Response to Reviewer 1

Reviewer comments are in black, authors' responses are in blue

The GeoMIP project has produce a lot of useful results so it is good to take a step back and think about lessons learned.

The review of past experiments and pointers to some of the resulting papers is very useful and interesting.

We thank Ken Caldeira for his comments. We have tried to include all of his suggestions in the revised manuscript.

Major notes:

This paper would be substantially improved with the addition of a section titled something like "Lessons learned". If you were to start this project over again knowing what you know now, what would you have done differently? [Since there might be multiple perspectives on what should have been done differently, this might be an opportunity to share several perspectives. In this reviewers perspective, this lessons learned section would be the most important section of this paper.

Similarly, it might be good to pull together what I would see as a section on cross cutting issues. The two issues that I see as being raised at several points are:

1. **To what extent should simulations attempt to be "realistic" and to what extent should the simulations be highly stylized aimed at facilitating more straightforward analysis?**
2. **Where should the balance be between simulations that may in some sense be "better", but be difficult for modeling groups to perform, versus simulations that might not be as useful, but might be easier for modeling centers to perform?**

There are probably similar cross-cutting questions that you might want to address, for example: When is it enough to have a small number of groups do a simulation and when do you really need a large number of groups to do a simulation? How do you draw a balance between then number of simulations that people need to perform versus the number of ensemble members for each simulation? How to think about the GEOMIP demands on people's time versus everything else they need to be doing?

I would not expect to see resolution on all of these questions, but maybe a couple of sentences showing the thinking on all of these questions might be useful.

Some of this material is already in Section 5.2 and the Conclusions section. Nevertheless, I think adding a "Lessons learned" section would be highly useful, and a "Cross-cutting issues" might be a place to focus discussion on some of the questions raised in Section 5.2 and the Conclusions.

The reviewers can accept or reject my "cross-cutting issues" suggestion but I hope would adopt my "Lessons Learned" proposal would be adopted.

We thank the reviewer for his suggestions. We feel like many of the topics he is discussing were already there, so in trying to avoid lengthening the piece too much, we have restructured the manuscript moving many of the opinions around past experiments from the end of section 2 and other sections to a new section with the title suggested. We have also tried to highlight more in the Conclusions the "cross-cutting issues" he mentioned.

Minor notes:

[line] comment

[24] Eliminate word "these" (stylistic) Done

[43-45] Provide citation for IS92a claim. We have provided a citation to Pedersen et al. (2020): their figure 1 clearly shows how close IS92a has been to mean growth rates during the 1990-2019 period.

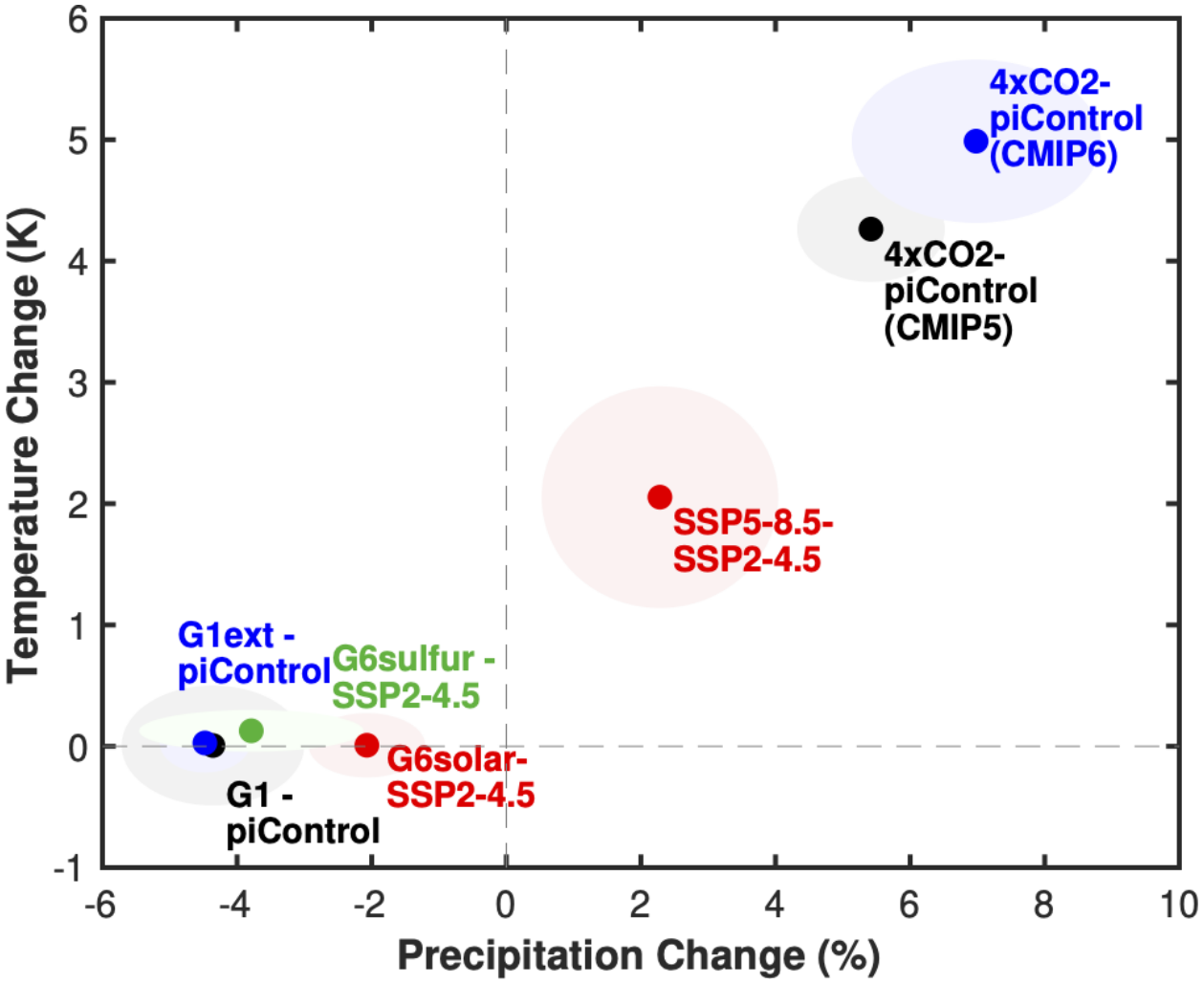
[67] Properly capitalize of project name. We followed the capitalization present in all of their reports (see i.e. https://cordis.europa.eu/project/id/226567/reporting) which never capitalized other words except the first. We now added ' ' to highlight the name, however.

