We would like to thank reviewer 1 for their positive and constructive review of our manuscript. We respond to each comment raised by the reviewer in the text below.

The introduction should contain a discussion about the uncertainty around the solar reconstructions and where it comes from, particularly in reference to the reconstruction from Shaprio et al used here. The study of Judge et al mentioned later should be discussed here as should the latest solar reconstructions made available for CMIP6 (as described in Jungclaus et al). In particular it is worth noting that the PMOD reconstruction which has still much larger amplitude than the Steinhilber et al reconstruction shown here, has less amplitude that the original Shapiro et al reconstruction and is described in Jungclaus et al as an "upper limit". I do not think that this in anyway invalidates the results shown in this paper – but they should be framed so that it is clear that the solar forcing used is now considered to have too large an amplitude.
Response: Thank you for this comment. We will modify the introduction to briefly discuss uncertainty in the solar reconstructions itself, as well as modify Fig 1b to show the more recent solar reconstructions from PMIP4.

I also think care is needed about the interpretation of the comparison of the model results to SAM reconstructions. For the main results the correlation to one SAM reconstruction is significant but isn't to the second reconstruction. I am not sure how to interpret this – particularly given the small number of degrees of freedom due to the strong 70-year filter used. How strong a result is this and how much does the result hinge on the low period of SAM in both reconstructions and models after 1400? Could one interpretation of the results be that the reconstructions have a very low SAM amplitude starting in the 15th century which cannot be explained by the model simulations unless a strong solar amplitude forcing is used? If so this seems to me to be a nice, strong and clear conclusion.
Response: Thank you for this comment. Indeed, our findings do suggest that the reconstructed minima in the SAM during the 15th Century can't be readily explained by model simulations unless they are forced with a strong amplitude of solar forcing (at least in the absence of interactive atmospheric chemistry). The reviewer's comment has prompted us to carry out two additional statistical tests to demonstrate that this is a robust conclusion of our work:

Firstly, we use a bootstrapping approach to randomly reorder the climate model simulated SAM data (number of iterations = 10000) to further assess the robustness of the correlation between the annual A14 SAM reconstruction and the annual Mk3L model SAM. Based on this, we find that the OGS-Shapiro ensemble mean SAM index is significantly correlated with the A14 reconstruction, as is the OGS-x2 ($p <$ 0.05 for both, relative to a random distribution of the model data). However, the OGS simulation is still not significantly correlated with the A14 reconstruction any more than could be explained by a random distribution of the model data. Similarly, the correlation between the annual D17 and bootstrapped OGS-Shapiro ($N = 10000$) is significant ($p < 0.05$), while it is not significantly correlated with the OGS-x2 SAM or the OGS SAM ($p > 0.05$).

To specifically explore how much the correlation is driven by the low in the SAM reconstruction and models during the 15th Century, we also remove the 1400–1500 portion of our time series and recalculate the correlation statistics between the A14 reconstruction and the OGS, OGSx2 and OGS-Shapiro SAM. We find that the OGS-Shapiro ensemble mean and A14 SAM reconstruction are still significantly correlated *(R = 0.52, $p < 0.05$)* even with the omission of the 15th century data, although we find the OGS ensemble mean is also significantly correlated to A14 (R = 0.52, $p < 0.05$), while the OGS-x2 ensemble mean is not (R = 0.39, $p > 0.05$). If we remove a larger window, i.e. 1300–1600, we find that all our simulations are significantly correlated to the A14 SAM reconstruction (OGS: R = 0.56, p < 0.05; OGS-x2: R = 0.56, p < 0.05; OGS-Shapiro: R = 0.8, p < 0.05).

Overall, this suggests to us that the low-frequency, very negative state in the SAM reconstruction and models during the 15th century is only somewhat driving the correlation in our stronger-solar simulations. We still find a significant correlation between the OGS-Shapiro and A14 SAM using a bootstrapping approach on annual mean data. Removing this interval of negative SAM (in both the simulations and reconstruction) actually improves the correlation across the OGS, OGS-x2 and OGS-Shapiro simulations, suggesting that the 1400s negative SAM state in the SAM reconstruction is a key factor in the lack of correlation between the OGS simulation and A14 reconstruction.

