while the MMM differences are as described, they arise from different biases from each model contributing to the CMIP6 MMM.

- Figure3: It's better to use different color pallet to make it more easy to identify the difference among models.

The YlOrRd colour pallet is used throughout this study as it is i) colour-blind friendly, and ii) the colour pallet identified for use by IPCC, and as such has been recommended for analysis of CMIP6 models. We agree with the reviewer that many of the panels look the same, but that is a positive – most CMIP6 models accurately capture the magnitude and seasonal evolution of TCO. However, differences can be seen – SAM0-UNICON and MRI-ESM2-0 have shallower ozone holes, while the CNRM models have lower springtime arctic polar ozone. This differences are further explored in the subsequent analysis, but figure 3 is intended to demonstrate that no model so significantly misrepresents TCO as to be a clear outlier.

- L380-381: Is the MMM TCO underestimation in SH polar region in polar winter real? The NIWA-BS data in this area in this season is mainly made by filling missing data as I correctly understand chapter 2.3 of the manuscript, so it might be artificial not real. Can you compare the MMM TCO with other data source, such as ground based TCO observation in SH polar region?

As the reviewer states, in the region of the polar night, where large baises are seen between the CMIP6 MMM and the Bodeker dataset, the ‘observations’ rely on a filling routine described in the manuscript. To prevent over-interpretation of this difference we have changed Figure 4 to use the unpatched version of the Bodeker dataset, and modified the text accordingly.

- Figure 5: Figures are a bit small and hard to recognize each symbol. Could you provide the detailed description how did you calculate statistics used in these Figures? It is not self-apparent what “spatial std dev” or “percentage bias” mean. There are several definitions to calculate those statistics. The descriptions can be in appendix or as a supplement material.

- L394-395: Why does a large “spatial” standard deviation for the SMA0-UNICON and MRI-ESM2-0 models indicate higher interannual, so “temporal”, variability? (Please also consider my comment for Figure 5)

- L406-407: Why do only these two models show no interannual variability? How about other models who used prescribed ozone fields? (Please also see the comment for “2.1 Models” [2])
We are not sure why these models show no interannual variability, and have passed this observation on to the relevant modelling groups.

- Table2:
[1] The number of ensemble member for each model should be summarized in Table1.
This information has been added to table 1.

Moreover, it must be described somewhere in the manuscript how the ensemble member was treated in all the analysis for this paper. Did you use ensemble means for all the figure?
Yes, we use all available ensemble members for each model to create a single ensemble mean for that model and then this mean is shown in all the figures. The CMIP6 MMM is then created from the individual model ensemble means. This process was adopted to prevent a model with many ensemble members dominating the ensemble mean – instead each model counts equally towards the MMM. This has been added to section 2.

[2] What does “errors” exactly mean?
It is the statistical uncertainty of the trends at 68% (1 sigma) confidence level. This has been added to the footnote of Table 2.

[3] Could you also provide the trend of observation (NIWA-BS) for 2000-2014?
Calculating the trends for the observations and comparing those accurately with the models is a challenge due to the issues surrounding TCO in the polar night. The models provide full lat lon domains throughout the year, and so the global mean annual mean trend is truly global. For the observations we have the option of using the patched NIWA-BS dataset (which is effectively using a model to fill in the values in the polar night), or the unpatched version, which has missing data in this region. Either way, we would not be comparing like with like, and as the recovery trends are so small (and generally not statistically significant) these differences are relatively important. A detailed evaluation of trends in the NIWA-BS dataset can be found at:
Bodeker, G. E., Nitzbon, J., Tradowsky, J. S., Kremser, S., Schwertheim, A., and Lewis, J.: A Global Total Column Ozone Climate Data Record, Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2020-218, in review, 2020.

- L411: What does "overall TCO decline" exactly mean here?
This has been changed to ‘the decrease in TCO between 1980 and 2000’

- L430: The modelled trends in TCO for 2000-2014 are small but not mostly nonsignificant.

