General comments

Quilcaille et al. have revised their paper to enhance the focus on model evaluation, leaving pure presentation of model behaviour for the appendix. I think these changes have improved the manuscript. The addition of Figure 1 was also very helpful for me.

My major concerns focus on a few key areas, many of which echo earlier comments on the manuscript. I think these can be addressed. I also still think the point of the paper could be made clearer (there were fewer tracked changes than I was expecting to see). Is it not: we have used OSCAR v3.1 in a few places already, here we provide a thorough evaluation of its behaviour over a number of experiments where we have something to compare against, here are the levels of agreement (quantified)? If the paper just stuck to telling this narrative, I think it would be much easier to read.

I would also note that many of the other reviewer's comments put a pretty high expectation on the authors. In my opinion, many of the questions asked about particular details and choices related to calibration are better explained by the code accompanying the paper (rather than duplicating this information in the paper) or in standalone papers. Adding such things into a pure evaluation paper (whatever that is worth, see comments below) makes it very hard to have focus.

Overall, I think the paper now achieves its aim of evaluating the behaviour of OSCARv3.1. However, I do think it could be greatly improved in terms of presentation and clarity.

Major concerns

Vague claims of goodness

The vague claims of goodness persist in this version of the manuscript (even in the abstract). Where they appear, they read like the authors want to be able to say, "OSCAR is good", which is particularly odd, because the authors are very honest about the limitations of their model in many other parts. Again, I would just remove any sentence that uses a subjective judgement, such as 'good' or 'satisfactory'. Just tell the reader what the difference is and they can decide what is good enough based on their own situation.

Behavioural description

The authors have retained their section that focuses purely on behaviour of the model, albeit as an appendix. I can see why they want to keep this section, but I have some further thoughts about this.

The first is this. In the revision process, the authors make statements like the following, "However, we highlight that we would not be able to invest the time to transform these deleted results into future studies, hence they would be lost." The implication is that this is the only chance to publish them. My

issue with this statement is that, by saying, "We won't have time", the authors are implicitly saying, "We won't make time". Put another way, the authors are saying that, "These results aren't interesting enough to be worth our time writing up". The problem with this is that it then raises the question, are these results worth anyone's time reading? I think it is ultimately an editorial question whether these pure documentation plots can be included in an appendix or not (they take up space and are disconnected from the main narrative of the manuscript, but you don't have to read them to understand the manuscript so they aren't a negative). However, I still struggle to see why plots of stuff, without any explanation of their implications, belong in the scientific literature (surely they are better captured as part of a tutorial on the model or the model's development repository, where they can be presented without any accompanying narrative?).

My second thought also follows from a comment by the authors, "We highlight that no other reduced complexity Earth system model has done such a thorough analysis before, and such a paper could be a first step towards better descriptions." I would agree with this (more or less) and I think it raises a fascinating question about how to document different model versions. The current practice of writing standalone manuscripts is clunky for a number of reasons. Firstly, a complete description of the model is not appropriate for any single manuscript so it never appears anywhere (rather, any user has to piece together the full picture from multiple papers). Secondly, description papers tend to be very long because they have to cover so much territory. Thirdly, they are very hard to write because they don't have an obvious narrative apart from, "Here is how the model looks/works" (and that narrative isn't very interesting to most people given models are for insight, not for numbers). Given that current practice is clunky, I would encourage the editors of GMD to give this question futher thought: How can model description papers be improved so that they are more useful for authors and readers alike? Are scientific papers even the right forum for such documentation given their focus on narrative and implications? Obviously these questions don't affect the publication of this paper, but given the authors' made the comments I thought I would reply.

Minor concerns

Diagnosis vs. evaluation

The authors refer to the new first section as diagnosis. This language seemed odd to me, I would have used the phrase evaluation because the authors seem to be evaluating the extent to which their model behaves in line with other available literature estimates over a range of experiments. An introductory paragraph at the start of section 3 (before the section 3.1 header) that re-clarifies the point of this section would be helpful (given how long section 2 is).

Reproducibility

The paper's reproducibility would be greatly enhanced if it was clear where an interested party could access the code that sits behind it, particularly the code related to constraining OSCAR. Having the model code available open-source is good, but it isn't enough to actually reproduce the paper's results by itself and the descriptions given in the paper are certainly not enough to reproduce the study by themselves.

Technical corrections

A selection are listed below, but I would note that the paper is still in need of a good proofread as many of the sentences are still missing words and use odd phrases, which makes reading the paper much harder than it needs to be.

page 1, line 12: 'spatial' → 'spatial and temporal' (noting that ESMs often run on sub-daily timesteps)

page 1, line 16-18: 'Overall, OSCAR v3.1 shows good agreement with observations, ESMs and emerging properties. It reproduces the responses of complex ESMs, for all aspects of the Earth system.' Sentence is meaningless without quantification, either delete or add numbers (and remove subjective measures of goodness like 'good')

page 2, line 43: 'is increased' → 'is also increased'

page 2, line 48: 'of CMIP6' → 'CMIP6'

page 2, line 56: 'to evaluate' → 'are used to evaluate'

page 3, line 65: 'meant' → 'is meant'

page 5, line 143: 'As illustrated in Figure 2', I don't see this at all in figure 2...

page 5, line 169: delete 'are used'

page 6, line 190: Suggest adding words like 'following' before the reference to Mcneall. The references don't illustrate the point, but they point in a direction for further improvement.

page 14, line 430: 'on carbon' → 'under experiments that examine'

page 19, line 690: 'resulting quantitative behaviorbehaviour of OSCAR remains largely satisfactory', suggest removing all these vague assertions