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**.

General comments:

ERA5-Land: The presentation of the ERA5 data in the manuscript seems pointless. The purpose of the manuscript is not to evaluate ERA5 Land against station data. ERA5 Land is also not used for additional validation of the model simulations on a larger spatial scale than the station observations allow. I suggest to remove the ERA5 Land contributions in the manuscript, or actually make use of them within their limitations (which would then warrant their evaluation against the station data).

We have revised our approach to the presentation of ERA5-Land data. We no longer use ERA5-Land as a "benchmark" or reference dataset for model evaluation. Instead, ERA5-Land is now treated as a widely used dataset that serves as a complementary 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 selected LS3MIP simulations (~0.9°). This adjustment allows us to assess how differences in resolution affect the results. We will also expand the Data and Methods section to provide further context for these changes.

Discussion: The manuscript lacks a proper discussion of its results in a comprehensive way, which also leads to conclusions that seem to have no basis. In the introduction, you state that: (1) "We will analyze the discrepancies between the same model in CMIP6 and LS3MIP to quantify the bias and uncertainty present in frozen soil regions, attributing them to land surface models versus those resulting from atmospheric forcings. With identical and more realistic atmospheric conditions, we anticipate that the LS3MIP models will more accurately simulate soil conditions. If these models fail to produce soil variable outputs that align better with observed data than the CMIP6 simulations, it is regarded as an error in the land surface models." (2) "We will discuss the variations among different models of LS3MIP and try to establish a connection between model performance and their specific features." However, the manuscript ends after the presentation of the results, without coming back to the analysis you promise in a comprehensive way. Since the you do not have a discussion section in the manuscript, the conclusions need to contain this discussion (or you need to make a discussion section). Please come back to both points from the introduction, and establish conclusions for both points rooted in your results in an understandable way. Right now, some conclusions and discussion are scattered throughout the results, but it is hard to puzzle them together to a coherent picture.

Indeed, we recognize that the current structure of the manuscript and the content within the conclusions do not sufficiently address these two important points. To enhance clarity, we will separate the results and discussions. We will include discussions that directly respond to both points outlined in the introduction. The additional sections will focus on “Discrepancies Between CMIP6 and LS3MIP” and “Differences in LS3MIP Model

Performance".

Specific comments:

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.

1 Introduction

The introduction loosely strings together statements on Arctic climate change and its impacts on the Siberian permafrost region, then jumps to factors that determine permafrost thermal state, and finally barely introduces CMIP6 and LS3MIP. Without knowing all of these things already, and how they are related to the uncertainties eg in the permafrost carbon feedback from climate model projections of the future, it does not tell the reader much, and does not coherently argue the importance of this study. Please outline a clear relation between the facts mentioned in the introduction and the statements about what the paper means to do from the last two paragraphs. I suggest rewriting the introduction completely. In addition, I have a number of specific comments on the introduction below.

We will adjust the structure of the introduction and add content to clearly explain the role of permafrost in climate change, the role of climate models in permafrost research, and the challenges of permafrost modeling. This will provide a clearer logical chain explaining why we are conducting this research.

Line 20: This is rather vague, please give specific numbers to the magnitude of Arctic Amplification, and cite their sources.

The data presented originated from calculations of temperature increase rates derived from the CMIP6 model ensemble across different regions; however, these contents were not specifically detailed in the manuscript. Moreover, our claim of "twice" referred specifically to areas within the Arctic permafrost zone, not to the broader Arctic region. To clarify and provide accurate information in accordance with your feedback, we will revise this statement.

Line 20: While I don't doubt the numbers for Arctic climate change cited from the two papers, and I acknowledge that they are permafrost related publications, these aren't the papers that produced the numbers, and they are seriously outdated. Please cite more recent publications on climate change projections for the Arctic, and cite the direct sources.

Climate simulations indicate that the Arctic warmed at a rate of 0.66 ± 0.32 °C/decade

from 1979 to 2014 (Cai et al., 2021), and high-latitude warming in a moderate scenario of 1.2 to 5.3 °C from 2005 to 2100 (Koven et al., 2013).