In our revised manuscript we propose to briefly add details on the bootstrapping tests described above to strengthen the findings already presented in the paper.

As well as the analyses shown here – one simple analysis which I would find very useful would be if you could show what the variability of the piControl simulations were (e.g. two standard deviations) – then you could quite easily make the point that the observed variability cannot be explained by internal variability (assuming that the model variability is correct).
Response: Great idea! Similar to Fig 3e in Abram et al. (2014), we show in the figure below the SAM index anomaly for each of the different CSIRO Mk3L runs (ensemble mean shown as thick lines, and the ensemble members as thin lines; including also for comparison the orbital-only [O], and orbital+GHG [OG] only simulations). The SAM indices are expressed as anomalies relative to the 1900–1999 mean and plotted as moving 70-year moving averages, and dashed lines to show the range of internal variability based on ±2σ from the orbital-only simulation. The OGS ensemble mean stays mostly within the range of unforced internal SAM variability during the last millennium, while the OGS-x2 ensemble mean has some intervals, including the 15th century, where the SAM is more negative than can be explained by internal variability alone. The OGS-Shapiro (panel e) ensemble mean SAM (and the individual ensemble members) is more negative than can be explained by internal variability prior to the 1700s, and the strong multi-century positive trend in the simulated SAM since the 15th century resembles this same characteristic described for the reconstructed A14 SAM (Abram et al., 2014). This demonstrates that the forced response from strong solar forcing on the simulated SAM cannot be explained by internal variability.

We propose to add this figure (before the current figure 7) and a short additional paragraph describing it to section 3.2 where we introduce the results of the transient simulations.

