We would like to thank the reviewers for their detailed and comprehensive comments and suggestions. We greatly appreciate the time and effort you have devoted to reviewing our manuscript. Below we respond to each comment point by point. The reviewers' original comments are in **black**, and our responses are in **blue**.

For RC2:

General comment

The revised manuscript shows clear improvements in structure and clarity, and ERA5-Land is now appropriately positioned as a complementary dataset. The introduction of model-family weighting also helps to reduce the over-representation of certain models. However, several figures (e.g., Figs. 5 and 6) remain visually dense, and the discussion retains a somewhat descriptive tone. Most importantly, the attribution of biases to specific physical processes in the models remains incomplete, with important explanations from the literature still missing. I therefore consider that the paper requires minor revisions before publication.

1. Inclusion of CNRM-CM6.1 and CNRM-ESM2.1

The authors have introduced model-family weighting to reduce the over-representation of the two CNRM versions. This change is a step in the right direction. Thank you.

2. ERA5-Land

ERA5-Land is no longer presented as a primary evaluation benchmark but as a complementary dataset, with a clear explanation of the regridding and its limitations. This fully addresses the initial concern, and no further adjustment is needed on this point.

3. Figures 5 and 6 too dense

Figure 6 has been partially simplified: tas/tsl variables are kept in the main figure, while pr/snd have been moved to the appendix, and graphic elements have been enlarged for better readability. However, the main figure still appears dense, with many overlapping symbols and layers of information, which limits clarity. Further simplification could improve understanding. The same applies to the new Figure 5.

For Figure 5 we now only show DJF results in the results section, which allows us to enlarge the figure, and put the original figure to the appendix, which includes all seasons.
For Figure 6, we restructured the labels in the bottom and enlarged the main figure. The label used for CMIP6 models are altered now to open circles instead of bullets,

thus the overlapping are reduced.

4. Discussion and conclusions

The revised manuscript now includes a dedicated Discussion section, organised into subsections that address the objectives set out in the introduction, as well as a more structured conclusion. This improves the link between objectives, results, and interpretations. However, it would be beneficial to include more perspectives on the implications for permafrost or climate modelling and to reduce the repetition of results.

We rewrote most of the Discussion section to reduce repetitions and added more supporting points.

5. Understanding of processes and literature review

A factual error regarding the formulation of snow thermal conductivity in JULES has been corrected, which is a positive step. Nevertheless, the main explanation for the cold bias in the CNRM models—linked to the representation of the snow-free fraction beneath tall vegetation—remains absent or superficially addressed, despite references available in the literature (Decharme 2019, Wang 2016). More broadly, the discussion still lacks an in-depth review of the processes specific to each model, which limits the ability to correctly attribute observed biases to their physical causes. In other words, this part remains superficial and would benefit from further development.

Thank you for specifically pointing to the 2 papers, which we revisited. The issue is now explained in more detail in the fifth paragraph of section 4.3.
We added further discussion to MIROC6 regarding their snow parameterizations. However, it is challenging to comprehensively relate all results to specific model features. It would need further sensitivity studies for each of the models. This statement was also added to the conclusions to highlight the need for further studies.

For RC3:

Major revisions:

- While the introduction has been significantly improved, it is still unclear how readers who are not familiar with the LS3MIP experiment will understand the objectives of the study.

For example, the statement:

“the differences between the two can be used to attribute model biases to either the land surface model structure and parameterizations or coupled atmosphere-land interaction”

requires further clarification. How could a larger bias in the land-only model (compared to the coupled models) be attributed to parameterization issues or missing processes in the LSMs? If the same LSM is used, there is no obvious reason why the bias should be reduced in the CMIP6 experiment. This is explained in the LS3MIP experiment but not here.

Similarly, the statement:

“Under identical, observation-based atmospheric conditions, the LS3MIP models are expected to simulate soil temperature more accurately than their CMIP6 counterparts.”

requires additional justification. From a logical standpoint, this is not self-evident, even if modellers may agree that this is often the case in practice. A clearer explanation is needed to substantiate this claim.

We added information about LS3MIP to help readers understand what LS3MIP is, why and how we used it in this research. You can find the relevant changes in the last paragraph of the introduction.

- The discussion section in the revised manuscript remains difficult to follow and is not well structured.

Several paragraphs contain sentences addressing multiple, unrelated aspects, which reduces readability. Furthermore, the section does not consistently compare the results with findings from previous studies, which would be essential for a proper scientific analysis.

The section also suffers from inconsistencies in tense (switching between past and present), and it uses qualitative terms such as “better” or “good,” which are not

scientifically precise.

While I have highlighted some of the most concerning sentences below, a thorough reorganization and rewriting of the discussion are necessary.

We went through and revised the discussion section and tried to make sure every statement is straightforward. We removed some repetitions and added more supporting points for our statements. We also addressed the tense and wording issues.

Minor revisions:

- Line 53 – Add a reference.

We added Koven et al. (2013) and Yokohata et al. (2020).

- Line 111 – The term _significant_ should not be used unless based on statistical analysis. Please be more precise.

Changed into „*Northern Eurasia contains more than two-thirds of the Earth's permafrost area (Groisman et al. 2007), with the majority located in Siberia.*"

- Lines 113–117 – These sentences could be shortened.