Line 22: While it is certainly true that the most distinct impacts occur in the permafrost areas where temperatures are already close to zero, the statement seemingly has no relation to your manuscript, since you focus your detailed analysis on cold regions, not the warmer edge of the permafrost zone, so I don't see the relevance of that statement.

While it is true that our primary focus is in permafrost regions, many of our selected stations are situated at the warmer boundary of permafrost degradation. Furthermore, in Section 3.4 (Figure 9), we examine the implications of reaching the critical temperature threshold of 0°C on model simulation accuracy. Our aim is to investigate whether models may exhibit greater simulation errors under climate conditions characterized by significant permafrost degradation. Therefore, although our main analysis centers on cold regions, understanding these warmer boundaries is crucial for assessing overall model performance.

Line 27: Again, this is very vague and lacks an appropriate reference. Please clarify.

Large amounts of soil carbon are stored in permafrost (Tarnocai et al., 2009; Fuchs et al., 2018), and as permafrost thaws, the soil carbon may be released into the atmosphere at a faster rate (Schädel et al., 2018).

Line 28: The point about abrupt thaw is that from models. we can usually only estimate the carbon emission effects of gradual thaw, but the effects of abrupt thaw are expected to be substantially larger than those of gradual thaw. However, instead of saying that, you simply line up facts with no connection or argument. Please rephrase, and cite appropriate sources.

In environments such as lakes and wetlands, the impact of thawed carbon on climate is even more pronounced due to the low-oxygen conditions, further increasing the proportion of methane emitted along with other greenhouse gases (Koven et al., 2015; Walter Anthony et al., 2018). Processes such as thermokarst result in sudden thaw events that greatly enhance the decomposition and release of frozen soil carbon, potentially increasing carbon emissions by up to 50% (Abbott and Jones, 2015; Turetsky et al., 2019).

Line 33: There needs to be at least one general, bridging sentence on how heat transfer through the soil is simulated, and that the following paragraph speaks about modelling.

In land surface models, heat transfer through the soil is typically simulated as one-dimensional vertical transport. Models adopt their specific soil layering schemes, where the thickness of soil layers generally increases with depth. By calculating the water and thermal balance at different depths, land surface models can derive the current state of soil moisture and temperature.

Line 35: “There are differences in the time scales of major physical processes between the soil and the atmosphere.” Vague, please clarify what you mean.

The time-scales of important processes differ largely between the soil and the atmosphere. For example, key variables such as temperature and humidity in the near-surface atmosphere can fluctuate substantially over hours or even minutes. In contrast, changes in water and thermal states in the soil are much slower with depth; for example, at depths of tens of meters below permafrost, soil temperatures may not vary significantly for decades.

Lines 33-42: These two paragraphs are a weird mix of processes and conditions controlling heat transfer through the soil, and how these are represented in models. Please separate clearly.

Line 43: Please state what CMIP6 means.

Line 46: There are a number of papers that describe these advances that should be cited here.

We include here the following references: (Ekici et al., 2014; Chadburn et al., 2015; Decharme et al., 2016; Brunke et al., 2016; Jafarov and Schaefer, 2016; Guimberteau et al., 2018; Cuntz and Haverd, 2018; Damseaux et al., 2025).

Line 47: Please state what LS3MIP means. Also, introduce what LS3MIP aims to do before you dive into the protocol.

2.1 CMIP6 and LS3MIP Simulations

Line 75: Land models treat input data differently, and may require different forcing data sets per se. A table would be nice, in particular since you look at tas, which can be close to/identical to the forcing, or quite different, depending on model setup.

We will double check the GSWP3 data that used in LS3MIP, if the models show strong differences we will add more information in the results (e.g. in Figure 2).

Line 89: Ménard et al only show snow properties in their paper. The way you cite the paper implies all information in you table can be found there, which is not the case. Please clarify.

