The manuscript presents an attempt to quantify the significance of the treatment of small-scale variations in soil properties on the simulated soil and surface temperatures, as well as the surface heat fluxes. The effects of heterogeneity and their impact on larger-scale averages are undoubtedly intriguing topics that warrant the attention of the modeling community. However, I think the manuscript would benefit from some revisions of the description of the model and analysis, and expanding of analysis to include the effects of interannual variability.

We are grateful to the reviewer for emphasizing the importance of this study and the constructive feedback. We will address the mentioned problems point by point.

To start, I would suggest changing the title of the manuscript. First, for readers not intimately familiar with the permafrost features and process, it is not at all obvious what "non-sorted circles" in the title refers to. As a very minimum, the title should make it clear that the study is specific to the permafrost processes. Second, the focus of the study is on the differences in results of several modeling approaches, but to the casual reader the term "aggregation error" implies comparison with observations (or the perfect representation of the processes), which is not the focus of the manuscript. I recommend avoiding this term, at least in the title.

We will think about another title, which will make the topic of our study clearer.

I think another problem is that the description of the model misses an essential part: the method used to calculate surface turbulent fluxes. Clearly, on such a small horizontal scale of ~10 cm that the described model uses, the Monin-Obukhov Similarity Theory (MOST) approach commonly employed in large-scale mosaic schemes would not work because a number of assumptions important for the MOST applicability are violated. Therefore, it is essential to describe what alternative approach was used to calculate surface fluxes, especially given that a significant portion of the manuscript is devoted to the analysis of differences in turbulent fluxes and energy balance. Without that, it is very hard to judge the validity of the results.

The method used to calculate the surface turbulent flues is indeed Monin-Obukhov Similarity Theory, but from our perspective, there are three reasons, which we can use MOST for our model: (1) All MOST assumptions are limited to the "atmospheric" part of our model that is only applied vertically (1D). Derived energy fluxes are then used for all columns individually. (2) For this specific study, we used only "bare soil" and neglected any topographical differences, which means no differences in height and/or roughness between columns. Snow is added to the soil scheme, which does not affect surface height in this model configuration. Consequently, soil roughness versus the reference height of 2m (lowest level taken from CRUNCEP data) is well enough the recommended value of 50. However, (3) if roughness differences) due to vegetation or topographical differences) are present, DynSoM couples MOST with a roughness sublayer parameterization following Harman and Finnigan (2007, 2008). We will clarify this in the model description.

The description also seems to contradict itself, saying in section 2.1.1 that "a prescribed skin temperature (see following section) serves as the upper boundary condition", while equation (3) implies that the surface skin temperature is calculated given the atmospheric meteorological forcing and prognostic equations of water and energy balance in the soil. I think providing more details about the calculation of surface fluxes and energy balance would help to resolve this confusion.

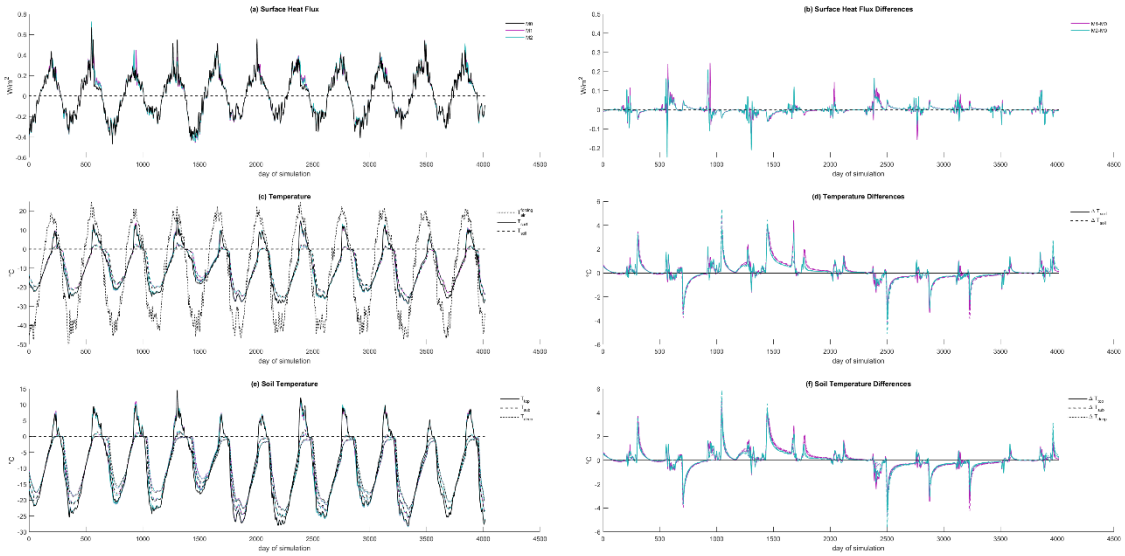
We took this method from Bonan (2019) and will refer to the specific chapter here.