Table 1. Maybe make another column for the background scenario. Thank you for the suggestion, we have added that and cleaned up the table.

Figure 1. It might be helpful in this figure or in an additional figure to make it clear which CMIP scenario is the reference "ungeoengineered" case relating to each geoengineering case. At the very least this could be in the figure caption. We have tried different options but that clutters the figure too much. Given that the reference scenarios are now more explicit in Table 1, we have added a reference to the table.

Figure 2: Expand figure caption to explain all labeled points in the figure. For what years are this? Is it really the standard deviation so low, or are these perhaps standard errors? If the values for G6Solar, G6sulfur, are compared against SSP2-4.5 values, might it be a good idea to show the SSP2-4.5 value on the figure? Do something to let people know which geoengineering case is related to which case without geoengineering. Thank you for your suggestions, we have tried to make the figure

clearer. We cannot include SSP2-4.5 values because they are the reference values, but we have tried to specify more things in the revised figure and caption.



New caption: *A comparison of global temperature (K) and precipitation (%) changes for some Tier 1 GeoMIP experiments across CMIP5 and CMIP6. Points represent the multi-model averages for each experiment, shaded areas represent 2 multi-model standard errors. Values for G1 and 4xCO2 (CMIP5, 13 models averaged) and G1ext and 4xCO2 (7 models) are from Kravitz et al. (2021), comparing against piControl values in the last 40 years of the experiment (years 11-50). Values for G6solar, G6sulfur and SSP5-8.5 (6 models) are from Visioni et al. (2021b), comparing against SSP2-4.5 values in the last 20 years of the experiment (2080-2099)*

[104-105] Please mention whether the reduction was the same in each model or different to achieve a temperature balance. We added some clarifications, and this phrase: "*For instance, such a comparison showed that between the models that performed*

this experiment across two generations, the value of solar reduction needed ranged from 3.80 to 5.00 % (Kravitz et al., 2021)."

[124] Please mention whether the reduction was the same in each model or different to achieve a temperature balance. Mentioned, as above.

[221-226] Some discussion of the use of SSP2-4.5 as a reference state rather than SSP5-8.5 would be appreciated. Was the choice to be more "realistic" worth having a higher signal-to-noise ratio? From the discussion on these lines, it seems researchers wanted to be more "realistic", but maybe it is better to hit models with a hammer to see how they behave with more extreme forcing. We have highlighted this trade-off in an additional phrase that reflects this point: "*Perhaps excessive focus on "realism" -whatever the current opinion on that is at a given moment- is good for communicating results and convincing modeling teams to consider performing a set of simulations, but might result in scenarios that perhaps do not hold the test of time as well as simpler, higher signal-to-noise experiments like G1.*"

[228-229] This discussion of "future proofing" might be expanded and discussed later along with the above questions. Is the goal to be "realistic" or to understand how models behave? How are these competing goals best balanced? Noted, refer to largest change.

[Section 3 and 4] For each of these subsections, it might be good to start each section with the main scientific question that each project is intended to address. (For example, line 509 mentions a question in a section that has no questions in it.) Good point. To find a better balance between length and descriptiveness, we have decided to expand the title of each subsection to make sure it reflects which question it is supposed to answer.

[619] Something akin to this boldfaced question should appear in each of the subsections of Sections 3 and 4. (Maybe not a bad idea to do this for Section 2 also.) See above.

[777-788] These kinds of questions about tradeoffs in design should get more prominence. We wholeheartedly agree!