We thank Reviewer 3 for their positive and careful review. As with Reviewer 1 and 2, we appreciate the attention that has been given to our manuscript, particularly regarding their comments related to the cloud adjustment and the comparison with He et al., (2023). Our responses are given in the blue font. Where necessary, we refer to the revised version of the manuscript.

Recent studies have argued that a large component of the CO2 IRF spread in CMIP models can be explained by documented large spread in stratospheric base state temperatures. This manuscripts serves as an important follow-up, testing whether differences in stratospheric O3 can explain the documented stratospheric temperature spread and thus the CO2 forcing spread. The authors find the answer is no. It's a well written, interesting study. But before it can be accepted I recommend the authors address my comments below that touch on interpretation of results and providing more details/explanations in certain places.

General: I agree that Stratospheric O3 differences cannot explain the spread in 4xCO2 ERF, based on the results presented here. But I think the authors are too quick to discount the effect of differing Stratospheric O3 on the spread in IRF. As noted a few comments below, I'd argue that a slightly deeper dive into the He et al. Results suggests the two papers may be more comparable than the authors suggest.

Introduction Section: To address the relevance of this study's results to the question of CMIP spread, it would be helpful answer how realistic is the range of 0.5 StratO3 to 1.5 StratO3 relative to the actual range of StratO3 across CMIP models. Is there a considerable spread in Strat O3 across CMIP models? I was under the (maybe incorrect) impression that O3 is prescribed in CMIP atmosphere-only simulations. In which case, the hypothesis that Strat O3 spread explains the CMIP Strat. T spread would be wrong. Some comments/explanation along these lines would be helpful.

We have now added another paragraph to the introduction/methods to better frame our motivation related to the spread in CMIP6 models (lines 66-73 and 91-93).

"One such cause could relate to stratospheric O_3 – a key constituent in modulating stratospheric temperature. Depending on the treatment of stratospheric chemistry, models adopt a range of methods to generate O_3 fields using either an interactive chemistry scheme, a simplified online scheme or a prescribed pre-simulated dataset. Consequently, the resulting spatial structure and regional distribution of concentrations can differ substantially. Keeble et al. (2021) evaluate long-term O_3 trends in 22 CMIP6 models and find poor agreement in the simulation of pre-industrial total column ozone (TCO), with a variation from 275 to 340 DU between 60˚N-60˚S. Further, a ~ 20 DU range is observed between 10 of the models that prescribe stratospheric O_3 according to the CMIP6 O_3 dataset (Checa-Garcia, 2018), highlighting that even the model-specific implementation of common input can lead to significant differences in TCO."

"Considering the substantial range in pre-industrial TCO noted by Keeble et al. (2021, Figure A3), we choose such a large, idealised increase/decrease in attempt to cover a broader range of stratospheric O_3 than shown by CMIP6 models, thus any effect on 4xCO_2 ERF would likely be amplified in comparison."

Here we refer to the Keeble et al., (2021 https://acp.copernicus.org/articles/21/5015/2021/) study that evaluated long term historical ozone trends in 22 CMIP6 models (6 of which used interactive chemistry schemes, 3 use a simplified scheme, 10 use the prescribed CMIP6 ozone dataset and 3 use prescribed ozone values from CESM-WACCM simulations). Despite good agreement between the CMIP6 multi-model mean and observations, this study shows substantial variation between individual models (see Keeble et al., (2021) Figures A1 and A2). Across the 2000-2014 period they report that notable differences occur in the uppermost stratosphere and around the tropopause, and across 1960-