Line 89-90: This is very vague again, and the table misses some of the processes mentioned here. Eg how is vegetation represented, are the Arctic specific vegetation types, are there shrubs? Please clarify and expand your table.

We will add necessary information to Table 2, such as the vegetation type options/distribution, as it may affect snow accumulation, albedo, and surface soil insulation.

Line 96: I find this sentence misleading, it implies that models that consider the impact of surface organic matter with a focus on hydro-thermodynamics don't include a carbon cycle, which is for example wrong for CLM5. Please rephrase.

The sentence will be rephrased for clarification.

Table 2: Power Function and Quadratic Equation: What does that mean? Either explain somewhere, or use a more descriptive term. What does snow conductivity depend on in these equations?

The terms "power function" and "quadratic equation" refer to the mathematical formulations used to describe the relationship between snow thermal conductivity and snow density (see Menard et al., 2021 and Wang et al., 2016). However, to improve clarity and address your concerns, we have replaced this ambiguous phrasing with explicit references to the specific formulations.

3 Results and Discussions

3.1 Winter 2-m Temperature in Target Area

Line 152: The definition used in Lawrence and Slater is the generally accepted definition of permafrost. Quite a number of the stations denoted as circles are actually situated on permafrost. Please explain potential reasons why they are not categorized as permafrost using this definition on the station data.

The definition of permafrost, 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 most 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, we will come back to the method of Lawrence and Slater.

Figure 1: The two triangle stations are very hard to see. In general, the figure would convey more information if the stations were colored by bias in comparison to the modelled data instead of their own mean states. Also, It would be useful to show the permafrost boundaries either from Brown et al or Obu et al in the map.

Thank you for your feedback regarding the visibility of the triangle stations. We have increased the size of the triangles to improve their visibility. Additionally, we will use the modeled temperature as a base map and represent the differences between station data and model outputs with color coding for clarity. So far we find that the model ensemble will have some grids being recognized as sea as the resolution is much lower than ERA5 Land. We cannot have the bias between model ensemble and observation of some sites due to this (especially triangles sites). So we will remove those sites from Figure 1. We appreciate your suggestion to include permafrost boundaries from either Brown et al. or Obu et al.; however, adding this could clutter the figure, so we do not plan to include it.

Figure 2: Bars need to be broader, median positions are not visible. Also, the labels have no positions, which makes them meaningless. For precipitation and tas, we could learn a lot from seeing where GSWP3 is, since it is the forcing data.

Regarding the readability of the figure, we have made adjustments to ensure that the bars are wider and that median positions are clearly visible. And I have altered the positioning of labels.
Regarding GSWP3, it is important to note that all LS3MIP models' temperature and precipitation results essentially derive from GSWP3 at varying grid resolutions. Therefore, we believe it is unnecessary to include an additional box for GSWP3, as this information can still be referenced without it.

3.2 Model climatologies

Line 160: Looking at figure 2, I don't see that.

As we improved the readability of the figure it should be more clear now.

Line 162: This statement is only true for pr. The LSMs compute their own tas. How close that actually is to the forcing depends a lot on what forcing is used (eg temperature at a reference height, or 2m air temperature itself), and on how complex the calculation within the LSM is.

We will clarify this.

Line 169: What about snow, soil moisture, vegetation? There is a distinct difference between soil temperatures in general and TTOP, which refers to (1) mean annual temperatures and (2) the top of the permafrost table.

We will add the impact of snow, soil moisture, vegetation and the function of permafrost on soil temperature.

Line 171: What does model family mean? Is it based on similarity of the atmospheres, or based on the atmosphere and land components? In your example, both land and atmosphere components of the models you put into one family actually share code history, but since you do not even state if you refer to the LS3MIP or the CMIP6 simulations, the statement is unclear.

We refer to Kuma et al. (2023) about the definition of model family. And here we define it by their land components. We will make a clearer statement about this.

Line 179: What is the reason for this difference in snow? Precipitation is similar, at least for winter, and air temperatures differ, but are so far below zero that the difference seems irrelevant. What drives this? Precipitation and temperature in autumn? And why does it only occur for this one model?

