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**.

The manuscript addresses the persistent discrepancies in frozen soil simulations within climate models and their land components. By comparing historical runs from seven land-only models participating in the LS3MIP with their coupled counterparts in CMIP6, the study aims to disentangle the contributions of land surface parameterisations and atmospheric forcing to these discrepancies. Given the importance of accurate frozen soil simulations for climate projections and land-atmosphere interactions, this study aims to explore valuable insights into the strengths and limitations of current land surface models and their coupling with atmospheric components.

However, while the study's ambitions hold significant value for the scientific community, the execution falls short in several key areas. The manuscript lacks the clarity and rigor expected in an academic publication, with issues in writing style, structure, and methodological justification. Furthermore, some interpretations of model results are overly speculative, requiring a more cautious and evidence-based approach. Addressing these shortcomings through major revisions will be essential to ensure the study’s findings are both robust and impactful.

We appreciate the reviewer's detailed feedback and constructive suggestions to improve the clarity and impact of the manuscript. We acknowledge the areas that require significant revision, particularly with regard to writing style and structure. We will carefully implement the recommended changes to ensure that the manuscript meets the standards of scientific publication.

Major revisions

- The writing style does not meet the standards of a scientific publication. Many sentences are overly long and vague, and overall consistency is lacking. Several sections—particularly parts of the introduction—are less rigorous and require significant clarification and development. To ensure the paper is accessible and acceptable to its readership, many statements need to be refined and expanded.
- The introduction requires a complete overhaul. I recommend restructuring it into distinct segments that provide:
- **Global/local context:** Outline the broader and specific contexts relevant to the paper.
- **Identification of issues:** Clearly identify and discuss the current challenges in frozen soil simulations.
- **Research question:** Present a specific, well-defined question that the study will address.
- **Study approach:** Detail how the paper intends to answer this question through its methods and analyses.

Thank you for your specific suggestions for restructuring the introduction. We will take them into account when rewriting the introduction.

- The manuscript would benefit from a clear separation between the results and discussion sections. This division would enable readers to first digest the novel findings in the results section and then understand their interpretation and comparison with previous studies in the discussion. Currently, comparisons with existing literature are scant; the manuscript should incorporate and reference a broader range of studies relevant to the subject matter.

  We will provide separate results and discussion sections to ensure that readers can focus first on the results and then on their interpretation and contextualization. We will also place the results in the context of other relevant studies.

- Regarding the use of ERA5-Land data, the paper's central question is not directly related to the quality of this dataset. The statement "this proves that ERA5-Land can be a solid benchmark that supports observation as gridded data" is somewhat misleading. Providing "tas" values that closely match observations does not inherently qualify ERA5-Land as an appropriate benchmark for studies focusing on soil temperatures. Instead of justifying its use, the authors should expand the "Data and Methods" section and include a more thorough comparison with LS3MIP simulations to support their choice.

  We have revised our approach to the use of ERA5-Land data. We will no longer present ERA5-Land as a "benchmark" or reference dataset, but rather as a widely used dataset that provides an interesting basis for comparison. Similar to the LS3MIP simulations, ERA5-Land is a land-only model driven by reanalysis. In addition, the ERA5-Land data have been regridded to match the resolution of the LS3MIP simulations (~0.9°), allowing us to evaluate how the resolution affects the results. We will also expand the Data and Methods section accordingly.

- When results observations, such as biases or qualitative values (e.g., high, low, warm, cold) are mentioned, include specific numerical values. This practice will enhance clarity and allow for a more precise interpretation of the data. This

  We'll include specific numerical values in the results section to improve clarity and precision.