2014 they show that ozone fields are both overestimated and underestimated by models that use interactive chemistry. They further report poor agreement in the simulation of pre-industrial TCO, varying between 275 and 340 DU (see their Figure A3) and furthermore that there is a ~ 20 DU range in pre-industrial TCO between the 10 models that prescribe ozone using the same CMIP6 dataset. 11 of the models analysed by Keeble et al. (2021) are used in the Smith et al. (2020) study (in Keeble et al. 7 of these models use prescribed ozone, 3 use an interactive chemistry scheme and 1 uses a simplified online scheme). However, even if all of the Smith et al. models used prescribed O_3 in their atmosphere-only set-up (i.e., piClim-control), we conclude that there could still be spread in stratospheric O_3 based on Keeble et al's. findings. Although Morgenstern et al. (2020; Figure S1 https://doi.org/10.1029/2020GL088295) actually demonstrate that there is a large spread in TCO between models that use interactive chemistry in piClim-control simulations (including GFDL-ESM4, MRI-ESM2-0 and UKESM-0-LL which are used in Smith et al., (2020), see their Table 1). Thus, we further conclude that at least some of the atmosphere-only simulations do include full ozone chemistry.

Line 80-90. Although it's clear if you read table S1, I recommend the authors make it clearer in this section of the text that stratospheric O3 is reduced (or increased) by 50% in both the perturbed 4xCO2 simulation and in the corresponding control pre-industrial simulation, thereby ensuring the new O3 filds act only to alter the base state and not as a forcing adjustments itself.

Thanks for this point, we appreciate that this will make our experimental design clearer to the readers and we have the following sentence (lines 94-95): "*Note that in each ERF experiment the O_3 increase (or decrease) is applied to both the control and 4xCO_2 simulation so that the new O_3 field acts exclusively to alter the base-state atmosphere and does not act as a forcing itself.*"

Line 150-190: I appreciated the authors nuanced discussion about the application of radiative kernels for diagnosing the stratospheric adjustment. Their arguments logically make sense. But do we have any proof that their estimate of the stratospheric adjustment using the CESM kernel is more representative of the model's true adjustment compared to the previous uses of kernels applied in e.g. Smith et al.? The fact that the residual-derived cloud adjustment matches the alternative Smith et al method is maybe promising, but as a residual calculation, it is difficult to pinpoint potential canceling biases. For instance, it's possible both the stratospheric adjustment and some other adjustments have equal and opposite errors.

We think that our estimate of the stratospheric adjustment is more representative of our NorESM2 experiment's adjustment because we do not extrapolate our data to levels beyond the uppermost atmospheric level, and so do not include the adjustment in temperature that occurs beyond these pressure levels (as done by Smith et al. 2020). Also, to a lesser extent, because the CESM-CAM5 kernel is built using the same radiative transfer code as NorESM2 i.e., RRTMG. With regard to the cloud adjustment, please see our response below.

If we assume the authors are correct in their statement that it is best to use kernels from the host model, it would be helpful if they also gave a recommendation about how kernels should best be used when being applied across multiple models to evaluate inter-model spread.

Although its not a focus of this paper, a brief comment would be helpful since this is a common use of kernels and there is not currently a radiative kernel available for every host model.

We appreciate this point, and it could be an option to recommend not extrapolating model data to kernel levels that are outside of the model's native bounds (see lines 216-219):

"If a radiative kernel is not available for a given model, or a kernel needs to be applied across multiple models to evaluate inter-model spread, then it could be more suitable to not extrapolate data outside of each model's native vertical bounds. However, the best use of kernels is likely quite case specific."

Line 220-226: First, the authors state that their range in IRF across experiments of 0.8 W/m2 is much smaller than the 4 W/m2 IRF spread that He et al. Finds across the online double calls. This is true. But the authors should keep in mind that He et al. Does not claim all of that spread is due to the base state. They claim "more than half" (presumably the r^2 ~ 0.67 is what they are basing this on) of the spread is explained by the base state but not all of it. I recommend the authors factor this in when comparing the He et al. result to their own findings.

Further, I'd argue that a fairer comparison between the spread in this paper and the spread in He et al. would be a comparison to the offline calculations of their figure 1C (rather than their 1B) where the IRF spread is subject only to base state differences and not to differences in radiative transfer algorithms across models, as is the case in this present study. In the He et al. figure 1C it appears ~14 degreesK of 10 hPa stratospheric temperature spread across models corresponds to 1.3 W/m2 of IRF spread. Since the 10 hPa temperatures in this study range from -3K to +4K relative to the standard case (line 205), and this corresponds to a 0.8 W/m2 spread in IRF across the experiments, it would appear the StratT vs IRF spread results are actually quite comparable between the two studies from this perspective. Does this impact the overall conclusions that the authors would draw about the importance of StratO3 spread to IRF spread?