We checked the HadGEM-LS3MIP data and found that in other seasons, the tas and pr

values of HadGEM-LS3MIP were not significantly different from those of other LS3MIP models (especially UKESM). Furthermore, in MAM and SON, the air temperature and soil temperature of HadGEM-LS3MIP were at higher levels in the model ensemble, but snd still showed the same high values as in DJF.
We also compared HadGEM-LS3MIP and UKESM-LS3MIP on a site-by-site basis, and the phenomenon of overestimation of snd was widely observed at most sites. We have not yet found a satisfactory explanation for this phenomenon. We believe that the phenomenon is likely related to the snow scheme of this model. We will upload our scripts on Zenodo for transparency and reproducibility.

Line 180: ±10 cm translates into a relative error of around 33%, which is massive! Please put into context.

3.2.1 Relative Spread and Relative Bias

Figure 3: Caption states you show all seasons, yet there is only winter and summer in the figure. Also, why is snow in winter similar between L and C even though precipitation differs considerably? Because the medians are the same, and that is what drives snow variability? Please expand.

We plotted the other seasons, but since we did not draw solid conclusion from these two seasons we decided to put them into appendix if necessary.
Figure 3 illustrates the relative spread of data over a 30-year period, representing the degree of variability. While precipitation is an important driver of snow accumulation, snow variability is also influenced by other factors such as air temperature and soil temperature. As that not all stations are located in permafrost regions. Therefore, higher temperatures and soil temperatures in certain models may limit the proportion of precipitation that accumulates as snow while they have higher relative spread of precipitation.
The ability of different land surface models to accumulate snow is constrained by their respective parameterization snow schemes. As shown in Figure 2, inter-annual variability for snow depth tends to be much lower than that for precipitation. C and L models derive their precipitation from different sources; thus their relative spreads naturally differ largely.

Figure 4: There is no shading. Correct the caption. Also, as Figure 3, this is not showing all seasons.

Line 186: In general, it is really hard to understand the summer parts of Figures 3 and 4 without an equivalent to figure 2. Maybe provide a summer version of figure 2 in the supplement. Specifically, I think this is meant to read "contrary to JJA where" or something similar. The sentence does not make sense as it is.

Line 193: “The pr in Group C exhibits more extensive group diversity than in Group L.”

Which is because in group L, the only difference between the different models is different interpolation of the forcing data set, which makes this statement meaningless.

We will delete this sentence.

Line 198: “the model’s bias is considered relatively small” I would suggest to rephrase that into something like "the model's performance is considered adequate", because if the IQR is big enough, very big relative biases could still lead to RBs around 1. In terms of model performance, because you only look at 30 years of data, I agree that this means model performance is adequate, however, the bias would not be small.

Line 200: “Almost all CMIP6 and LS3MIP models have a positive pr-bias but a smaller relative and non-systematic snd-bias in winter.” Since snow is not a pure winter phenomenon and snow build up starts in autumn, so I am not sure how much meaning this comparison has. This analysis needs to be extended to snow build up in autumn.

We will expand this analysis to whole snow accumulation period.

3.2.2 Spatial Heterogeneity

Figure 5: I find this figure extremely irritating. Figure out the orientation, and resort so that maybe there are eight rows and two columns, so that the figure can be read. Also, for tas, the spread in the CMIP ensemble is bigger than the spread in the LS3MIP ensemble. For tsl, it is the other way around. Why? Please expand in the manuscript text.

We have changed the structure of this figure. For better viewing, we will split it into 2 figures and make them vertical.
There is an explanation for this question in 218. It is due to the compensation phenomenon of climate models. In general, the land models have the opposite bias to their atmospheric models. And so they have more neutral and realistic results. Otherwise, if the atmospheric model has a warm bias and the land model also has a warm bias, it will produce unrealistically warm land. But if the models have the same forcing, the land models will show their natural bias.
Besides, modeled soil temperatures are more diverse in absolute value/average state (as in Figure 2). As ensemble spread counts for absolute values, tsl should has larger sizes than tas.