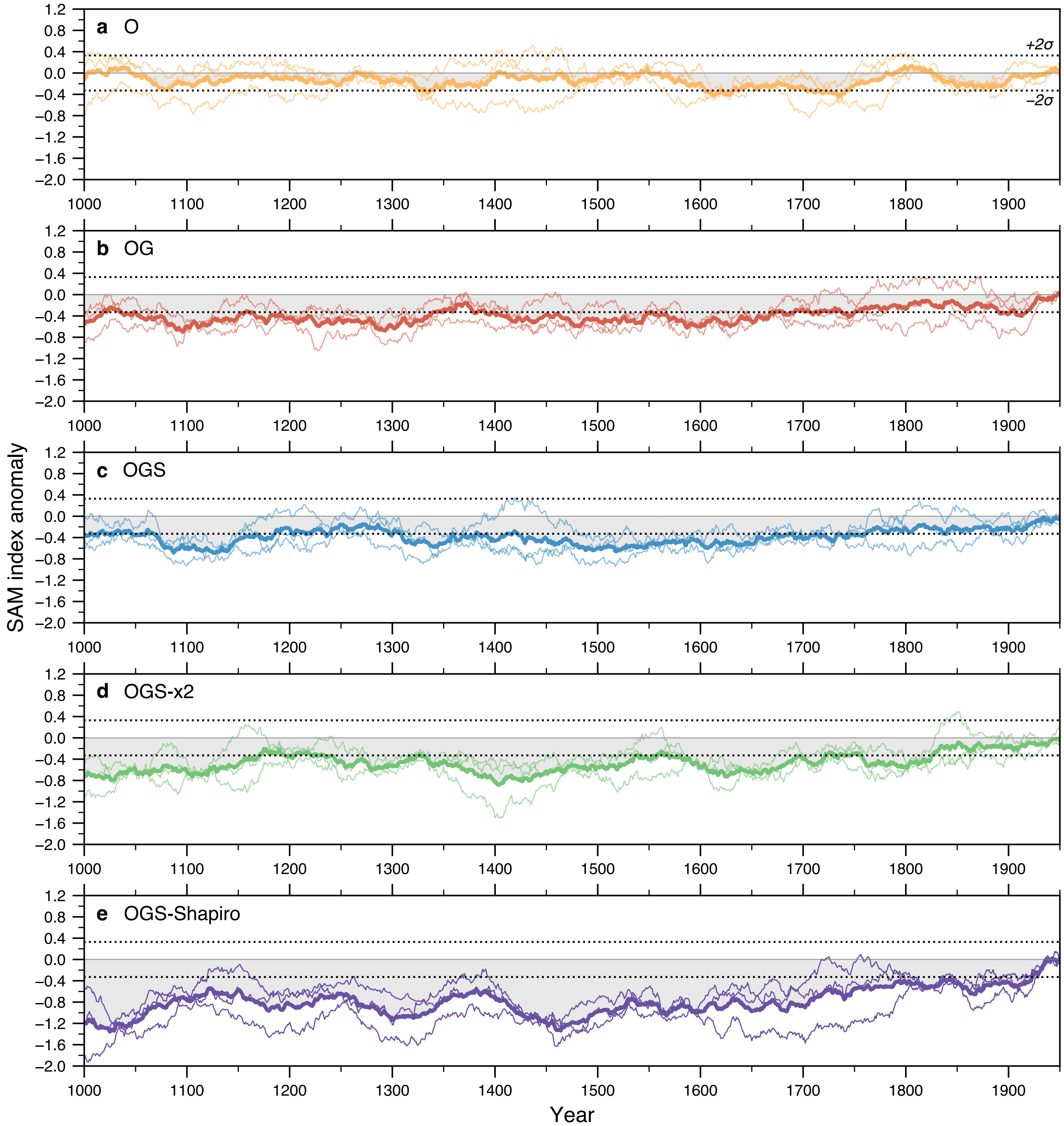

Figure: SAM Index anomaly, calculated relative to the 1900–1999 mean and shown as 70-year moving averages. Thick lines refer to the ensemble mean, while thin lines denote the individual ensemble members. Dashed lines on all the subplots show the ±2σ range based on the orbital ('O') only simulations, representing internal variability.

L34-35 The Neukom et al references are of course fine here, but I wonder if a reference to the PMIP3-PAGES2k paper doi.org/10.5194/cp-11-1673-2015 would also be useful.

Response: We will add the additional reference.

L205. Although mentioned later – I think it would be useful to mention here that HadCM3 also does not have interactive ozone.
Response: We will modify the methods to mention that HadCM3 does not have an interactive ozone.

L206. Which reconstruction is this calculation of the SAM consistent with? The one using monthly means or annual means as a calibration?
Response: Our calculation of the SAM is consistent with the reconstructions that use the annual mean as their calibration (e.g. A14 and Marshall in Fig 3a).

L248 and fig 7 – is the radiative forcing just for solar, or does it include all forcings?
Response: This includes all forcings (orbital + greenhouse gases + solar) for the simulations. We will clarify this in section 3.2, by modifying (green text) the text to:
The simulations also include transient orbital and greenhouse gas forcing, and so we express the results using radiative forcing (*i.e., the combined radiative forcing of orbital, greenhouse gas, and solar;* instead of solar irradiance only) and focus on pre-industrial times (i.e., prior to 1900).

L353 – it would be worth checking the latest PMIP4 simulations runs to confirm if this is definitely still the case.
Response: Unfortunately, there are currently no PMIP4 *past1000* simulations available on ESGF that include interactive atmospheric chemistry.

No mention is made in the results about the 11-year cycle despite the fact you mention that it could have an effect on the SAM in the introduction. I know that the title makes it clear that you are interested in "long-term changes" so the 11-year cycle may be outside the scope of this paper but I wonder whether this is something you have looked into? Is there any evidence of an effect in the transient simulations? Alternatively is there much of an effect in the first decade of the constant forced simulations, how long does it take for the climate to react to a change in forcing? I think this would be a useful addition to this study, but appreciate that there is already quite a lot of work in this paper already so may not be something the authors wish to pursue.
Response: We have not looked into the 11-year cycle in our models, as we are not confident that our models would capture a reasonable response here due to their lack of interactive chemistry, and as we were really interested in the larger, long-term change in the SAM and solar. We agree that this would be interesting to investigate, but unfortunately it is beyond the scope of this study.