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 1:

The manuscript presents an analysis of the evolution of ozone and stratospheric water vapour from the pre-industrial to the present-day (2000 – 2014) and out to 2100 from a number of coupled chemistry climate models that were submitted to CMIP6. In addition, the present-day distribution and seasonal cycle of ozone and water vapour from these models are compared to a number of observational datasets. While the factors controlling the projected evolution of ozone and water vapour seen in these simulations are well known, the presentation of CMIP6-era chemistry climate model simulations is an important update of the literature. In particular, the future projections for the new set of CMIP6 scenarios (SSPs) is welcome.

The paper is well written and the presentation of the results is clear so I do not have any significant concerns about the content. I will point out that one difficult aspect of the overall presentation is the mixing of models of varying complexity in their representation of atmospheric chemistry. There are five models that include what could be considered a fully prognostic representation of chemistry, with others using specified ozone or linearized chemistry. To complicate the situation further, a couple of the models with specified ozone are closely related to models with prognostic chemistry, having derived ozone from one of the five models with prognostic chemistry in different ways. It makes interpretation of what exactly the multi-model mean represents a bit difficult to fathom. The authors have been generally clear in the description of the models and it is easy to figure out which models contain a full representation of chemistry and which do not. But I would suggest a few minor modifications to help the reader better understand the composition of the multi model ensemble and where the models with prognostic chemistry are significantly different than the other models.

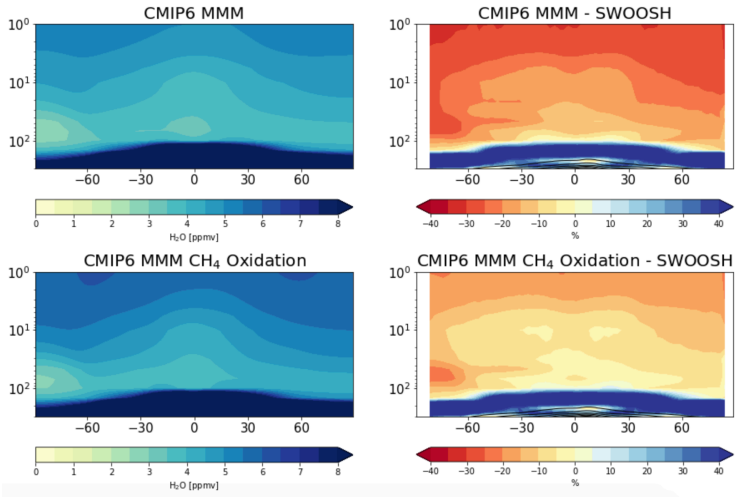
For one, there are a number of models that specify stratospheric ozone using the CMIP6 dataset and yet, see lines 341 – 343 and lines 448-449, some of the models that use the specified CMIP6 ozone show significant differences with each other. Is it possible to include the CMIP6 ozone dataset in a few of the figures comparing the different models? Having the zonal average ozone from CMIP6 dataset shown in Figure 1 and the difference to the multi-model mean in Figure A1, would be very helpful. Perhaps Figure 5 as well?

We have added the CMIP6 ozone dataset to Figures 1 and A1. We have also added the CMIP6 ozone dataset to Figure 3 and created a new Figure (A2) which shows the climatological (2000-2014) total column ozone differences between each model and the CMIP6 ozone dataset with respect to the CMIP6 MMM. Together, these figures give a clear indication of how different the CMIP6 ozone dataset is to the CMIP6 MMM. Overall, there are only modest differences between the CMIP6 ozone dataset and the CMIP6 MMM, consistent with the large number of models included in this analysis which prescribe the CMIP6 dataset dominating the MMM.

Secondly, starting at line 538 there is discussion of water vapour in the models, including the behaviour of the CMIP6 multi-model mean that is shown in Figure 12. Given that a number of models do not include a chemical source of water and the fundamentally different behaviour that omitting methane oxidation produces, as seen in Figure 11, I would suggest defining the MMM in Figure 12 as only including those models that include the chemical source of stratospheric water. You really are mixing apples and oranges when you take all ten models that provided water vapour outputs and included them in the MMM. This is much less of a problem for the remaining plots that focus on lower stratospheric water (70 hPa), but for the zonal cross-sections I would suggest either a separate plot of the mean of only those models that include CH4 oxidation or redefining the MMM plot to only include those same models. I would also suggest a similar segregation of models for Figure 17 since the zonal cross-section of PI to PD changes in water vapour will be fundamentally different depending on whether or not the models account for a chemical source of water vapour. Is this

approach already used in Figure 18, where only five of the 10 models are used to construct the multi-model means of water vapour shown there?