Line 218: Why would there be a compensating effect like that? The ensemble spread is not particularly strong in your figure. Please explain.

You can refer to our last reply. Please focus on the subplots in the upper left and upper right of Figure 5, which represent the tas simulated by the atmospheric model in CMIP6 and the tsl simulated by the LS3MIP land surface model (where the biggest circles are more than 10 °C standard deviation), respectively. In the model ensemble, the two have opposite biases at most sites in Siberia. In addition, you can observe the bias tendencies of each model in Figures 7 and 8, where this phenomenon can also be observed.

3.3 Permafrost Region

Figure 6: It is impossible to read the labels. If all variables are to be presented in one Taylor diagram, they need to be distinguishable.

We will separate the figure into two, one with tas and tsl, another with pr and snd (which will be put into appendix). Thus it can be more readable.

3.4 Climate Dependency of Modeled Temperatures

Line 245: In the figure caption, it says 50th quantile, eg median, instead of the mean, which actually makes more sense. Please check.

This is an oversight. For statistical reasons, we have used the median rather than the mean in this study.

Line 268: I think this needs to read “Four models …”

Line 270: The reference is misleading, Dutch et al 2022 only discuss simulations with CLM. Please correct.

Line 272: “There is an excessively low tsl shown in Fig.8, possibly due to insufficient geothermal (functions as upward energy flux from the bottom of soil columns). As the decrease in tas has a limited influence on the tsl through high snd, the main source of error is likely from the other side of energy transportation (thermal conditions in the bottom of the soil column).” If that was true, models that consider a non-zero flux condition at the lower boundary would have to perform better than those with zero flux conditions, which is not the case. The depth of the column plays an important role here, as eg discussed in Alexeev et al, 2007 https://doi.org/10.1029/2007GL029536 and more recently Hermoso de Mendoza et al (2020), https://doi.org/10.5194/gmd-13-1663-2020.

Yes, the low tsl may be due to other factors. We will carefully analyze it and provide a more reasonable explanation.

Line 275/276: What about the strong underestimation of variability in summer? What is the reason for that?

We didn’t add information about the variability in JJA here. However, in Figure 4 the LS3MIP underestimate summer tsl variability, and tas having almost identical spread with Obs. One reason could be without snow in summer, there is less uncertainty in tsl simulation. The models have less uncertainty considering impacts of soil conductivity and vegetation.

Line 285: “In contrast, …” I don’t understand that sentence. Please reformulate.

What I mean is that in LS3MIP's Set Intermediate and Set Warm, the standard deviation of the tas difference between model and obs is relatively small compared to CMIP6. And there is just a slight cold bias.

Line 292: That is a really important statement, it should be explicitly taken up in the conclusion, and the implications should be discussed!

3.5 Snow Insulation

Figure 10: It would be really useful to have horizontal grid lines (maybe in light grey) in the figures so the reader can better understand how close to the observed values models are.

Figure 10: CESM: This actually looks a lot better than what is Burke et al, 2020, for just winter. I wonder why.

Our assumption is that the CESM2 in the CMIP6 historical run has a very strong warm bias in tas. The CMIP6 results are much worse than LS3MIP for CESM2, and for the colder 2 categories, CESM2 doesn't have enough high snow depth samples.

If we focus only on the CMIP6 results of CESM2 and add up 3 categories, it should look like the result in Burke et al, 2020. As it goes from 5 to 10 ◦C in the -25 to -15 category.

Line 297: From your figure caption, I assume that you use monthly mean values from all months, not just the winter months, for your plot. However, I assume the classification is still based on the DJF 30 year average of the station?

No. For example, the obs sample size is 12months*30years*236stations. The classification is based on the air temperature value of every sample not on the DJF 30 year average of the station. We will clarify the caption

Line 303: "under sufficiently thick snow, the tsl gradually convergences near 0 ◦C and is primarily impacted by tas in a limited manner." I don't understand that statement. Please reformulate.