We have edited this paragraph to say 'Over the period 2000-2014, generally, models show non-significant (at the 95% confidence level) positive trends in TCO. However, nine models show significant albeit weak positive trends in global TCO, of which three are INTERACTIVE models (CESM2-WACCM, MRI-ESM2-0, and UKESM1-0-LL). The significant positive trends calculated in these models show the largest positive trends in both the NH and the SH high latitudes and moderate positive trends in mid-latitudes (Table 2). The INTERACTIVE models collectively show stronger positive trends in all regions, compared to the all-model mean. Significant and the strongest positive ozone trends in the SH high latitudes occur in MRI-ESM2-0, NorESM2-MM, and UKESM1-0-LL, whereas significant and the strongest positive trends in the NH high latitudes occur in CESM2-WACCM, NorESM2-MM, and UKESM1-0-LL. Significant but weaker positive trends also occur in SAM0-UNICON and CESM2 at SH high latitudes. Here, the significance is the consequence of small variability in those models without interactive chemistry.'

- Figure7 (and for some other figures): Why did you use "standard error" not "standard deviation" for indicating the model spread? The standard deviation is appropriate for this purpose.

The standard error (of the mean) is shown in all the line plots to better represent the uncertainty associated with the CMIP6 MMM. In contrast, the standard deviation provides a measure of the spread about the mean, and we feel that this can be appreciated by seeing the individual models which comprise the MMM.

- L440: I could not see the TCO increase of "20-30DU" from 1850 to 1960 in NH in Figure 7. Could you revise the number?

We thank the reviewer for pointing this out – the correct range is 10-15 DU. This change has been made in the revised manuscript.

- L440-441: English is too complicated for me to understand correctly what it means.

This sentence has been modified to now read 'In the NH, TCO values increase by 10-15 DU between 1850 and 1960. This increase in TCO is larger than the TCO depletion that occurs from 1960 to 2000 in response to the emission of halogenated ODSs, resulting in higher NH mid-latitude TCO values in the late 1990s than in the pre-industrial.'

- L446: How did you evaluate the "ability" of models to simulate pre-industrial TCO? Since we don't have TCO observation in that era, we cannot ensure the model's ability through comparing model results with observation.

This has been changed to 'There is poor agreement in the simulation of pre-industrial TCO across CMIP6 models'

- L448-454: Could you make more discussion about the difference among the models in simulating the past TCO changes. Discussions on why the models prescribing CMIP6 ozone data showed such a large discrepancy and those on the overestimation of ozone decline by some models are desirable.

We have added a new section to the manuscript (section 3.4) comparing models with prescribed vs interactive ozone fields which explores in more detail the historic changes in TCO across the models.

- Figure 8: Is the SSP370 scenario simulation result necessary for this figure. This part was never referred in the manuscript.

We feel that inclusion of the SSP370 scenario in this figure gives important information about the future changes to stratospheric column ozone in the models explored in the figure, highlighting the fact that future TCO increases are driven by significant increases in upper stratospheric ozone, while lower stratospheric ozone increases occur much more slowly (due to increases at high latitudes being offset by decreases in low latitudes, see our Figure 10.). We have added more text to the discussion of Figure 8 to include this point and cover the SSP370 scenario projections shown in the figure.

- L458: Typo. "TCO seen in Figure 8" -> Figure 7
Corrected

- L460: Where did you describe about a large tropospheric ozone bias of UKESM1-0-LL in the manuscript? Which figure show that?
In the manuscript we state that 'It is also clear from Figure 8 that much of the high TCO bias for the UKESM1-0-LL model (Figure A2) comes from elevated stratospheric ozone mixing ratios, rather than a large tropospheric ozone bias.' While we do not plot the tropospheric ozone column, we see in Figure 8 that the UKESM1-0-LL model has a significant stratospheric ozone bias which is contributing to the TCO bias. If Figure 8 had shown UKESM1-0-LL stratospheric partial columns in agreement with the other models then the only location that would be left would be the troposphere. So the argument being made here is based on inference rather than being directly shown.

