General comments

Quilcaille et al. present a summary of a number of CMIP6-style simulations performed with the reduced complexity earth system model OSCARv3.1. They briefly describe the model, how they constrained it to create probabilistic results and then go through the results of their numerous experiments.

The paper is clearly the result of a massive amount of effort, and having this description of OSCAR in the literature is beneficial. My issue with the current presentation is that it lacks any punch. There are so many results, and comparatively so little discussion, that it's hard to know what I'm meant to take away from this apart from, "We ran our model heaps". That's not to say that there aren't really interesting results in the paper, it's just that they're overly hard to find.

I think the paper would benefit greatly from improved focus. Many of the experiments require extensive discussion and further exploration. There doesn't appear to be space for all of them in this paper. For those that cannot be fully explored, I think it is better to save them for future papers where they can be explored appropriately rather than having partial explorations (also because I don't think partial explorations belong in the scientific literature).

Major concerns

Lack of focus

As discussed above, the paper lacks focus. Results are presented from experiment after experiment, with no space to actually explore their implications or what to make of them. Obvious examples are Sections 7.2, 7.3 and 7.5. However, many of the sections left me wondering, "What is the point?"

This is unfortunate, because many of the results are very interesting. For example, it is surprising that the ZEC is much higher for 2000 PgC experiments but overall warming doesn't have the same non-linearity. However, there is no further exploration of this. Similarly, "the carbon stocks still increase in G6solar, even more than in ssp585 thanks to the lower GSAT and despite lower global precipitation." Is this what we would expect? Or does this point to a clear limitation of the model if we have less rain but somehow more carbon stocks? Eight different spin ups are done: how do they compare? What does this tell us about the way we make climate projections and any potential bias in the CMIP-style of doing things. The authors also write, "Our results cannot be compared to the final CDRMIP results yet, for they are unpublished, but they are consistent with those obtained with a model of intermediate complexity (Zickfeld et al., 2021)." However, they have a great tool to evaluate the questions: if they know how much to trust their model (which they should at the end of this evaluation), then they don't need to wait for the CDRMIP results and could write a great (separate) paper on their results now.

I would recommend the authors reconsider which results to present in this paper,

which belong in their own paper and which are best left out. This would probably significantly reduce the length of the paper. It would also improve the abstract, which is currently too long and contains too much detail (it could be re-written to just focus on key points: Emulators are needed, emulators need to be validated, here we examine OSCAR, strengths are X, weaknesses are Y, last sentence could stay as is).

Writing

The writing is very slow, i.e. it doesn't always make clear what the point is. I think this is partly due to a lack of focus as discussed above. It's partly due to being repetitive (the point about the need for validation is mentioned three times in the first paragraph). However, I think it is also partly due to phrasing. It might help to swap phrases like, "As illustrated in Table 4, OSCAR v3.1 estimates a ZEC (in the reference case of the esm-1pct-brch-1000PgC experiment) that is within the range of ZECMIP (Macdougall et al., 2020), although the long-term decrease seems to happen later in OSCAR." with more active formulations (that move all the table and figure references to parentheses) like "OSCAR v3.1 estimates a ZEC (in the reference case of the esm-1pct-brch-1000PgC experiment) that is within the range of ZECMIP (Macdougall et al., 2020), although the long-term decrease seems to happen later in OSCAR (Table 4)." (Yes, I acknowledge the irony of giving advice about punchy writing when my review is probably slightly rambling.)

The paper is also in need of a proofread. Reading it this time was overly difficult due to typing and other phrasing errors. This will take some effort, but it will greatly improve the experience for the reader.

Vague claims of goodness

The paper has some very vague claims of goodness. Two examples, "It reproduces the responses of complex ESMs, for all aspects of the Earth system.", and, "the resulting quantitative behaviour of OSCAR remains largely satisfactory". Both these claims are vague and subjective. I would simply remove them and all others like them from the paper, the reader can judge the quality for themselves based on the results (and likely, the 'good enough' level will change depending on the application of interest).

Specific comments

1. "Similar reasons explain why a pulse of carbon removal cools the atmosphere slightly more over the short term than a pulse of emission warms it, but less over the long term." I can see why you get more cooling in short term (cause of the logarithm). What I can't rationlise is why you get less over the long-term (which feedback is causing the issue, it's the concentration feedback right because the climate feedback makes the sinks even more efficient)?

2. It seems computationally expensive to run all 10 000 combinations for all experiments, before then throwing a bunch away based on a few experiments (mainly ssp585 and 1pctCO2 I'm guessing). Would it not be faster to just run everything for ssp585 first, do some exclusions. Then 1pctCO2 (with whatever configurations remain), do more exclusions. Then run the rest of the experiments? That would save almost a factor of 10 in computing time or do I miss something?