- Most importantly, the authors' attempt to link the biases observed in this manuscript to the physical processes represented or the parameterizations used in the models is largely misleading and lacks sufficient nuance. To make such claims, the authors could rely on sensitivity experiments, as they themselves acknowledge: "It is challenging to identify how model features influence the vertical energy transportation process without conducting sensitivity experiments." Without such analyses, the attribution of biases to specific model processes remains speculative. The manuscript would benefit from a more cautious interpretation of results, clearly distinguishing between observed discrepancies and their potential causes. Additionally, a more thorough review of existing literature on model parameterizations and process representations would provide a stronger foundation for discussing the origins of biases. Instead of drawing causal

conclusions prematurely, the authors should consider discussing alternative explanations, acknowledging uncertainties, and explicitly stating the limitations of their approach. For now, the results provided and their analysis are simply not sufficient to support the interpretations and conclusions.

In the revision we will take a more cautious approach, explicitly distinguishing observed biases from speculative causes. We will enhance the discussion by reviewing relevant literature on model parametrizations and process representations to provide better context. We will also clarify uncertainties and limitations in our methodology to ensure that interpretations remain balanced and well-founded.

Figures and tables revisions

- Verify and, if necessary, adjust the color palette to ensure it is color-blind friendly for all figures (I doubt figure 2/3 are for example).

  We tested the color blind readability of images on the inspection website suggested by The Cryosphere. The current palette is friendly to all color-weak users. However, since we used more than three models, we were unable to meet the reading requirements for red-green blindness.

Table 1:

- Specify which version of CLM5 is being used (e.g., CLM5.0).

  Yes, CLM5.0. Also adjusted everywhere else.

Table 2:

- Replace “snow conductivity” with “snow thermal conductivity.” Note that all values are density-dependent except for MATSIRO. It would be more informative to share the default scheme used (e.g. Yen 1981)and reference the relevant publication rather than only providing the mathematical formulation.

  Initially we adopted the “wording” from Menard et al. (2021) and Wang et al. (2016), now we included the references of the used scheme as suggested.

Figure 1:

- Clearly indicate that “perma” refers to frozen soil at -20 cm, as opposed to permafrost at deeper layers elsewhere.

  We used a much strict way to select permafrost sites. As mentioned in the reply to RC1, the definition of permafrost in Lawrence and Slater, which requires at least two consecutive years of soil layers in a frozen state, poses two problems when implemented in this study. 1. Many sites lack deep soil layers below 0.8 m, and the soil above 0.8 m at these sites almost always thaws annually. 2. The study spans 30 years, during which some sites had frozen layers for two consecutive years, but the majority of the time was characterized by freeze-thaw cycles. We have therefore adopted a very strict permafrost screening scheme: a site is defined as permafrost only if the observation data at a depth of 3.2 meters (with sufficient observations) remain completely frozen throughout the 30-year period. Of course, it is not

rigorous to define only these sites as permafrost, so we will come back to the original definition when revising.

- Specify what is displayed by the color bar adjacent to the figure.
- Consider adopting a visualization approach similar to that in Figure 5 for representing color classes—despite the current use of a linear color scale, note that employing too many classes (a maximum of 7×2 is recommended) can be problematic. If a diverging color scale is preferred, it should be centered around 0.

Figure 2:

- Ensure consistency across figures with respect to units, labels, and color schemes (e.g., using either "Celsius" or "°C" uniformly). The ERA5 Land color scheme should be consistent (currently it varies from gray to black), and model labels should be uniform. Choose ERA5 Land instead of ERA5_Land.
- The model labels and colors are not clearly aligned. It would be preferable for the legend to explicitly represent the color assignments and to differentiate offline land with a distinct style, as demonstrated in Figure 6.

  The labels are altered. But we don't consider now making offline boxes another style, it will add complexity to the readability.

- Clarify (in the figure caption and potentially also elsewhere in the text) that the site-average corresponds to the average of all selected station locations.

Figure 3:

- Ensure that the colors match exactly those used in Figure 2.

  Double checked, they are the same.

- The colors assigned to HadGEM and UKESM are too similar, making them difficult to distinguish.

  This is on purpose as they are using the land model from same family, just similar to CNRM models. We now use different markers for them, so they can be distinguishable.

