Review of CMIP7 documentation paper, by Dunne et al.

Firstly to say that the CMIP panel and authors here are to be congratulated on the way they have approached the task of developing CMIP7 plans in a complex landscape of requirements. CMIP has had a lot of success historically but requirements have grown and that growth is not sustainable so the new approach to consult with both users and providers and hence prioritise a more manageable, but still vital, set of simulations has been extremely welcome. The outreach, consultation and dissemination of information has been excellent throughout and this paper contributes to that process.

CMIP is a huge undertaking and changes the deployment of resource (both personal and computing/technology) in many, many modelling and research centres around the world. Careful design of what is requested and why is essential.

I perform this review mainly in the context that the main aspects of CMIP7 and the Fasttrack, are already determined and too late to make substantial changes. Therefore, I focus on the presentation and explanation aspects with a few suggestions of things which could still be tweaked or clarified.

My major comment is to ask for more details on where the "Guiding Research Questions" came from? Are these the result of a consultation on the priority climate science questions? They resemble, but are not the same as, past WCRP grand challenges (e.g. on extremes or carbon cycle). The way the paper is presented implies you started with these as a guiding set of questions and designed CMIP7 to answer them. But in practice that wasn't how I recall it happening – so have these questions been retro-fitted to the experiments? E.g. line 132 says that CMIP7 design came from consultation and surveys – this is certainly true of the experiments – but did this consultation also take place for the science questions?

When I look over the CMIP7 web page there are lots of details and further links to the experiments, the task teams, the data request, the REF etc. Your figure 3 is replicated on the website, which mentions the science questions linked to each FT experiment - but I cannot see the questions described or explained anywhere. It feels like these questions have been added after the experiment design. If these really are "guiding questions" that have guided, and are intended to keep guiding, CMIP I think they and their derivation need more prominence.

It is not clear, for example, why you identify SST patterns over, say, cloud feedbacks, as a key driver of system sensitivity? Also, when you discuss a "carbon-water nexus" – is this just a catch-all for things not included in the other questions? The paragraphs of description of this question (sec 2.3) don't appear to cover interactions between carbon and water cycles as implied by the "nexus" tag.

So overall it would be good to articulate maybe how these priorities were arrived at. I am not querying the importance of these questions – they are clearly crucial. But other aspects (for example on aerosol forcing and cloud processes) could also be seen as equally important, and CMIP7 will address many more than just these. Maybe it is better to present the experiments first and then give some example high priority questions as examples of things which CMIP7 may help address – but it feels to be overselling the tag of "guiding questions" to imply that these came first and led to the CMIP7 design.

Other suggestions I think are important:

Model/simulation quality

i. Lines 374-375 – it feels reasonable to suggest a degree of stability of a control run: +-5ppm is probably OK – but better as a rate than an absolute – is this +-5ppm *per century* for example? In CMIP6 C4MIP requested drifts of less than 10 PgC per century in the main pools. But it would be consistent to also request stability criteria for other metrics – e.g. global T must drift by no more than +- XX degrees, or AMOC within XX Sv. It would be good to treat all major climate components similarly.

ii. More importantly – I think it is unwise, however, to suggest arbitrary quality criteria for historical runs. Many ESMs may not hit the historical CO2 within 5ppm. See e.g. Hajima et al (https://egusphere.copernicus.org/preprints/2024/egusphere-2024-188/) for thorough evaluation of CMIP6 models in this respect. What happens if a model does not hit you 5ppm bounds – is it excluded from analysis? Again – as above, will you also specify acceptance criteria on other measures? – e.g. goodness of fit of the historical temperature record? This would be a big change for CMIP – to specify acceptance criteria – I think it needs much more consultation before you introduce this.

Ensembles – do you have any recommendations around generation of ensembles (from each model)? I realise you don't want to rule out models by requiring large ensembles, but some experiments may benefit more than others from ensembles. Line 510 says that the FT "promotes the generation of ensembles" – but it is not clear how? FT does not appear to mention ensembles at all – but it could be a good opportunity to do so.

It might be useful to provide guidance on this without mandating. Likewise you could guide on choice of initial conditions (e.g. branch points best taken >XX years apart from the control run).

