The updated manuscript has been returned to me after a second round with reviewers. One reviewer has made a couple of minor comments which I post below in case you don't have access to the report. Please respond to these comments on the open discussion page of the manuscript. I look forward to receiving your revised manuscript.

Regards,

Penny

We thank the anonymous reviewers and the editor for their time and comments, In addition to our response below we note that we have also added the following.

Thermocline depth for as many models as possible (see Table 1). We note that this data has taken much longer to process than planned due to the unexpected outage of our supercomputing facility. This data is now fully processed and is currently being uploaded to the public archive. Should the editor wish to wait to this upload to be complete before accepting this review – we completely understand. We have also added Jemma Jeffree as an author to acknowledge her contribution of processing 3D ocean temperatures to get this depth.

We have added the GISS-E2 model which is described in the text on lines 126-133.

Line 4: Observations sitting within the ensemble spread is not a sufficient test of the models - the credibility of the model's response to forcings and simulation of internal variability is crucial. This needs to be highlighted in the abstract to prevent misinterpretation of the models.

The abstract highlights what we did in the paper as such we have left this as is. We note that we talk about a fair comparison of models and observations on line 4 not an evaluation of the models credibility for it's simulation of individual factors. We have added additional information around this point on line 27 expanding on the information in the abstract which reads 'To this point, fairly evaluating projections in single runs of climate models, particularly for highly variable climate quantities against this single realisation of the real world is only possible by taking long time averages, to effectively smooth out natural climate variability and allowing for the assessment of the model's forced response. The advantage of using a LE is that we can additionally evaluate whether observations sit within the model's ensemble spread. This is a necessary, although not sufficient, condition for model evaluation that makes LEs invaluable tools for such evaluation.'

Lines 36-51: Thanks for adding this text. However, the need to test for potential signal to noise errors should be explicitly highlighted, because if they exist it will fundamentally undermine the straightforward interpretation of large ensembles. Furthermore, whilst I completely agree with the aims of ForceSMIP, if I understand the protocol by using model simulations to evaluate the methods it will not assess signal to noise errors. If so, please make this clear.

Line 52 has been added to read ' Finally, the recent discovery of the signal-to-noise paradox (Scaife and Smith, 2018) where an ensemble can better predict observations than their own members, is also an avenue where LEs can provide valuable insight into model behaviour (Weisheimer et al., 2024).'

Line 179: please add 'modelled' in front of 'response to external forcing'

This has been done.

Line 270: but large ensembles are necessary to test for signal to noise errors. Please amend this

paragraph taking this into account.

Lines 369-370: but large ensembles are needed to test for signal to noise errors - please amend

For the previous two comments we agree – large ensembles are needed to test for signal to noise errors, to quantify variability and potentially many other factors. We do not argue that this is not the case. We argue that for a fair comparison of a observational value to a modelled one if the variability is low, we need less ensemble members. We have changed the wording to comparison in both cases rather than assessment/evaluation to make this point clearer.

Lines 373-376: but detrending does not remove higher frequency signals forced by aerosols, volcanoes, solar, ozone etc. Please be clear about this.

Dentrending by removing a quadratic fit does not remove these signals, removing the ensemble mean does. This is why we highlight that both methods can be used in the package and compared with each other.

Fig 4: I needed 400% magnification to read the text - would it be possible to make it clearer please?

This will be rectified when the Figure is bigger in the final manuscript and the caption can be moved to make more room for the Figure.

Fig 6: the caption still says 'The dark blue curve shows the model’s ensemble mean timeseries' but surely it is the ensemble mean power spectrum?

Thanks for catching this – we have fixed it.