Figure 4:

- There are no shaded areas.
- Include ERA5 Land in the legend.

Figure 5:

- Align the orientation of both the figure and its caption.
- The current figure is challenging to interpret due to an overload of information. It would be beneficial to label each column or row directly within the figure, rather than forcing the reader to refer continuously to the caption.
- Include a legend that explains what the circle radii represent. Reducing the maximum radii is advisable to avoid overlapping circles, as both color and circle size are important for interpretation.
- Maintain a consistent cartographic projection for all maps; the projection used in

Figure 5 currently differs from that in Figure 1.

Valuable point, we will make them the same.

- Some data points are not legible. Although this may have been intended to portray small biases and standard deviations, it could be misinterpreted as insufficient data. Using an edge color (e.g., black) for the points may enhance clarity.

The comments above will be considered and fixed into Figure 5.

Figure 6:

- The volume of information in this single figure could be overwhelming for readers. It is recommended to separate the diagrams by variable—focusing on the most relevant ones in the main text, while relegating the less critical diagrams to supplementary material.
- For enhanced clarity, consider dividing the legend into distinct sections: one indicating color (for different models), another for symbols (to distinguish between ESM and stand-alone LSM), and a third for numerical indicators (indicating variable significance).
- The “REF” is not adequately described; it should be clearly explained in the figure caption or within the main text.

We will separate the four variables into two groups and present them across two distinct figures, while also modifying the legends for greater clarity.

With regard to the term "REF", it is important to note that this refers to the point at which the 1 time normalised standard deviation, and the correlation coefficient to observation is 1.This description will be incorporated within the figure caption and discussed in the main text.

Figures 7 and 8:

- Although the manuscript states that “the model simulation outputs are binned at 2°C intervals,” the x-axis labels and ticks reflect a 5°C interval.

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 that would arise from using 2°C ticks.

Figure 9:

- The significance of the violin plots is not explained. The authors should clarify what the areas, dots, and black lines represent.
- It remains unclear what the x-axis represents. Specify which “temperature” is being shown (e.g., air temperature or soil temperature, observations versus model results, and which dimensions are averaged).
- The current style of the figure does not adequately support the subsequent discussion. For example, statements such as “Set Frozen and Set Warm tas data sample sizes are more than twice as large as in Set Intermediate” and “the difference in LS3MIP runs is negligible, suggesting that the climate model will likely

have a cold and small deviation under these temperature states" are not easily decipherable. Focusing more on the standard deviation as a visual might improve interpretability.

We will enhance the figure by adding clear explanations of the violin plots, including what mentioned here.

We will try to improve the style of the figure, and add more info via extra table or data in plot, to ensure that our statements can be easily understood.

For your reference, the X-axis the categorized by observation temperatures(tas for the left subplot and tsl for the right subplot). We calculated 30-year mean at every month and station, so the data sample here has the size of 12month*236stations.

Figure 10:

- It is surprising that this figure includes data from all seasons. Isn't it only winter data? If the intent is to follow the approach of Wang et al. (2016), the data should be restricted to winter conditions.

  Initially, the selection of DJF data was undertaken in accordance with the approach of Wang et al. (2016). However, due to an inadequate amount of samples for this period, the analysis was expanded to include data from other months as well. The present study focuses specifically on near-surface temperature data points below -5°C, utilising monthly average values. Consequently, it is hypothesised that the prevailing selection criteria can provide a satisfactory and reliable sample size for analysis.

- Position the model names at the top of each panel to enhance readability.

Minor revisions

Thank you for your valuable and specific comments, which we greatly appreciate. We will take them into account in our revisions to improve the manuscript. In particular, the introduction will be rewritten and the results and discussion sections will be separated. Below, we address some of your comments individually where further clarification or alternative perspectives are needed; for other comments, we agree with your suggestions and will revise accordingly.

