We are grateful to the reviewer for their valuable suggestions that have enhanced this work. Below, we indicate the reviewer's comments in blue and our response in black.

There are a few inconsistencies and a lack of clarity related to the transition temperature, which I summarize here:

1. How is the transition temperature defined? I assume when the feedback becomes zero?

We added clarifications on that aspect in the Methods section. Yes, the transition temperature is when the feedback becomes zero, but more precisely it is also when the TOA imbalance is the least negative, as any other year would be in a more negative imbalance than this one.

2. If my interpretation is correct, how is the value and the uncertainty determined? E.g. looking at Fig. 3A in the 1/64 x CO2 line, I could see why one would argue the transition temperature to be at -20 K, but at the same time -33 K also seems reasonable. Similar reasonable ranges seem to exist for many other simulations.

See comment above.

3. How does this uncertainty affect the uncertainty of the emergent constraint?

We added the effect of the uncertainty on Fig.6. Even with a large uncertainty of 5 K on the transition temperature, this affects the upper bound of ECS estimate uncertainty by a bit less than 2 K. We emphasise that we are constraining the upper bound of ECS here. Much larger values on the upper bound are often given, and with some lines of evidence the upper bound is basically infinite. Therefore, even providing a value of 10 K with uncertainty brings valuable information to the community effort of constraining ECS.

My second set of comments refers to the emergent constraint:

1. Fig. 6: Taking out all the CESM-2 points (the blue ones and, as I read from the text, the upper left white one) leaves no relationship at all. In fact, the authors write that "the relationship lacks robustness" (caption Fig. 6). Considering only the remaining PMIP models, there seems to be no linear relationship at all, which seems to indicate that all the emergent constraint actually comes from CESM2.

Only the left-most point is CESM2, the others belong to the CESM family, as written in the methods section (CESM1.2, CESM1.3). After reviewing the text, we removed "the relationship lacks robustness", as this is not entirely true. In fact, the relationship is robust with higher ECS models than those of the PMIP4 ensemble. This is shown in Renoult et al. (2023): in fact, CESM2 is not necessary for the robustness of the relationship, as CESM1.2 and CESM1.3 suffice (Fig.14 of Renoult et al., 2023).

2. The instability threshold was previously argued to be around 0 degC, in this figure it starts around -8 K SST anomaly relative to preindustrial. Given that pre-industrial SST were around 15 degC or so, how does this go together?

We agree that this figure is indeed confusing. We used SSTs here because we did not have land surface temperature values for the CESM model family (blue dots), so we instead converted the results we got from Fig.3 into SSTs (which would provide a transition temperature at around -8 K SST relative to pre-industrial). We decided to entirely remove the CESM model family, as described above and in Renoult et al. (2023) they are in fact not necessary for the robustness of the relationship: they and CESM2 are almost interchangeable. Instead, we made a new figure which uses this time only TS to be consistent with Fig.3 and Fig.6 and adjusted our conclusions accordingly.

3. How is the emergent constraint different from simply excluding models that freeze over in the LGM when estimating ECS on the grounds that they would be too sensitive? Would that method lead to the same results?

This is indeed a very good point, however this would only work efficiently if there were enough models with high ECS that would publish their runs. We have very limited information regarding higher ECS models that went to snowball Earth states when trying to simulate the LGM, as explained in the introduction. This is why we try to motivate modelling centres to share and publish those simulations. Using a method similar to an emergent constraint allows to use all the published LGM runs, not just those that fail. Regardless, if we look at the ECS of the models we suspect might be entering snowball Earth instability under LGM conditions, then our constraint appears somewhat conservative. As this is anecdotal evidence, it is difficult to implement in the method..

The third set of comments refers to the threshold of snowball transition at 0 degC:

1. Some arguments are very hand-wavy. For example, the argument that this supposedly generalizes across models (l. 118-119) and the “geometric argument” in l.122-124.

The fact that the state-dependency could be general across models is hypothetical ,but not hand-wavy: sea-ice is expected to freeze at the same temperatures in any model and consistently with our physical understanding, so we believe state-dependency, which is the controlling factor of a true inception towards a snowball state, is similar across models. The geometric argument is admittedly slightly hand-wavy, but not necessarily wrong either: it is a simplification and illustration of the point above.