We will reformulate. What we meant here is that tsl has an almost constant value (0°C or lower) if there is about 0.2m or more snow depth.

Line 325: "UKESM1.0-LL consistently demonstrated similar snow insulation effects in both ensembles" From just looking at the figure, so does HadGEM, which is not surprising since the land models are similar. MIROC and IPSL also have very similar curves regardless of the forcing. Please quantify your distinction in model performance.

Thank you for pointing this out, we will revise the text and quantify the performance of the models more accurately.

Line 329: "Despite cold conditions, an increase in snd still affects the snow insulation effect of LS3MIP CESM2." I don't understand the statement. Please reformulate.

What meant was that in the two categories (-25 to -15 and less than -25), the snow insulation effect shows an increasing trend with increasing snow depth at all depths. This phenomenon is not evident in Obs. We will clarify that in the revision.

Line 336: I cannot follow this statement. Both in Wang et al 2016 and Burke et al 2020, previous versions of CLM5 (CLM45 stand alone in Wang et al, CLM4 in the CMIP5 analysis of CESM1 in the supplement of Burke et al) clearly outperform CLM5 with regard to the snow insulation curve. Please explain further what your statement is based on.

Burke et al. (2020) analyzed coupled models. And in Wang et al. (2016) the curve of category -5 ◦C to -15 ◦C is going up with a stronger trend than obs. Here we see a rather close distribution of snow insulation effect of CESM2-LS3MIP. However, as we didn't incorporate CLM4.5 data in this article, so we will remove this statement and reconsider.

3.6 Impact of Land Model Features on Performance

Line 343: "show good performance in reproducing accurate snd" Actually, in Figure 2, the observed median value for snow depth is within the interquartile range of 1!! model in the LS3MIP forced simulations that supposedly do not suffer from biased precipitation. I would not call that good performance. Please add context.

Line 345: "Although IPSL-CM6A-LR employs a simpler spectral averaged albedo scheme than other land surface models, it does not have an observable impact on its tsl simulation." What data in your analysis is this statement based on?

In Figures 2, 3, and 4, the tsl of IPSL-CM is not worse than other models in terms of bias and spread. Only the bias in JJA is slightly higher, but it is still within 1 times the Obs IQR. However, we will phrase this claim more carefully.

Line 348: While vegetation is certainly important for accurately calculating albedo, in terms of the surface energy balance in general, the timing of snow cover is important. Please discuss the impact of a wrong timing of the onset of snow cover and melt.

This is an important point, and we will include more analysis on this.

Line 349: "Considering snow conductivity, the Power Function could be why CNRM-CM6.1 and CNRM-ESM2.1 have a negative bias of larger than -6 °C in the SON (figure not shown)" In table2 , both the models with best snow insulation performance (the versions of JULES) and the model with the worst performance (Surfex) employ a power function, so this seems unlikely as the reason for the difference in performance. Please explain your conclusion in more detail.

In reviewing the references, we found conflicting information between Menard et al. (2021) and the primary references for JULES. Contrary to our initial understanding, both versions of JULES use a quadratic formulation of snow density to calculate thermal conductivity, as described in Wiltshire et al. (2020) and Calonne et al. (2011), rather than a power function. This correction requires a revision of our discussion, where we will re-evaluate the potential reasons for the observed biases in CNRM-CM6.1 and CNRM-ESM2.1.

Line 352: Especially in autumn, this could also be an effect of incorrect timing in snow. If snow cover is late in the models, the soil will release heat to the atmosphere for a

prolonged time, which could also explain an underestimation of soil temperatures. Since you have not looked at the timing of snow cover, and snow rmse is large for all models in autumn at least in comparison to the stations considered in Figure 6, I think you need to extend your statement.

Line 363: Since you cannot compare the performance of these models to versions that do not contain organic matter, I don't see how you can draw that conclusion. Please explain further.