General

- Some terms need to be more explicitly defined:
- **Models:** Clearly define which type of models is under discussion (e.g., Earth System Models, land surface models, etc.) in the introduction.

  We will make brief introduction of Climate Model, Earth System Model and Land Surface Models in the introduction.

- **Model Ensembles:** When introducing model ensembles, particularly in reference to CMIP6 and LS3MIP, specify this term explicitly.

We will revise the text to explicitly define and specify the term "model ensembles" in the context of CMIP6 and LS3MIP for clarity and precision.

- **Land-Only vs. Offline land surface models:** Adopt and consistently use one term throughout the manuscript.

  We use term "land-only".

- **Permafrost:** At line 152, the manuscript classifies certain locations as permafrost, which may be misleading because it may imply that permafrost is present exclusively at these sites. The explanation provided in lines 223–227, clarifying that these locations exhibit permafrost soils at a depth of –20 cm, should be introduced earlier and applied consistently to avoid ambiguity

  The classification problem is answered above in Figure 1 comment. Here, this was an error in expression. We will clarify a solution and maintain consistency throughout the text.

- Use negative numbers consistently when referring to depth and cold bias values. For example, if cold biases are sometimes shown as negative values, ensure that all such instances follow that convention.
- The authors are encouraged to provide the code used to produce the figures in addition to the underlying data. This will improve transparency, reproducibility, and the ability for readers to further explore and validate the results.

  In the revised manuscript, we will provide a link to Zenodo where we will upload the relevant scripts.

Introduction

- **Lines 20, 44, 46, 78, 94–97:** The authors should include additional references at these lines.
- **Lines 21, 171:** Update the references with more recent publications to reflect the current state of research.
- **Line 24:** The sentence is overly long and should be rewritten into two or more concise sentences to improve clarity and readability.
- **Line 28:** If abrupt thaw is mentioned, a clarification is needed to explain its relevance to this study. The authors should clearly delineate the connection between abrupt thaw processes and the objectives of the manuscript.

  Our research does not specifically address abrupt thaw. It is mentioned here to illustrate the importance of permafrost research in the context of climate change. We will consider reducing this content or expressing it more logically and coherently.

- **Line 29:** Clarify the phrase “such carbon emissions” by specifying what these emissions refer to, ensuring that the meaning is unambiguous.
- **Lines 30–32:** The implications of “changes in surface vegetation types” should be further developed. While the focus on permafrost thaw is noted, the authors need to expand the discussion to include other consequences of permafrost thaw on Earth’s ecosystems, not just those related to vegetation.

- **Lines 34–35:** Revise the sentence for clarity. The term “frequently varying” is ambiguous, and the concept of “thermal offset” should be explicitly defined and contextualized.
- **Lines 35–37:** The statement regarding differences in time scales between soil and atmospheric processes and the role of the soil surface as the interaction window needs further refinement. A clearer explanation of these dynamics, with concrete examples if possible, will help strengthen the argument.
- **Lines 36–38:** Should examples be mentioned here, the authors are advised to include at least a couple of specific cases. For instance, emphasizing “excess ice” conditions (as noted by Burke et al. (2020) and other studies) would help illustrate the point effectively.
- **Line 41:** The statement is currently too vague. The authors should detail what each mentioned characteristic does and how it impacts the study.
- **Line 42:** Specify which conditions are being referred to (including the time-scale, any specific event, and the relevant soil depth), providing the reader with necessary context.
- **Line 44:** The phrase “the most suitable” is subjective and should be replaced with more objective language.
- **Line 45:** The text mentions horizontal resolutions; it should be clarified why high resolution is necessary to distinguish between different frozen soil regions. A brief explanation or supporting evidence is recommended.
- **Line 47:** Provide further justification for the necessity of high-resolution data at this point in the paper.
- **Line 56:** Describe the potential consequences (or cite relevant studies) that are being discussed. It is advisable to move this sentence to an earlier position in the introduction, before the datasets are introduced, to frame the context properly.
- **Line 59:** This sentence is overly long.
- **Lines 61–62:** The delineation of which “characteristics” are being assessed is too vague. In addition, the concept of a “benchmark” is not developed. The authors should specify which benchmark is being applied.
- **Lines 62–64:** Although the general concept is sound, this section should be expanded. The authors need to elaborate on the differences between CMIP6 and LS3MIP that could lead to the observed biases and uncertainties in frozen soil regions. In particular, clarify how these differences might be attributed to discrepancies arising from the land surface models versus those caused by atmospheric forcings.
- **Line 64:** The phrase “with identical and more realistic atmospheric conditions” is ambiguous. Clarification is needed regarding what is meant by this and why, under such conditions, LS3MIP models are anticipated to simulate soil conditions more accurately.
- **Lines 65–66:** This sentence, which is key to the study, requires further development. The rationale behind regarding certain discrepancies as “errors in the land surface models” needs to be clearly explained, with supporting arguments that make the rationale accessible to all readers.
- **Line 67:** Clearly specify which features are being referred to. Providing examples where applicable will help avoid ambiguity.