3. page 6, line 169: It is sort of surprising that the fit with cumulative net ocean carbon flux gets worse from unconstrained to constrained. It would be great to have more text on why this could be. Could it be because there is a lack of covariance between likelihoods in the current calculation? It seems like the cumulative CO2 metrics get very high weight (because they're used 4 times effectively as there's 4 scenarios), although that should make the cumulative CO2 agreement better, not worse.

4. page 7, line 214: "OSCAR's overall ability to simulate the RF of short-lived species compares well with the IPCC AR5 values." The methane RF is outside the shown AR5 range, that doesn't seem like comparing well? It's also surprising that the methane RF is lower than AR5 (given Etminan revised it even further upwards). It's also not clear to me how total ERF agrees with AR5 (the lower bias from CH4 and aerosols seems to large to be neatly cancelled by the high bias of tropospheric ozone).

5. page 10, line 318: "the unconstrained carbon cycle of OSCAR v3.1 is well in line with CMIP exercices". Should 'unconstrained' → 'constrained'? Also, while Table 3 sort of supports this, Figure 1 shows pretty clear differences in cumulative compatible CO2 emissions between OSCAR and the constraint it is targeting so I'm not sure I would use 'well in line'.

6. Standard deviation seems inappropriate for skew distributions, can you show plumes of 5-95% and report 5th and 95th percentiles where distributions are skewed (e.g. aerosol ERF) and the appropriate data is available?

7. A curiosity, where is OSCAR's wildfire modelling best described?

Figures

Figure 3 caption: 'middle panels' → 'right panels' and 'right panels' → 'middle panels' ?

Figure 3: "Pearson's moment coefficient of skewness are provided in the legend" I can't see anything, am I missing something?

Figure 6: the arrows don't really help with knowing which direction things are moving. It's possible to intuit, but fixing the arrows would be preferable.

Figure 7: 'breached' → 'branched' ? 'Year after breach' → 'Year after zero emissions' ?

Technical corrections

page 1, line 15: comma missing, should read 'such a model, the newest'

page 1, line 15: 'observations from ESMs'? Can an ESM produce observations?

page 1, line 18: missing 'is' before 'that'

page 1, line 19: 'unstability' → 'instability' (and throughout)

page 2, line 41: 'relative' → 'relatively'

page 2, line 42: 'diagnose' → 'diagnosis'

page 2, line 48: 'increase' → 'increases'

page 2, line 54: 'some of' → 'some'

page 2, line 61: 'the DECK' → 'the CMIP6 DECK'

page 2, line 64: 'to the solar' → 'to solar'

page 2, line 67: 'exercice' → 'exercise'

page 3, line 74: 'and meant' → 'and is meant'

page 3, line 74: '(Gasser et al., 2017)' → 'Gasser et al. (2017)'

page 3, line 74: '(Gasser et al., 2020)' → 'Gasser et al. (2020)' (this referencing error is repeated in multiple places throughout)

page 3, line 74: 'We pinpoint that v3.1 is still calibrated on CMIP5 ESMs, then not meant to emulate CMIP6 models.' What does this mean?

page 3, line 77: 'It means that' → 'As a result,' (one suggestion, other phrasings would also work but the current formulation is clunky)

page 3, line 86: 'and with an' → 'with'

page 4, line 104: 'calibrated on' → 'calibrated to'

page 4, line 120: Should it read 'AR6 volcanic radiative forcing'? If no, have you used AR5 volcanic forcing with CMIP6 experiments, or what is going on?

page 5, line 127: delete 'may'? Or is there something else that can happen that hasn't been described?

page 5, line 128: As above, why do you use may? Are there other options? If yes, they should be described. If no, 'may' can be deleted.

page 5, line 130: 'represent' → 'represents'

page 5, line 149: "We acknowledge that when a significant fraction of the configurations is excluded, confidence in our model's result is lowered, but such a limitation of the validity domain is inherent to reduced-complexity models." Why does this lower confidence? It seems entirely normal for things to explode if you

put in particularly exotic parameter combinations (particularly in a non-linear model like OSCAR). It'd suggest deleting or rephrasing this line.

page 6, line 173: "the model returns after constraining 0.55 K", there is a word missing here somewhere, what does this mean?

page 7, line 192: The text says IPCC AR5 has a standard deviation of 18 PgC, but in the figure it looks more like 35 PgC. Can you please check as this impacts the impression of how close (or not) things are?