- L490: Is this for SSP1-1.9 scenario not for SSP1-2.6?
No, we mean here SSP1-2.6 as TCO values do not return to the 1960 or 1980 baseline in the northern midlatitudes under SSP1-1.9.

- L515: Could you add the changes in the BDC simulated in the CMIP6 models to Figure 10?
Unfortunately, this cannot be done as the models have not widely output and made available the diagnostics required to do this. However, the reviewer makes an interesting point about explicitly identifying the changes in stratospheric circulation in CMIP6 models, and it is hoped that as more diagnostics become available in the future this can be done in other studies.

- L529-530: Figure A4 should be cited here if you want to refer to the percentage difference, since Figure 11 cannot show it.
This has been done.

- L535-536: How is the temperature at the tropical tropopause in the CNRM models? Is there any low temperature bias there which can cause the dry bias in the stratosphere in those models?
The CMIP6 model data used here is only available on 19 pressure levels, and many models do not provide the tropopause pressure as a diagnostic. As a result, it is very difficult to determine the cold point tropical tropopause temperature for individual models. We agree with the reviewer that it is likely that the low H2O mixing ratios seen in the CNRM models in the lowermost stratosphere is associated with a cold tropical tropopause. However, it is not possible to state this definitively, and so we do not make this conclusion in the manuscript.

- L538" As for ?
This has been changes to 'As with'

- L555: What does "CH4" exactly means in this equation? Concentration? Mixing ratio? What is its unit?
Mixing ratio – this sentence has been amended to read 'the tropical stratosphere H_2O mixing ratio will equal 7.0-2.0*CH_4 mixing ratio'

- Figure 14: There are no reference to Figure 14 in the manuscript. The figure capture does not include the description of color bar. What does each point in the figure represent? Are they annual mean? Horrible lack of information for this figure.

The discussion of Figure 14 is in the final paragraph of section 4.1 – we have now explicitly referenced the figure here. The figure caption has been significantly expanded, and now reads 'H2O vs CH4 scatter plots of the six CMIP6 models for which both H2O and CH4 mixing ratio are available from the historical simulation. The data shown here is monthly mean, zonal mean H2O and CH4 mixing ratios (in ppmv) for the years 2000-2014. The coloured shading of the points represents the altitude (in hPa). The black line gives gradient for all model points above 70 hPa, while the dashed black line gives SPARC estimate (H2O = 7-2*CH4)'.

- Figure 15: Why you did not comment anything on the comparison with the observation. The CMIP6 models apparently underestimate the observation and the modelled increasing trend in the stratospheric water vapor can not be seen in the observation. You should discuss about those comparison in the manuscript.
We have added more text describing Figure 15, particularly focusing on how the models compare to the SWOOSH observations, to the revised manuscript.

- L580-582: How is the temperature change at the tropopause not at the 100 hPa? Is the increase at 100 hPa temperature and the increase in water vapor quantitatively consistent with each other?
It is a significant challenge to calculate the temperature change at the model tropopause rather than at a given pressure level as few models have provided data on the pressure/height of their tropopauses and the vertical resolution of the data used here would make calculating a lapse rate tropopause from the temperature fields inaccurate. Generally, we can see that the 70 hPa H2O mixing ratios agree well with the 100 hPa temperatures by comparing Figures 15 and 16, but of course these annual mean temperatures at a fixed pressure level are only a very rough estimate of the tropical tropopause cold point temperatures, and so identifying a quantitative relationship between them (for example based on Clausius–Clapeyron) is a significant challenge.

- L588-590: Could you show separately the relative contribution of 100 hPa temperature rise and CH4 concentration increase for the stratospheric water vapor increase in the CMIP6 models? Both for historical simulation and future projections.
This is an interesting point, but very difficult to do as we lack the diagnostics to achieve this. Only a handful of models have output CH4 mixing ratios, and even from these models it is clear that there are significant differences between the amount of water vapour formed per molecule of CH4, even for models which report including CH4 oxidation, and so we cannot generalise across the models which do not provide CH4 mixing ratios.