Data and methods

- **Line 76:** Consider introducing the concept of climate “feedback” in the introduction, as it is a key component of this study. This will help set the stage for its later use in the analysis.
- **Lines 78–80:** The sentence in these lines is unclear. A revision is needed to improve clarity and ensure that the intended meaning is conveyed unambiguously.
- **Line 81:** Replace “climate models/earth system models” with “Earth System Models (ESMs)” to maintain consistency and accuracy in terminology.

  We removed the term “earth system models” and kept the more general term “climate models”. ESM would not apply to CNRM-CM6.1, MIROC6 , and HadGEM3 (e.g. see definition by Eyring et al., 2016 or the nice overview graphic from Kuma et al., 2023). So we keep the more general term “climate models” throughout the manuscript.

- **Lines 82–83:** Rephrase “cannot be considered” to clarify whether the models lack a freeze option when turned off or if they do not adequately represent frozen soil processes.
- **Line 84:** Remove any repetitive wording to improve the flow of the section.
- **Line 91:** Use the term “snow thermal conductivity”.
- **Lines 91–92:** The current description is somewhat misleading. It should be clarified that all formulations are either (1) empirically derived and density-dependent or (2) assigned fixed values. This nuance is important for understanding the parameterizations.
- **Line 100:** Replace the vague phrase “assist our assessment” with a more precise description of how the method contributes to the analysis. Additionally, the term “numerical” is too vague; the authors need to clarify and explain the differences between ERA5-Land and other land surface models, including a brief definition of what reanalyses entail.
- **Line 103:** Provide details on how the available depth data are interpolated to the target depth of the study. This clarification is necessary for understanding the data processing methodology.
- **Line 107:** Specify which quality flag is being referenced and explain what it represents regarding data quality or processing.
- **Line 108:** Reconsider the rationale for using user-defined values of longitude and latitude to determine warmer climates. Note that areas east of 120°E do not correspond to Siberia. It may be preferable to use average air surface temperature measurements to define these regions.
- **Line 128:** The phrase “in the central tendency of the data” is ambiguous.
- **Line 139:** Clearly define the seasons used in the analysis. For example, if DJF is employed, specify whether it covers December 1st to February 28/29th or follows another seasonal definition.