We intended to refer to the "surface layer" rather than the "organic layer". In months without snow cover, the surface insulation effects shown by the models are lower than the observed value, which is probably due to insufficient/missing representation of surface organic matter. Therefore, we made this assumption. But other factors like vegetation and model soil texture could also have impacts, so we will rethink and rephrase our explanation.

4 Conclusions

Please see my general comment on what the conclusion should contain. Specific comments below.

Line 373: "Except in summer months, 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." What is the reasoning behind this conclusion?

As can be seen in Figure 5, the direction of model ensemble bias of CMIP6-tsl is commonly the same to LS3MIP-tsl but opposite to CMIP6-tas. And the ensemble standard deviation of LS3MIP-tsl is even larger than CMIP6-tas although LS3MIP is forced by same forcing, which indicates larger variability caused by the land surface model than the atmospheric model. We will add numbers here to support our conclusion.

Line 378: "The largest model biases of tas and tsl are witnessed under -5 °C." What does this refer to? Winter, summer, LS3MIP or CMIP6 models? And to what do the -5 °C refer? Climatological mean of winter temperature? MAGT? Please provide more context to explain the statement.

This conclusion is drawn from Figure 7/8/9. Regardless of season, when the temperature state itself is going under -5 °C, the standard deviation of Bias(model-Obs) grows. We will add more detailed explanation.

Line 379: "These indicate a weakness for models reproducing the tsl relationship with tas in freezing conditions" Which could point to deficiencies in soil moisture, which you have not discussed at all, even though it has a profound impact on latent heat during freeze and thaw. Please extend the discussion accodringly.

Line 381: "Land models tend to simulate lower tsl when overlying snow exists." Do you mean lower than observed? Because as a general statement, that is wrong. Please explain

more clearly.

Here the expression is not clear. We meant that the parametrization of snow insulation is insufficient for most models, so when there is snow, the model overestimate the energy loss from the land surface to the atmosphere. Causing the soil temperature to be lower than Obs under same tas/snd condition. We will rephrase.

Line 383: “Note that the scope of this study is limited to soil depths down to 0.2 m” You never state anywhere that all tsl metrics you show only refer to tsl at 20cm. Since the RosHydroMet data provides temperatures at 20, 40, 80, 160 and 320 cm depth, I assumed all metrics referred to comparisons of all depths, and that only the snow insulation analysis is restricted to tsl in 20cm depth as proposed in Wang et al., 2016. This would have to to be clearly stated in the data description, but actually, I don't see any good reason for excluding the other depths from the general analysis, especially because you argue the relevance of the soil temperature analysis with the climate change impacts on permafrost, and 20cm depth is above the active layer thickness in large parts of the northern hemisphere permafrost area. Please extend the tsl analysis using all depths from the station data.

We will expand our scope (adding figures to appendix), and to make sure we have a clear statement. Still we will mostly focus on 20cm depth.

References

Cai, Z., You, Q., Wu, F., Chen, H. W., Chen, D., & Cohen, J. (2021). Arctic warming revealed by multiple CMIP6 models: Evaluation of historical simulations and quantification of future projection uncertainties. Journal of Climate, 34(12), 4871-4892.

Calonne, N., Flin, F., Morin, S., Lesaffre, B., du Roscoat, S. R., & Geindreau, C. (2011). Numerical and experimental investigations of the effective thermal conductivity of snow. Geophysical research letters, 38(23). https://doi.org/10.1029/2011GL049234

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.

Koven, C. D., Riley, W. J., & Stern, A. (2013). Analysis of permafrost thermal dynamics and response to climate change in the CMIP5 Earth System Models. Journal of climate, 26(6), 1877-1900.

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.

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.

Wiltshire, A. J., Duran Rojas, M. C., Edwards, J. M., Gedney, N., Harper, A. B., Hartley, A. J., Hendry, M. A., Robertson, E., and Smout-Day, K.: JULES-GL7: the Global Land configuration of the Joint UK Land Environment Simulator version 7.0 and 7.2, Geosci. Model Dev., 13, 483–505, https://doi.org/10.5194/gmd-13-483-2020, 2020.