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**.

Review of "Assessing Climate Modeling Uncertainties in the Siberian Frozen Soil Regions by Contrasting CMIP6 and LS3MIP"

Firstly, I would like to commend the first reviewer for their thorough and detailed evaluation of this manuscript. The major strengths and particularly the weaknesses of the study have been clearly identified, leaving little room for additional comments. I fully agree with the assessment that this work would greatly benefit from major revisions to enhance its methodological rigour, clarity, and overall structure. Below, I provide a few additional comments that address secondary but potentially relevant concerns.

While this study provides valuable insights into the performance of CMIP6 and LS3MIP models in frozen soil regions, it would benefit from revisions to improve its methodological rigour, clarity, and overall structure. Removing redundant models, excluding ERA5-Land, providing a clearer discussion of results, and improving the presentation of figures would significantly enhance the manuscript's quality. I encourage the authors to consider these points in their revisions.

Thank you for your thoughtful review and constructive feedback. We appreciate your recognition of the insights provided by our study and the points raised for improvement. In response, we will carefully revise the manuscript to address the concerns listed below. Specifically, we will re-evaluate the inclusion of ERA5-Land, refine the discussion to ensure clearer communication of the results, and improve the presentation of the figures for better visual impact. We value your suggestions and will incorporate them to strengthen the overall quality of our work.

1. Inclusion of Two Nearly Identical Models (CNRM-CM6-1 and CNRM-ESM2-1)

One methodological issue that warrants attention is the inclusion of both CNRM-CM6-1 and CNRM-ESM2-1 in the analysis. These two models share an extremely similar structure and code base, making them effectively redundant. Their simultaneous inclusion biases the statistical assessment of model diversity and artificially reinforces certain trends. The authors should consider removing one of these models to ensure a more balanced and independent analysis.

We acknowledge the structural and codebase similarities between these models. However, we believe that analysis of both models also provides an opportunity to examine where their results diverge, which may provide additional insight into the impact of specific model components or configurations. Similarly, we note that the HadGEM and UKESM models are also closely related, with UKESM essentially being the ESM version of

HadGEM. To address your concern, we will explicitly highlight the similarities between these models and emphasize their differences in the relevant sections of the manuscript. In addition, we now adopt a weighting method proposed by Kuma et al. (2023). Consequently, Equation 3 in the manuscript is updated to read as follows:

$$EB_{i,s} = \sum_{m=1}^{M} \frac{1}{F\,N_m} \left(med_{m,i,s} - med_{o,i,s}\right)$$

where *m* represents each climate model, *F=5* is the total number of 'model families' and N_m is the number of 'family members' within each family. In practice, in our case, this results in CESM2, IPSL-CM6A and MIROC6 being given weights of 0.2, while CNRM-CM, CNRM-ESM, HadGEM3 and UKESM are given weights of 0.1. This adjustment also affects Figures 5 and 9 in the manuscript.

2. Questionable Use of ERA5-Land Data

ERA5-Land is a reanalysis product, not an independent observational dataset. While reanalysis can sometimes provide useful large-scale validation, its role in this study appears unjustified. The manuscript already includes direct observations, which are far more suitable for model evaluation. Furthermore, comparing models against another model-based dataset (ERA5-Land) does not provide meaningful validation or evaluation. Removing ERA5-Land from the analysis would streamline the results and improve the manuscript's focus on actual observations.

We will no longer position ERA5-Land as a "benchmark" or reference dataset for model evaluation. Instead, ERA5-Land is presented as a widely used dataset that serves as a complementary basis for comparison. Since ERA5-Land was generated similarly to the LS3MIP simulations, a land-only model driven by reanalysis, ERA5-Land provides an opportunity to explore how reanalysis-based datasets perform alongside direct observational datasets. In addition, the ERA5-Land data have been regridded to match the resolution of the LS3MIP simulations (~0.9°), allowing us to assess the influence of resolution on the results.