The reviewer raises an important point about the inclusion of models without a water vapour source from CH4 oxidation in the CMIP6 MMM. Excluding models from the multi-model mean is always a challenge, and highlighting that the CMIP6 MMM has lower water vapour mixing ratios in the upper stratosphere, and that this bias is arising from the fact that some models do not include CH4 oxidation, is an important conclusion, and one that climate and Earth system models can use in the future development of their models. As a result, we have left the CMIP6 MMM in Figure 11 as the mean of all the available models. However, to address this comment, we have added a second row to Figure 12 which shows the CMIP6 MMM, the mean of models which include some method of accounting for water vapour formed from the oxidation of CH4, and the differences for both of these means with respect to the SWOOSH dataset. In this way, the revised manuscript now shows both the CMIP6 MMM using all models, and the mean using only models with CH4 oxidation as the reviewer suggests. This new figure is included here and replaces the original Figure 12 in the revised manuscript.



As can be seen from this new Figure 12, there is a steeper vertical gradient in water vapour mixing ratios, and the upper stratospheric bias is reduced with respect to the SWOOSH climatology, but even when models with no treatment of CH4 oxidation are explicitly excluded from the mean, water vapour mixing ratios in the upper stratosphere remain smaller than those in observations. This discussion has been added to the text in the revised manuscript.

For the reviewer's final comment, only models which have run all four SSP scenarios are used for Figure 18, rather than sub-setting of the models based on processes. This was done in order to prevent differences between the future projections arising from the inclusion of different models in each projection.

My other comments are minor and are itemized below. Lines 205-206: 'form' should be 'from' in 'instead prescribed form the CMIP6 dataset.'

This has been corrected.

Line 259: The text states that the SAM0-UNICON model uses specified ozone but doesn't say what the source of the data is: 'Stratospheric and tropospheric ozone is prescribed as a monthly mean 3D field with a specified annual cycle.' Is it CMIP6?

Yes, the SAM0-UNICON model prescribes the CMIP6 ozone dataset – this has been added to the model description for SAM0-UNICON.

Lines 341 – 343: A couple of models are found to have large differences in ozone in the upper stratosphere relative to the multi-model mean, as shown in Figure A1. In particular BCC-ESM1 and FGOALS-g3 are singled out for having much higher concentrations of ozone in the upper stratosphere, but both of these models base their ozone on the CMIP6-specified ozone dataset in some manner. Does this indicate problems with the CMIP6 ozone dataset or problems in how the models used the data?

This issue is related to how models treat ozone in the top boundary levels. We can see from the updated Figures 1 and A1 that there are no significant problems with the top levels of the CMIP6 ozone dataset. However, what

is clear from the analysis presented in the revised manuscript is that models prescribing the CMIP6 ozone dataset are not conserving local ozone mixing ratios of the total column, despite using the same forcing file. Where these differences are coming from is a challenge for each of the modelling centres prescribing the CMIP6 ozone dataset and beyond the scope of this paper to evaluate. We have added to the conclusions of the revised manuscript a few sentences on the differences seen between models ostensibly prescribing the same ozone and highlighted that it is a challenge for the global modelling community to do this in a more robust manner. These sentences state:
'Models which prescribe stratospheric ozone from the CMIP6 ozone dataset show surprisingly large variation in TCO, particularly in the pre-industrial period, at which time there is a ~20 DU range in pre-industrial TCO values between those models prescribing the CMIP6 ozone dataset. There are also large percentage differences between zonal mean ozone fields output by the individual models and the CMIP6 ozone dataset, likely connected to the treatment of the top boundary conditions in different models. Together, this evidence suggests that TCO is not conserved after model implementation of the CMIP6 ozone dataset, and instead small differences are introduced between the models. A future challenge for modelling centres is to prescribe ozone concentrations in such a way as to preserve local mixing ratios and the total column abundance.'

Lines 632 – 633: 'However, there is poor agreement between the individual CMIP6 models in the pre-industrial and throughout the historical period, with model TCO values spread across a range of ~60 DU.' To make this clearer I would suggest adding a few words along the lines of 'However, there is poor agreement between the individual CMIP6 models for the absolute magnitude of TCO in the...'
This change has been made, and the sentence now reads 'However, there is poor agreement between the individual CMIP6 models for the absolute magnitude of TCO in the pre-industrial and throughout the historical period, with model TCO values spread across a range of ~60 DU.'

Line 1268: The caption on Figure 8 should state that the average is over 90S – 90N.
This information has been added to the Figure caption.