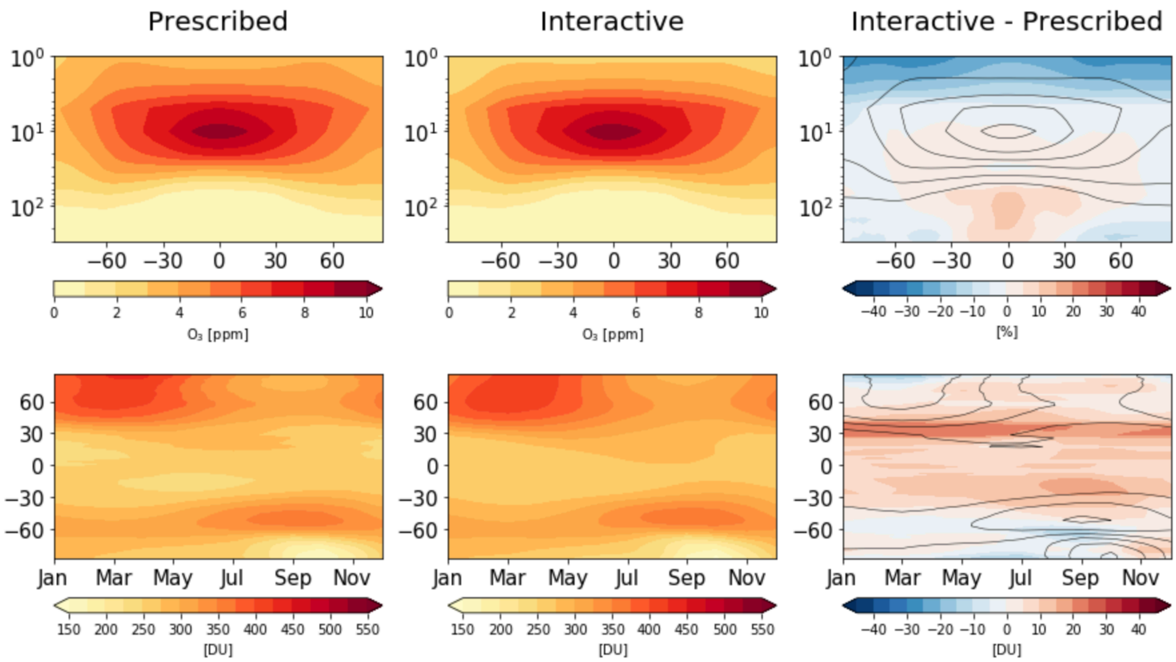
We thank the reviewer for their time reviewing the submitted manuscript and for their insightful comments. Please find our responses to each comment below - original reviewer comments in bold, author responses beneath. In addition to addressing the comments from the reviewer, it has also been possible 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 2:

The paper presents a comparison of ozone and H2O fields between different CMIP6 models. It also includes their evaluation against observations. It provides a lot of information that is necessary for the analysis of the climate simulations, notably the contributions of stratospheric ozone and H2O to radiative forcing in CMIP6 coupled climate simulations. I recommend publication after addressing several points. I think the difference between models that calculate interactively ozone and those that specify it should be even clearer in the analysis. The spread among models with specified ozone reflects how important the implementation process of a given ozone climatology into a model can be whereas the spread in the interactive models is of a different nature. I would suggest to add the “"Mean (specified)” in the Tables and add where possible “CMIP6 MMM (specified)” And ‘CMIP6 MMM (interactive)” plots (e.g. Figure 4, 6). Here are below more minor comments.

We have added a new section to the manuscript exploring the differences between models with interactive and non-interactive ozone (section 3.4). As we discuss in that section, some conclusions can be reached regarding these two subsets of models. However, of the 22 models evaluated in this study, only 6 have interactive chemistry. Comparison is further complicated by the fact that the prescribed ozone fields are taken, in part, from a forerunner to the CESM2-WACCM model (in combination with fields from the CMAM model). As a result, significant caveats exist for the conclusions drawn when comparing models with interactive and prescribed ozone fields.

When we calculate a CMIP6 MMM (interactive) mean, it is not dissimilar to the MMM presented in the original manuscript. We have updated Figure A3 in the revised manuscript to include both the CMIP6 MMM and the mean of models with interactive chemistry. We have also produced a version of Figure 4 comparing the climatological (2000-2014) mean of models with prescribed ozone fields to the climatological (2000-2014) mean of models with interactive chemistry:



For the zonal mean, throughout much of the stratosphere differences between these model groupings are less than 5%, although large differences are seen in the uppermost levels and the tropical troposphere. Expressed as a total column difference, the mean of models with interactive chemistry is ~10 DU higher than the mean of models with prescribed ozone in the tropics and midlatitudes. This is most likely due to the inclusion of the UKESM1-0-LL model, which has a significant high TCO bias.

However, while the means are similar, this masks a large range of TCO and zonal mean ozone mixing ratios across the models with interactive ozone chemistry. We have added the Figure above to the revised manuscript, and included a new section (section 3.4) which discusses the figure and the points made here.

341-343: "Notable differences between the models occur in the uppermost stratosphere, and around the tropopause (Figure A1). The BCC-ESM1, CESM2, FGOALSg3 and SAM0-UNICON models all simulate much higher ozone mixing ratios in the upper stratosphere."

This has been amended to say '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.

L75: "which, while consistent with the IPCC-AR5 estimate, represents an increase of ~80% compared to the CMIP5 ozone forcing dataset". Why would a difference of almost a factor 2 indicate consistency between these 2 estimates?

