I commend the authors on a lucid and concise document. I also appreciate how difficult this can be for such a complex undertaking. Hence I offer my comments and invite them to consider them or not, as they wish, and as the editor sees necessary. I don't wish to impose requirements or additional burden on what is already a difficult enterprise.

The document nicely describes efforts to document how the latest efforts to fit global climate models to past observations influences differences in their behavior. In this regard, and given that there has been no real breakthrough in the way models are fit to data, isn't it important to acknowledge the need to document how they behave as compared to the models in past intercomparison projects, and which experiments contribute to that. If a model differed from what was inferred from the ensemble of models in CMIP5 would we be able to document why and how in the same way that we are trying to document how models in CMIP7 differ?

I very much liked the emphasis on the science questions. While these helped understand how experiments were chosen to document the model behavior related to each question, I was wondering whether they are physically motivated. If so, highlighting the physical hypotheses that informed the experimental design would strengthen the scientific contribution.

Related to this, the manuscript acknowledges little how far physical understanding of state dependence has progressed in recent years. For example how the IWV influences the width of the window and hence the degree of water vapor masking (Bourdin et al., (2022)), how cloud masking of the surface window response depends on cloud temperature and coverage (Stevens and Kluft (2023)), and how the distribution of clouds in temperature space can be equifinal in CRE but lead to different radiative responses to surface warming (Kluft et al.,(2025)). These area all examples from my near field, but these ideas have also been highlighted in the work by Jeevanjee and collaborators.

Table 1 is a nice overview, but it seems like some experiments don't relate to any question or application. Probably this is not true, but the manuscript gives this impression. Maybe it can be corrected by adding another category of purpose.

There have been some very large and ambitious cloud and cloud circulation field studies over the past years, (AC)3, CAESAR, EUREC4A, ORCESTRA, OTREC. The manuscript gives the impression that the modeling community sees no value in these efforts. Certainly their relevance is not reflected in the paper, where I believe that they could guide the experimental design. Consider comparing ORCESTRA with GATE in AMIP, for instance.

Aerosol was not mentioned, do they enter the protocol in any way? Given how strongly many models represent such effects, is there a strategy for documenting their impact on cloud responses. Or is there a reason (may be covered by other communities) that can be shared as to why not.

I didn't understand (§3.2.2) the use of HADSSTs. It made it sound like the AMIP SSTs are not being regularly updated. Given how emphasis the CMIP community has placed on updating the forcing datasets this seems hard to believe. Asking modeling centers to do extra work to assimilate other SST datasets just because the AMIP SSTs are not being updated misplaces the application of community persuasion.. I do understand the value of looking at the sensitivity to small changes in SST patterns, but then a different experimental design would be more appropriate (climatological SST with HAD1SST and AMIP II SSTs)

I appreciate that CMIP tends to focus on one type of model, and to the extent CFMIP is about designing experiments and output to help document the behavior of this type of model the experimental design works well. If however the purpose is to understand how clouds contribute to climate change (not just in CMIP models), a greater process focus would be beneficial. It would open the door to model diversity, from LES, to DNS, to kinematic cloud microphysical models, to storm-resolving models, to regional models, to observations.

It is not clear to me where I would find the COSP diagnostics. Is this still the Swales et al paper? If so its use of ISCCP simulator output makes no physical sense. Clouds don't depend on height, and not really on pressure, but they do depend on temperature. If one wishes to distinguish

between cloud temperature and surface temperature then this should be included in a dimension of the simulator, rather than pretending it is somehow captured by height. Not the least because passive sensors on satellites don't see clouds in this way, and the conversion to temperature space is straightforward for active sensors.

Is there a strategy for dealing with the artificial treatment of clear-sky as 'cloud-cleared-sky' which is not observable. The latter arises because models don't treat the clear sky different in scenes where cloud fraction is less than one, which of course is wrong. One way to do this could be to output the radiation as histograms by cloud fraction, as that might get closer to what we observe, and it seems like that would be more desirable than contorting the observations to be wrong in the same way as the models.

Maybe Li et al (2013) for the importance of accounting for slow ocean responses (line 168).

The use of the word "opportunity" rather than CFMIP data output "protoco"l or "request" reads like a politically motivated change in language that a small group of people communicating in narrow circles has invented to help display their virtue. For most others I suspect it will just be confusing.

Bjorn Stevens