2. In particular, I am not sure that the transition temperature would be 0 degC across all models. Already in Fig. 3A) I can see a transition temperature range of ~ 15 K across simulations performed with the same model. Similarly, in Fig. 5, the transition temperature seems to be at or below -20 K anomaly to pre-industrial, which would be well below 0 degC, too. The statement that all

models share a similar transition temperature to snowball Earth also seems to be at odds with the statement in l. 28-29, which points towards different models having different transition temperatures. Also the statement in l. 147-148 points toward different transition types and temperatures even within the CESM model family.

The transition temperatures vary a lot in Fig.3 due to time-dependency effects, as written here: “Nevertheless, the transition temperatures of each phase show a slight shift to lower values under stronger negative forcing. Therefore, the climate system deviates from pure state-dependent behaviour as the strength of the radiative cooling and the speed of transition to snowball Earth increases.”. However, we also emphasise that slower simulations are the closest to any real transitions, as illustrated in a schematic we added in the Methods section, and because they are the least affected by time-dependent effects. When running a 50 ppm run: “When abruptly decreasing the CO2 concentration to 50 ppm (around 1/4 of pre-industrial CO2), we find hints of instability near the global mean temperature of 0°C”. Unfortunately those simulations can run for thousands of years, which is why many studies have used “fast” transitions with strong abrupt forcing, like we also show in our paper. In this case it is then interesting to compare those simulations and discuss time-dependency as we did in our study. L28-29: “These climate models start transiting to a snowball state at temperatures substantially cooler than indicated by LGM reconstructions” does not indicate that models have different transition temperatures: 0°C, which is the transition temperature as calculated in our study, is well below any LGM reconstructions. Some models might have different transition types, for instance they might show a stable waterbelt state. However this does not mean they necessarily would have a different transition temperature. This only indicates that they have a third possible state which exists between today and a complete snowball Earth state.

l. 28 – 29: I found this surprising, and from skimming through Zhu et al. 2021a, I didn’t find that information. From my understanding, they look at only one climate model family (CESM), albeit in different configurations. While CESM2 definitely goes to a snowball state, I can’t see at what temperature the transition would happen, as their Gregory plot (Fig. 2 (d)) seems to indicate a stable regime throughout (with small, but nevertheless negative feedback). Furthermore, I was under the impression that cloud feedback accelerates the transition to the snowball state, or, as stated in l.95-97, doesn’t change the transition temperature. All of these statements seem to be inconsistent with each other. A similar statement is found in l.107. How do these conflicting statements go together?

CESM2 enters runaway below its simulated LGM temperature (so somewhere below -11.3 K). Our sentence is indeed confusing and we modified it. Cloud feedbacks accelerates the transition, but this is not inconsistent with them affecting the transition temperature. Cloud feedbacks can affect the cooling rate of the simulation (how fast the simulation will reach the transition temperature), but the transition will still happen at nearly the same temperature because it is mainly controlled by the strengthening of the sea-ice albedo feedback, as shown in Fig.2 and in the text.

l. 32: independent from what? If the intended meaning is “independent from models”, then I think independence is a strong claim, which should be further justified

Independent from other estimates, clarified.

l. 99-100: This is almost exactly the same finding as in Abbot 2014 (https://doi.org/10.1175/JCLI-D-13-00738.1), please cite

Done.

l. 145: “universal”: I suggest rewording, given that it was tested on only two selected models

“Universal” is used here in a hypothetical way.

Why is there no summary, conclusions, or discussions section? I don’t want to insist on the traditional structure, but some wrap-up and putting-into-context at the end of the paper might be helpful.

We discuss our results within the text and Section 5. We believe the structure of our paper clear is enough to answer our scientific question.

The authors did a great job motivating the study in the introduction. A few additional sentences about ECS and its uncertainty would be helpful, since the emergent constraint on ECS is one of the main purposes of the study. Also, the emergent constraint could already be clearly set as a goal for the paper, as well as the logic that is behind it.

Added.

l.99 I suggest “locked clouds” rather than “locked cloud feedbacks”. I find “locked cloud feedbacks” not wrong but misleading, because with locked clouds there is actually zero cloud feedback.

We do not think the term is misleading, as the method applied here is called “feedback locking”.

l. 130: “state” here refers to temperature, not CO2 concentration, I guess? If so, then I suggest making this clearer, since the CO2 concentration technically belongs to the state.

To avoid confusions we changed “state-dependency” to “temperature-dependency”.

l.156-158: I am not sure the sentence is correct grammatically, at least it’s hard to digest

Fixed

l. 171 and following: I suggest to implement the call for the new experiments to the abstract.

Added