Results

- **Line 142:** Instead of beginning every sentence with “Fig. x…”, the authors should directly present the scientific point. This repetition occurs multiple times and could be streamlined to improve readability.
- **Line 143:** The term “matching” is superfluous and should be removed.
- **Lines 147–148:** Clarify the rationale behind the observations made here. Consider moving this explanation to an earlier portion of the section so that the context is established before the results are discussed.
- **Line 150:** The term “outcomes” is vague.
- **Line 162:** The term “interpolated” is ambiguous. The authors should specify the interpolation method used and indicate how this process might affect the results.
- **Lines 162–163:** The sentence stating, “Fig. 2 shows slight differences between different land models because of interpolation uncertainties using different model grids with different setups,” uses the terms “differences” and “different” in a repetitive way. The authors should (a) avoid rushing to conclusions by providing supporting quantitative evidence (e.g., specific values or statistical measures), and (b) rephrase the sentence to clearly explain the potential impact of grid differences and interpolation uncertainties.
- **Lines 163–164:** The statement, “This illustrates how carefully a comparison of coarse-grid model output against point-like station data has to be interpreted,” requires further explanation. The authors should provide concrete evidence or reasoning to demonstrate how this conclusion was reached.
- **Lines 170–171:** Rephrase and clarify the material within the brackets to ensure that it is concise and informative.
- **Line 171:** The phrase “same family” should be explicitly defined. The authors need to clarify what criteria determine the grouping of models into the “same family.”
- **Line 172:** The phrase “their ability to simulate tsl” is too vague. The authors should detail why a particular performance in simulating soil temperature (tsl) is expected and what underlying processes or parameterizations support this expectation.
- **Line 174:** I am not a statistician, but the term “diversity” is not adequate in this context and in the rest of the manuscript for me. Could it be replaced with a more specific term such as “variability” or “spread” in model performance to better describe the differences observed?

  Yes, using “diversity” here is inappropriate. We will change it to “spread.”

- **Line 175:** Clarify what is meant by “with most sites lacking insulating snow.” The authors should specify the criteria or observations underpinning this statement and discuss how this influences the results.
- **Line 177:** Specify which models are being referred to at this point to avoid ambiguity.
- **Lines 207–208:** The sentence is redundant or not essential to the discussion.
- **Line 211:** The assertion that “Differences in grid cell scale among models can lead to biases in the tas state over the grid” is unclear. The authors should expand on this point—explaining how grid cell scale differences can affect biases in near-surface air temperature (tas)—and support the statement with references to the literature or numerical values.
- **Lines 218–220:** While the discussion of differences in tas EB and tsl EB between

LS3MIP and CMIP6 simulations offers an interesting perspective regarding the compensation between land surface and atmospheric processes, this section needs further clarification. The authors should:

- Provide explicit numerical values or clear graphical support for the claim.

We will provide more data (numbers) to support our analyses.

- Reconcile this discussion with other observations in the manuscript (e.g., the statement in line 234 regarding Group L versus Group C performance).
- Elaborate on the physical reasoning behind the offsetting errors observed in the CMIP6 ensemble.
- **Line 223:** The focus on the "shallow soil response" should be clarified and introduced earlier in the section or even in the introduction to offer better context to the reader.
- **Line 230:** The use of the term "better" is vague. A more precise descriptor or quantitative measure of comparative performance should replace it.
- **Line 234:** It is unclear whether the reported tsl values pertain to all seasons. The authors should clearly state which seasons are included and, if possible, quantify the differences observed.
- **"Climate Dependency of Modeled Temperatures" Section:** This entire section could be confusing as it references two kinds of "temperature values":

1. Cold/warm biases (often without specific numerical values)
2. Temperature values from observations or model outputs (without clear designation) It is recommended that the authors adopt a consistent approach similar to that used by Wang et al. (2016) and further in the manuscript, where temperature regimes are clearly defined and specific numerical ranges are provided for each regime.