As an example, quantifying TCRE from flat10 is a relatively large signal-to-noise activity. Ensembles may add little value to this. But quantifying ZEC from the flat10-zec simulation is a very small signal-to-noise and ensembles of this run could be really useful. See e.g. Borowiak et al (https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2024GL108654) which shows that ZEC derived from CMIP6 ZECMIP are subject to a level of uncertainty which CMIP6 did not consider due to lack of ensembles.

Spin-up. I'm not sure I understand the request to submit numerical results from the spin-up of the models. What is the goal of this – how will they be used? "for curation" sounds like an odd phrase – why do these need curating? And what does "curation" involve – is this the same as archiving on a public database like ESGF?

Model selection. I think you are very wise not to do any prior screening or selection of models. The "hot models" paper you cite in Appendix 3 by Hausfather et al is rather simplistic to provide a table of "Y" and "N" on model screening based on sensitivity. A more nuanced analysis by Swaminathan et

al (https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2024EF004901) shows clearly that many metrics of crucial interest are not related to ECS. Many high sensitivity models have very good evaluation scores on many metrics and vice versa – having a lower ECS is certainly not a measure of quality. Any screening or selection needs to be much better understood and carried out case-by-case for the application in question. It cannot (yet) be done at the scale of CMIP which has so many down-stream uses of the outputs.

Minor comments

- Lines 102-107. This is a nice description of how CMIP has expanded and refined focus as both the expertise and need evolves. It feels that more knowledge of reversibility and symmetry is a big gap in our understanding of the climate system, and here could be a good place to articulate the need for more process exploration of how the system behaves under reversing of forcing.
- Line 216 says that CMIP7 focus on emissions-driven runs allows for more exploration of extremes under stabilisation – can you explain how so?
- Sec 2.4 on points of no return – is there a reason not to call this either "tipping points" or "irreversibility" which have become much more common phrases for these topics. Wood et al (2023 - https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2022EF003369) is a good reference here for the framing of high impact/low likelihood outcomes and the need for research spanning different dimensions of this topic.
- Line 297 onwards – describing the CMIP7 DECK intent. It is worth being explicit here that the goal is only to characterise the response to _increasing_ forcing. It was a deliberate decision not to add a DECK experiment to characterise the system response to reducing forcing. (This remains a gap in CMIP7 – noting that flat10-cdr can only be performed by ESMs)
- Table 1 is important. A couple of notes/suggestions
    - For esm-piControl the forcing is described as "emissions" - I wonder if this should be better described as "interactive CO2" or "simulated CO2" because of course there are no emissions. So even though we informally describe this as "emissions mode" it risks implying that there are some emissions being applied. Or at least specify that CO2 emissions are zero.
    - Typo – looks like the 1% and historical lines have transposed the solar/volcanic forcing entries
- Line 355. Can you clarify the need for 100 years of control run before any experiments are branched off? I don't recall this being requested in CMIP6
- Line 364 – can you explain why conc-driven control run is required if the esm-control is stable? That seems redundant
- Table 2 is useful – but it feels odd to name individuals. What happens as/when a person moves job etc? maybe a named group in an organisation is more useful.
- Table 2, N deposition. Will this be speciated into dry/wet and oxidised/reduced reactive nitrogen?
- Line 405. The section on spin-up – it is not clear how the strap line "characterising model diversity" is relevant to this sub-section. Maybe just call the section "ocean and land spin-up" (where land here includes land ice/cryosphere?)
- Line 470 – is "SCP" a typo? "SSP"?

- Table 3 is super useful and important – it will be a very good easy-look-up of the whole set of FT simulations. But it is really big! It is important that it is produced and typeset to be easily readable given how big it is. I feel this comment may be more for the journal/typesetters than the authors – I hope you can find a way to make it well readable.
- Table 3 – scenario time period. You quote that scenarios run to 2100 – is this decided? I thought it would be 2125, or at least this was still being discussed. (personal opinion – it drives me mad that IPCC figures and values can only ever quote a climate – i.e. 20-year average – for 2090. So an extension to a minimum of 2110 seems vital so that we can actually quote a 2100 value for projected results!)
- Appendix 1 – requested spin-up metrics. As per my comment above I'm not yet convinced why you need to request these. But if you do, then to close the land carbon cycle you should also requested cProduct. Even if the control run has no land-use _change_ it will still have land use, and the product pools may well be non-zero. cLand is then the sum of cVeg+cLitter+cSoil+cProduct