3. Overly Complex and Unreadable Figure 6

Figure 6 is too dense and difficult to interpret, as it combines multiple variables (tas, tsl, pr, snd) in a single diagram. This makes it hard for the reader to extract meaningful insights. A better approach would be to separate this into multiple figures, each focusing on a single variable. For example, sub-figures or distinct panels could be used for each variable, with clear titles and well-defined legends. Additionally, clearer labelling and improved visual representation would greatly enhance readability.

A more effective approach would be perhaps to create a separate figure for each season, with four distinct panels for tas, tsl, pr, and snd. This structure would allow for a clear comparison of model performance across different seasons and variables, enhancing readability and facilitating the identification of trends and anomalies. Using box plots or violin plots in each panel would effectively display the distribution of data, making the figures less cluttered and more insightful.

We will seperate the figure into two, one with tas and tsl, another with pr and snd (which will be put into appendix). Additionally, we will explore incorporating distribution plots to enhance readability in the figures.

4. Lack of Discussion Section and Unstructured Conclusions

As previously noted by the first reviewer, the manuscript lacks a dedicated discussion section, and its conclusions do not sufficiently synthesise the results in relation to the stated objectives. In particular, the authors should:

- Revisit the key research questions outlined in the introduction and explicitly address them in the conclusions.
- Provide a clear synthesis of the main findings, rather than scattering them throughout the results.
- Offer a more structured discussion, especially accounting for the following comments 5.

To address these issues, we will introduce a dedicated discussion section to thoroughly analyze the findings, situate them within the existing literature, and address their broader implications. In addition, we will revisit the key research questions outlined in the introduction and explicitly answer them in the conclusion section to ensure alignment with the study's objectives. The conclusions will also be restructured to clearly consolidate and synthesize the key findings and avoid the current scattered presentation.

5. Lack of In-Depth Understanding of Model Processes and Literature Review

One of the most concerning aspects of this manuscript is the apparent lack of in-depth understanding of the processes simulated by the six analysed models. Throughout the text, the authors make causal claims about model behaviour that are either too vague or lacking sufficient references, sometimes even incorrect, suggesting that they have not thoroughly studied the literature on these models. A deeper engagement with existing research would improve the accuracy of the study and prevent misleading conclusions.

Before attempting to diagnose model biases and uncertainties, the authors should conduct a more comprehensive literature review on each of the models they analyse. This would allow them to:

- Properly attribute biases to the correct physical processes,
- Avoid making incorrect causal inferences,
- Provide a more nuanced discussion of model differences.

A clear example of this issue is the discussion of CNRM-CM6-1 and CNRM-ESM2-1 in lines

350-353. The authors claim that the cold bias in these models is due to snow conductivity, when in reality, it is mainly caused by the way these models simulate snow cover fraction as a function of vegetation (see section Snowpack Processes and Appendix B in Decharme et al. 2019). Unlike observations, which assume a fully snow-covered ground, these models allow for a snow-free fraction where soil is directly exposed to atmospheric forcing, leading to an artificially cold soil temperature. This is well-documented in the literature (e.g., Wang et al. 2016). For example, Decharme et al. (2019) states: "In addition, the specific snow fraction over tall vegetation is generally very low, annihilating the soil insulation effect of the snowpack." This explanation is completely absent from the manuscript, despite being a critical factor in the model's behaviour. Wang et al. (2016) also provide a robust and clear discussion of this problem in their "Model Processes" section.

In summary, the authors would benefit from reviewing Wang et al. (2016), which provides an excellent discussion of snow/temperature processes in models (see the "Model Processes" section, page 1733). Writing a discussion of equivalent quality about model processes is essential if this paper is ever to be accepted.

We recognize the need to strengthen our understanding and discussion of the processes simulated by the six models analyzed, and will address these concerns in our revision. We will conduct a more comprehensive literature review, including the work of Decharme et al. (2019) and Wang et al. (2016), as suggested. The discussion will be revised accordingly, and vague claims will be removed.