The values of the radiative forcing estimation by R.Checa-Garcia et al, 2018 rely on the CMIP6 ozone concentrations dataset from input4MIPS (and the previous corresponding to CMIP5). These estimations are consistent with the multimodel estimation of IPCC-AR5 report. This is shown in the figure 1 of the paper (lines black and red of the figure on right panels).

https://agupubs.onlinelibrary.wiley.com/cms/asset/e1926f6c-7cf8-4104-b828-b5f183436083/grl57057-fig-0001-m.jpg

However, if we compare the radiative forcing of CMIP5 ozone concentrations dataset and CMIP6 ozone concentrations dataset the last one has an increase of 80%. In other words, the radiative forcing estimated with CMIP5 ozone concentrations dataset was not consistent with the multimodel estimation of IPCC-AR5 report, meanwhile with CMIP6 we have values close to that multimodel mean.

L77: "The relative uncertainties in radiative forcing estimates for both stratospheric ozone and water vapor are large due to the challenges in constraining the concentrations of both during the pre satellite era". That's not the only problem. If 2 modelling groups are provided with the same preindustrial and present-day fields for ozone and H2O, the radiative forcings calculated by the 2 groups would differ substantially, depending on their radiative schemes, on how the forcing is implemented (notably with respect to the tropopause adjustment), etc... (see literature about model intercomparisons including literature from some of the authors)

We agree with the reviewer that there are a number of challenges in calculating radiative forcing associated with ozone and stratospheric water vapour changes. This sentence has been amended to say 'The relative uncertainties in radiative forcing estimates for both stratospheric ozone and water vapor are large due to the challenges in constraining the concentrations of both during the pre-satellite era. As a result, the current radiative forcing estimates rely on ozone and water vapor fields derived from simulations performed by global climate models and Earth system models. However, models use different radiation schemes and model (in the case of models with interactive chemistry scheme) or prescribe (in the case of models without interactive chemistry schemes) ozone differently, further contributing to the uncertainty estimates.'

L80-86: I am not sure that this short history of stratospheric chemistry in the middle of the introduction is necessary.

We include this short section on stratospheric ozone chemistry as the paper is aimed at the CMIP6/IPCC community, which covers a diverse range of academic backgrounds. The results of this study are of interest, for example, to the radiative forcing community, which may not be familiar with the details of stratospheric ozone chemistry. As such, we feel the inclusion of these 6 lines provides broader background material, but is short enough so as not to disrupt the flow of the manuscript.

L100:" However, recent research (Polvani et al., 2018, 2019) has shown that stratospheric ozone depletion caused by increasing ODSs has accounted for around half of the acceleration of the BDC in recent decades." It is an estimation based on model simulations. However, there is no clear agreement or at least quantitative agreement between BDC changes in model simulations and observations (BDC tracer proxies). The authors should be more cautious: recent model simulations indicate. . .ODS may have accounted.

We have added this to the manuscript. The sentence now reads 'However, recent research (Polvani et al., 2018, 2019) has shown, using model simulations, that stratospheric ozone depletion caused by increasing ODSs may have accounted for around half of the acceleration of the BDC in recent decades.'

L110: "followed by a sudden decrease of ∼10% after 2000 (e.g. Solomon et al., 2010)." Can models reproduce it? If not, again, one should be cautious with model results regarding H2O evolution.

Solomon et al. highlight that the decrease in stratospheric water vapour after 2000 is associated with, among other things, anomalies in sea surface temperatures. As the models do not use historic sea surface temperatures, we do not expect them to have a similar decrease at the same time. However, the reviewer is correct to say that when evaluating stratospheric water vapour in Earth system models asking, 'do the models capture the observed annual to decadal variability in H2O mixing ratios?' is an important question. Here we focus on longer term (decadal to centennial) changes in water vapour, and so it is beyond the scope of this study to evaluate this fully, and we hope that other research on CMIP6 models will assess annual to decadal variability, and the drivers of this variability, in detail.

L159-163: Some caveats should be added here. The fields are "unique, consistent" but it is just 2 models! In view of the inter-model differences and differences with observations (see Figures in the

paper), there are biases in those recommended fields, large uncertainties, notably in the projections. Again more caution is required.
We have added the following text to the end of the section discussing the CMIP6 ozone dataset 'Not that, as the CMIP6 dataset uses values from model simulations, it has biases with respect to observations and uncertainties associated with the projections of stratospheric ozone beyond the period observations exist for, both from the pre-industrial to the start of the observational record, and from the present day to the end of the 21st century under the different SSP scenarios.'

L446: ". . .CMIP6 models to simulation (simulating) pre-industrial TCO".
This has been changed to 'There is poor agreement in the simulation of pre-industrial TCO across CMIP6 models'

L446-447: Add a comment. The large differences for preindustrial are not totally surprising. Models are tuned to reproduce the observed ozone evolution (during the satellite era), not the preindustrial era when no observations are available.
We feel it is not necessary to explicitly state that differences in the pre-industrial are not surprising – as the reviewer says models tend to better agree during periods of observations and differ outside of these periods. Whether this is an explicit tuning the models undergo or an unconscious result of the development of the chemistry scheme is harder to say. Additionally, while we may expect the models to disagree during the pre-industrial, the degree to which they disagree (spanning a range of 75 DU) perhaps is surprising.

L449: "Surprisingly, there is a ∼20 DU range in pre-industrial TCO values between those models prescribing the CMIP6 ozone dataset." Add: suggesting that the TCO is not conserved after model implementation.
This has been added.

L615: The conclusions do not differentiate between ozone interactive and specified models. The spread among models with specified ozone reflects how important the implementation process of a given ozone climatology into a model can be whereas the spread in the interactive models is of a different nature driven by differences between model schemes. The analysis provides important information on both approaches. I think the conclusion should also summarise the conclusions for each approach.
We have added a paragraph to the conclusions section summarising the new section which compares models with interactive and non-interactive ozone fields.