Revised as *„Within this region, we used the soil temperature observational dataset at standardized depths, as provided by the All-Russian Scientific Research Institute of Hydrometeorological Information-World Data Center (RIHMI-WDC) "*

- Figure 5 – Align the orientation of the figure and its caption.

We simplified Figure 5 and put the previous to appendix (Figure A4), where we aligned the orientations accordingly.

- Figure 7 – In the authors' response, it is stated: _"The data have been grouped into intervals of 2°C, as illustrated in the histogram pairs. However, the x-axis ticks have been set at 5°C intervals in order to accommodate the wide range of temperatures and to avoid the cluttering."_ However, it remains unclear how a 4–6°C bin would be treated—would it belong to the 0–5°C or the 5–10°C interval? This should be clarified, as the current explanation is not consistent.

We thank the reviewer for pointing out this potential ambiguity. To clarify, the data is strictly grouped into 2°C intervals . The x-axis ticks are displayed every 5°C only for visual clarity and do not represent the binning. For example, a bin of 4–6°C is a distinct 2°C interval on its own, The center value of each bin can be seen by the center

x value of E5LC subplot and center x value of group pairs (between purple and green bars) in other subplots. We revised the figure caption to explicitly state this, so that the distinction between bin width (2°C) and axis tick marks (5°C) is unambiguous.

- Table 3 – Most readers may not be familiar with the meaning of _Kurtosis_. It would be useful to provide a short explanation in the discussion, including what it represents and why it is relevant in this context.
We added introduction of Kurtosis in Method section 2.5, and in discussion we further discussed about it.

- Section 4.1 – This section is not necessary or should not be part of the discussion.

Moved into method section 2.6.

- Line 531 – Rewrite this sentence for clarity.

Revised in the first of Section 4.1 and added more information.

- Line 545 – Avoid the use of _better_, as it is not scientifically precise.

We revised the betters with clearer wording where necessary.

- Line 566 – The sentence _“Moreover, better snow simulation ability improved soil temperature simulation performance”_ needs clarification. Does this refer to results observed in this study? If so, where in the results is this demonstrated? If it is instead linked to the previous sentence regarding summer data, the connection is not evident and should not be made without further justification.

This general statement is actually not true for all models in our results, so we removed it.

- Line 568 – Provide a stronger motivation for this statement. How exactly does the data indicate deficiencies in representing surface energy exchange?

Rewrote as in the sixth paragraph of Section 4.2.

- Line 576 – The reference to Dutch et al. is problematic, since that study does not address soil insulation but only snow, and it focuses on a single site in Alaska outside the study domain.

After revising we no longer use this reference at this part of the article, but later and related only to snow.
Although its outside our domain, we still can use it in the context of improved parametrization for snow thermal conductivity, as the characteristic of frozen soil

snow should be similar.

- Lines 577–582 – This section begins with observational results (which should not appear here), then transitions into a speculative explanation involving thermal resistance or low soil moisture, which is insufficiently explained. This part requires substantial revision for clarity and coherence.

We have rewrote the discussion part. The relevant content could be found on the last paragraphs of Section4.2 and Section 4.3. And we removed redundancy of results in discussion.

- Line 597 – Since Dutch et al. is already cited for snow insulation earlier, consider consolidating the discussion here to avoid redundancy.

We rewrote the contents according to this reference and now it is not repeating.

- Line 602 – While snow does indeed have low thermal conductivity, the statement should be made more specific. The value varies considerably by region and conditions.

We now include a range of values, from fresh to dense snow.

- Line 609 – This sentence is unclear. If it is merely a description of the figure, it should not be part of the discussion.

We rewrote it, including more detailed discussion (Section 4.3, third and forth paragraph).

- Lines 615–619 – Consider integrating relevant discussions from Dutch et al. and Damseaux et al. here.

Thank you for highlighting the relevance of these references for this part of our discussion. We included them, noting their primary finding that using the snow thermal conductivity parameterization described by Sturm et al. (1997) enhanced Arctic soil temperatures in CLM5.0.

- Line 649 – Clarify what is meant by _confirmed by the results_.

We removed this wording.

- Line 691 – This sentence requires clarification.

Rewrote as in Section 4.3, paragraph 5.

- Line 717 – This finding is also demonstrated in Damseaux et al. and could potentially be added here.

Yes, this supports our conclusion, we referred accordingly.

Comment on the authors' response:

- The classification of "Permafrost stations" remains problematic. The revised definition using the two-year criterion is an improvement, but the map still shows inconsistencies—for example, some northern locations are excluded while some more in the south are included. This likely results from the criterion requiring "at least 300 days of valid data every year." To avoid confusion, it would be clearer to rename these sites as "valid stations", or something else, rather than "permafrost stations," since the current label may lead readers to believe that excluded stations are not underlain by permafrost.

We now use the name ‚valid permafrost stations'.

- In the response, the authors stated: _"In the revised manuscript, we will provide a link to Zenodo where we will upload the relevant scripts."_ However, no such link is currently included. Please add this reference.

We have sorted the scripts and uploaded them. The Zenodo link is now provided in the manuscript.