- **Line 243:** The term "state" needs to be clearly defined. Furthermore, introductory words such as "So," (and "And" further in manscrupit) should be removed in favor of a more formal tone. This writing style is very very preoccupying.
- **Line 247:** The phrase "simulate similar histograms" is misleading because histograms represent sample counts.
- **Line 248:** The statement "However, a slight cold bias below -30 °C exists in Group L" requires clarification. If such a bias is observed, the authors should provide the exact numerical value, discuss its significance, and ensure the figure clearly demonstrates this bias.
- **Line 261:** The phrase "more likely distributed" is unclear. The authors should:
- Specify how the distribution of tsl was characterized.
- Indicate the specific temperature range of the cold bias.
- **Lines 266–268:** This passage needs to be rewritten for clarity. For instance:
- Clearly specify that the values refer to the minimum tas (if that is the case).
- Rewrite "than that of the land-only run."
- Explicitly describe how differences in the lower extremes of tas translate into corresponding gaps in tsl values, supporting this statement with numerical evidence.
- **Line 269:** There is a typographical error: "underestimate" should be corrected to "underestimating."

- **Lines 269–270:** The sentence is overly long and ambiguous. It should be restructured to clarify the differences between the "snow insulation effect" and the "surface insulation effect". Also, note that Dutch et al. (2022) discuss only CLM5.0. The authors should clearly explain whether the two effects are distinct or closely linked, and provide additional literature to support any claims regarding deficiencies in the modeled snow insulation effect.
- **Line 272:** Missing the term "flux".
- **Lines 273–274:** The claim—that a decrease in tas has a limited influence on tsl due to high snd, implying that the primary source of error stems from thermal conditions at the bottom of the soil column—requires stronger substantiation. More data, evidence, or references should be provided to support the assertion regarding the role of geothermal flux.
- **Lines 277–280:** The categorization of states (i.e., the thawed state, the freeze–thaw transition state, and the frozen state) needs further clarification. It is recommended to (a) reference literature that has adopted a similar categorization technique and (b) ensure consistency in nomenclature between the text and figures—for example, aligning terms like "Set Frozen, Set Intermediate, Set Warm" with the descriptive categories.
- **Line 282:** Remove "Results are shown in Fig. 9".
- **Line 287:** Clarify whether the observation that "the tsl samples are mainly concentrated in Set Intermediate and Set Warm" is directly evident from Fig. 9. If not, provide either numerical summaries or additional explanation within the text or figure caption.
- **Lines 290–292:** The claim that "the mean and minimum value of tsl bias is much lower than that of tas bias" and that this negative bias appears across all sets requires further explanation. The authors should provide precise numerical values and discuss how these figures support the conclusion that the land models simulate tsl as being too cold relative to expectations.
- **Lines 292–293:** The key point that improved tas accuracy in Group L models (in Set Frozen and Set Intermediate) does not necessarily yield better tsl simulations, and that tsl variability below –5 °C is higher in Group L than in Group C, needs to be expanded.
- **Line 303:** The statement that "the tsl gradually convergences near 0 °C" is not clearly supported by the observation figure. In addition, the phrase "and is primarily impacted by tas in a limited manner" appears contradictory. The authors should re-examine the data, reconcile these inconsistencies (especially in light of their earlier remark in line 273), and clearly articulate the influence of tas on tsl with supporting evidence.

  We will rephrase and perform additional calculations to provide supporting evidence.

- **Lines 305–307:** The discussion of snow shielding effects—claiming that thicker snow (snd > 0.3 m) strongly relates ΔT and tas—needs to be clarified. The authors should elaborate on how thicker snow modifies the impact of overlying tas and provide additional data or references to substantiate this relationship.
- **Lines 316–317:** The claim that "all other models fail to reproduce the observation-

like curve, underestimating the snow insulation effect under most conditions" seems overly generalized. For instance, CESM2 appears to capture the curve reasonably well. The discussion should differentiate among models and provide detailed evidence to support such claims.