Reviewer 2 also raised this point and we agree that the comparison to Figure 1c of He et al. (2023) provides a better evaluation of our results against theirs. We now compare against the near 2 W m^{-2} spread in IRF from Figure 1b of He et al. (2023) and then discuss the similarity between our change in temperature at 10 hPa (between the 50% increase and decrease base-state) and the spread in IRF and Figure 1c from He et al. (2023) (see lines 261-269). Our conclusion remains the same with respect to the impact of stratospheric O_3 increases/decrease on 4xCO_2 IRF, i.e., that our idealised experiments show a dominant impact on the IRF magnitude.

Line 231-244: It is interesting to consider the relative importance of StratO3 effect on CO2 forcing through overlap with CO2 vs through stratospheric temperature effects. Do the authors have a relative sense of this? It's difficult to imagine a setup that could address this for ERF, but for IRF one could presumably perform an offline radiative transfer calculation with PORT where StratO3_x0.5 or StratO3_x1.5 is imposed but Stratospheric temperatures are prescribed in all cases from a StratO3_x1 climate as a way to isolate the spectral overlap effects from the stratospheric temperature base state secondary effects.

We previously performed LBL tests (using the GENLN2 code) to compare the relative importance of these effects. We found that a 50% reduction in O_3 leads to a +0.07 W m^{-2} increase in 4xCO_2 IRF (from 4.2 W m^{-2}), whilst a reduction in the temperature of -2K across the whole stratosphere leads to a 0.16 Wm^{-2} increase. Based on these results the effect of stratospheric temperature dependence is stronger than the effect of spectral overlap. Furthermore, as shown in Figure 2, decreasing stratospheric ozone by 50% results in widespread cooling of the stratosphere with ΔT values largely more negative than -3 K and peaking at -9 K. We have now added these results as a footnote on page 12 of the revised manuscript:

"Tests performed by the GENLN2 line-by-line (Myhre et al., 2006) show that a decrease in temperature of 2 K across the whole stratosphere leads to a 0.16 W m^{-2} increase in 4xCO_2 IRF, whilst a 50% reduction in stratospheric O_3 leads to a 0.07 W m^{-2} increase in 4xCO_2 IRF."

Line 253-255: There appears to be a misinterpretation here. The fact that A_Tstrat remains the same size across experiments would actually support the IRF enhancement/reduction extending all the way to ERF rather than prevent it (As ERF = IRF + A_Tstrat + other adjustments). In order for the IRF enhancement/reduction not to extend to ERF, there needs to be an equal but opposite compensation in the enhancement/reduction of a different adjustment. In Figure 3, this seems to occur largely through the cloud adjustment term. i.e. the IRF is larger than the standard experiment for the StratO3x0.5 case while the A_c is smaller than standard. Likewise the IRF is smaller than the standard experiment for StratO3x1.5 while the corresponding A_c is larger than standard. I recommend the authors rephrase this section to emphasize the A_c term changes rather than focusing on the static magnitude of A_Tstrat. I further recommend the authors explore why A_c has this apparent sensitivty to StratO3. It would help us understand whether the ERFs lack of sensitivity is due to the specific characteristics of stratospheric O3 or if ERF is just not sensitive to stratospheric temperature base states more generally.

This is a great point and we now realise our apparent oversight of the possible importance of the effect on the cloud adjustment.

However, following this comment (and similar points from Reviewer 1 and 2) we decided to change our method for calculating the cloud adjustment to the adjusted cloud radiative effect method of Soden et al. (2008; https://journals.ametsoc.org/view/journals/clim/21/14/2007jcli2110.1.xml). As shown in the updated version of Figure 3, the magnitude of the cloud adjustment in each experiment is now much more similar, demonstrating that this adjustment doesn't offset the impact of ozone increase/decreases on IRF and SARF as implied previously.