page 7, line 193: 'constrain' → 'constraint'

page 7, line 208: Which value do you end up plotting? AR5 rel. to 1850, rel. to 1750, rel. to 1850 using CMIP6 concentrations?

page 8, line 224: Could you clarify (perhaps earlier in the text) which experiment was used for the constraining? I'm guessing concentration-driven historical?

page 8, line 228: How do you come to the conclusion about ocean heat content? You only show one observational data point and it doesn't overlap with the timeseries from OSCAR.

page 8, line 226: "we note differences caused by volcanic eruptions" There don't seem to be any plots or other evidence to help evaluate the size of these differences. Can you please provide some move information on what you're referring to here please?

page 8, line 237: Delete "These results are shown in Figure 3." Sentences such as this aren't needed, just refer to the figure where relevent.

page 8, line 240: 'In regard of' → 'In regard to'

page 8, line 246: 'participates in' → 'contributes to'

page 8, line 252: 'follows the' → 'follows'

page 9, line 258: 'at the end of the 1,000 years' → 'at year 1,000'

page 9, line 259: 'about' → 'related to'

page 9, line 265: "The higher values for the ECS from some CMIP6 models are significantly reduced when constraining (Nijsse et al., 2020; Bonnet et al., 2021), with ECS even lower – 1.38K with a likely range of 1.3-2.1K – than those shown by OSCAR here." There are some words missing here (and maybe one of the ECS is meant to be TCR) so I couldn't follow this sentence.

page 9, line 268: "provided by OSCAR remain consistent with the litterature" Can you please include the OSCAR values and literature values so it's easy to judge whether these are indeed consistent? Or more clearly refer to Table 2 in the text so the reader knows where to look.

page 9, line 269: 'unconstrained' → 'constrained'? Again, please provide values so the reader can judge the level of consistency.

page 9, line 286: ‘are strongly’ → ‘are more strongly’

page 10, line 293: ‘on these’ → ‘in these’

page 10, line 320: “In any case, this suggests that our carbon cycle may be too optimistic”, meaning that the uptake is too high for the same level of warming?

page 11, line 340: ‘of the hydrological’ → ‘on the hydrological’

page 11, line 343: ‘fully coupled ESM’ → ‘emissions-driven reduced complexity model’, OSCAR isn’t an ESM

page 11, line 346: ‘to its’ → ‘at its’

page 11, line 350: ‘at the same rath than the ramp-up period’ → ‘that is the reverse of the ramp-up period’

page 11, line 351: ‘over’ → ‘for a further’

page 12, line 357: ‘return within’ → ‘return below’ ?

page 12, line 367: ‘pinpoint’ → ‘note’

page 12, line 376: ‘aims at’ → ‘aims to’

page 12, line 384: ‘First, the ZEC in 385 branched experiments is systematically lower than the one in bell experiments.’. I see the opposite in Figure 8. The branched experiments are all higher in the corresponding year e.g. the branched 2000 PgC peaks around 0.2C, whereas the bell 2000 PgC never gets that high.

page 13, line 399: ‘group of forcing’ → ‘group of forcings’

page 13, line 412: ‘For the simulations under natural forcings, the range from the constrained OSCAR is smaller than the ones from (Gillett et al., 2021), which may suggest an over-constraining.’ Is the range from Gillet quoted or am I missing something? It could also be that Gillet is under-constrained?

page 13, line 416: ‘cannot conclude as to’ → ‘cannot comment on’

page 14, line 447: ‘preindustrial’ → ‘quasi-preindustrial’

page 15, line 463: “Both shifting cultivation and wood harvest have no impact at all on the land sink, by construction of their formulation in OSCAR (Gasser et al., 2020).” Is this what we would expect based on first principles? Or is it a quirk of OSCAR?

page 15, line 466: should ‘increases’ be ‘decreases’? Please clarify the experimental design and why it is one way or another (my intuition was that having grasslands instead of croplands would increase the sink as written but I can’t see how to infer this from Figure 9, where negative numbers imply a reduction in the sink).

page 15, line 472: ‘an land’ → ‘a land’

page 15, line 481: ‘increase of -17 ± 13 PgC in the land’ → ‘decrease of 17 +/- 13 PgC in the land’ would be clearer

page 17, line 524: ‘warming’ → ‘causing less effective radiative forcing’ because you’re not talking about warming

page 18, line 552: ‘ability of properly isolating’ → ‘ability to properly isolate’

page 19, line 582: “four tested scenarios, respectively”, which order (SSP585-RCP585 first or SSP126-RCP26 first)?

page 19, line 587: ‘Follows’ → ‘What follows is’

page 20, line 612: delete ‘would’

page 38, line 1069: ‘cases’ → ‘case’