- **Lines 319–321:** The explanation that low-resolution land surface models hinder accurate determination of surface organic matter distribution—thereby leading to errors in calculating the surface insulation effect—requires further detail. The authors should clarify the underlying processes, reference additional studies (e.g. 10.1175/JCLI-D-24-0267.1), and distinguish this issue from concerns related to snow insulation.
- **Lines 319–324:** This section appears to focus on surface insulation rather than directly addressing snow insulation. Given that other studies (e.g., 10.1038/s41467-019-11103-1) have highlighted that the warming effect of soil organic matter is less significant in winter because of insulating snow cover, it might be advisable to revise or remove this passage.
- **Line 325:** The statement "Similar conclusions can be made to HadGEM3" should be clarified by explicitly detailing which aspects of the analysis are similar and providing supporting evidence for this comparison.
- **Lines 334–340:** More numerical evidence is needed to support the claims made in this portion of the discussion. The authors should include specific numbers, statistical measures, or comparisons to strengthen their argument.
- **Line 340:** The phrasing "where a substantial reduction in the lack of snow insulation is seen" contains a confusing double negative.
- **Lines 341–368:** There is a lack of quantitative support throughout this portion. It is recommended to supplement the discussion with numerical values that back up the claims. In particular, when addressing snow thermal conductivity, compare the performance of different parameterization schemes (rather than solely presenting their mathematical formulations) and clarify the impact of low snow depth on thermal behavior.

Conclusions

- **Lines 373–382:** The authors need to substantially develop and clarify nearly every sentence in this section. Several statements are not consistently supported by the manuscript. For example:
- The last figure concerning CESM shows results that contradict the claim that "inaccurate inter-annual variability in the simulation of soil temperature by CMIP6 models is mainly caused by deficiencies in the land surface models and less inherited from atmospheric components." This discrepancy is not adequately addressed in the discussion.
- The statement that "biases in the land surface model even partially compensate for the influence of air temperature biases" lacks sufficient evidence or numerical backing.
- The claim that "better precipitation simulation does not ensure snow depth results improve, especially in winter and spring" is not clearly linked to the rest of the manuscript.
- Terms such as "weakness," "near-surface energy transport process," and "snow

amount" are imprecise.
- The recommendation for "further improvement of parameterization" is vague. The authors should identify which specific parameterizations (e.g., those related to snow dynamics, soil thermal properties, or energy exchange processes) require refinement.

References

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5), 1937-1958. https://doi.org/10.5194/gmd-9-1937-2016

Kuma, P., Bender, F. A.-M., & Jönsson, A. R. (2023). Climate model code genealogy and its relation to climate feedbacks and sensitivity. *Journal of Advances in Modeling Earth Systems*, 15(7), e2022MS003588. https://doi.org/10.1029/2022MS003588

Menard, C. B., Essery, R., Krinner, G., Arduini, G., Bartlett, P., Boone, A., Brutel-Vuilmet, C., Burke, E., Cuntz, M., Dai, Y., Decharme, B., Dutra, E., Fang, X., Fierz, C., Gusev, Y., Hagemann, S., Haverd, V., Kim, H., Lafaysse, M., Marke, T., Nasonova, O., Nitta, T., Niwano, M., Pomeroy, J., Schädler, G., Semenov, V. A., Smirnova, T., Strasser, U., Swenson, S., Turkov, D., Wever, N., and Yuan, H.: Scientific and Human Errors in a Snow Model Intercomparison, *Bulletin of the American Meteorological Society*, 102, E61–E79, https://doi.org/10.1175/BAMS-D-19-0329.1, 2021.

Wang, W., Rinke, A., Moore, J. C., Ji, D., Cui, X., Peng, S., Lawrence, D. M., McGuire, A. D., Burke, E. J., Chen, X., Decharme, B., Koven, C., MacDougall, A., Saito, K., Zhang, W., Alkama, R., Bohn, T. J., Ciais, P., Delire, C., Gouttevin, I., Hajima, T., Krinner, G., Lettenmaier, D. P., Miller, P. A., Smith, B., Sueyoshi, T., and Sherstiukov, A. B.: Evaluation of air–soil temperature relationships simulated by land surface models during winter across the permafrost region, The Cryosphere, 10, 1721–1737, https://doi.org/10.5194/tc-10-1721-2016, 2016.