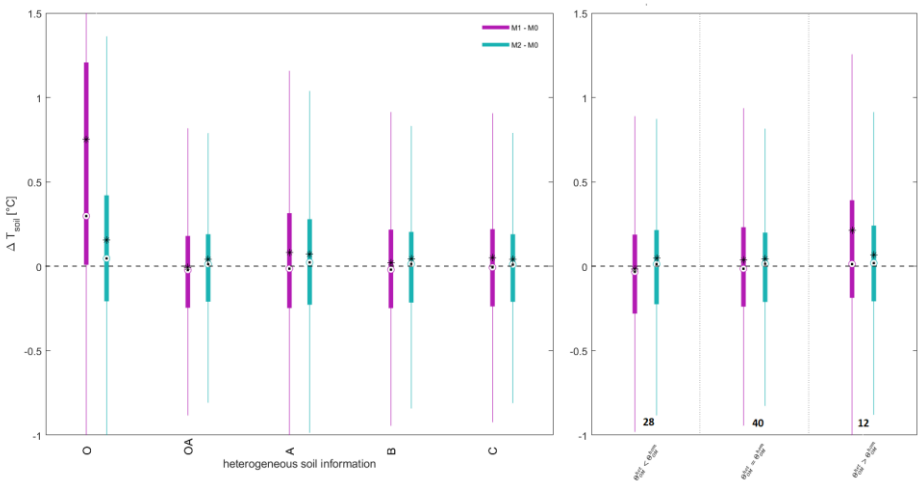
In the section describing the results of the simulations, the authors chose to limit the analysis of the results to just one year, but the motivation is not entirely clear to me: the text in lines 191-192 says "to preserve the current atmospheric forcing signal, which is strongly superimposed when another

averaging method is used." I am not sure about the meaning of this phrase; especially since it somewhat contradicts the one-week smoothing applied to the results to "avoid overly fuzzy near-surface" values. I think analyzing and presenting the statistics across all available years of forcing is essential to take into account the interannual variability and to have confidence in the robustness of the results.

This does not, of course, preclude using the results from one year (or one season) as an example illustrating physical processes, if necessary.

We agree that founding our study one a single year limits its significance. We will the full timeseries of figure 2 to the SI (see upper figure below this text) to show that the general observed pattern, i.e. simulation differences that are solely caused by soil heterogeneity and their increase by lateral heat fluxes, is not dependent on single years, whereby of course the absolute differences between models (inter-model ranges) differ across years. For figure clarity, we will leave out the intra-model ranges here, whereby these obviously also differ across years. To keep this information, we will replace figure 4 (see lower figure below this text as example; as well as related figures in the SI) by figures showing boxplots for the entire period and rewrite section 3.4 accordingly.





In the description of the energy-related fluxes and balances, the authors frequently use the confusing phrase “kW/m^2 per day”: the fluxes are typically measured in W/m^2, and it is absolutely not clear what “per day” refers to. Similarly confusing is the phrase “°C per day” in the description of the range of temperatures.

We will change this.

Assuming that my understanding of the units used in the analysis is correct, some of the energy balance numbers seem to be unreasonably large. For example, on lines 221-223, the manuscript says “the total simulated annual heat budget … (M0: 13.7kW/m2, M1: 14.1kW/m2, M2: 13.5kW/m2)”. Of course, the total long-term average energy balance at the surface of the well-spun-up land model should be close to zero, so it is not clear what these numbers represent.

I can only assume that these results represent the annual average sum of sensible and latent heat fluxes (which should compensate radiative fluxes), but even then the numbers seem excessively high. For comparison, the solar radiation incident on the area perpendicular to the sun rays at the top of the Earth atmosphere is ~1360 W/m2. It is not clear how it is possible that the annual heat balance at the site (~69N latitude) can be so large, given attenuating factors due to site latitude, annual averaging, and absorption/reflection/scattering by the atmosphere and clouds. Unless this is a typo, an explanation must be provided. Likewise, the intra-model differences in heat fluxes are on the order of hundreds of W/m2: that of course is not impossible on the short time scale, but would strongly depend on the way the turbulent fluxes are calculated, and needs to be discussed.

The reviewer is right. This is a unit error. All energy fluxes are actually given in W/m². We will change this in the manuscript.

In figure A1, soil water content is measured in kg/kg; this is kilogram of water per kilogram of what dry soil or wet soil? Or per dry soil+ice? Why the commonly accepted definition of volumetric water content is not used?

It is soil water (A1c & d) and soil ice (A1e & f) in kg per kg dry soil accordingly to Bonan (2019), chapter 16. We will clarify this.

Technical comments:

Line 38, and elsewhere replace “snow height” with “snow depth”

We will changes this.

L 171: Provide coordinates for Cherskii site

We will add this

L 178: Typo: “growthto” should be “growth to”

We will change this.

L 203: Replace “horizon-wise averaged” with “horizontally averaged”

We disagree here, because "horizontally averaged" would only imply the horizontal averaging over single rows, which have (in our model configuration) a vertical extension of 10cm, whereas the "horizon-wise averaging" that we applied implies a larger vertical extension (in our model configuration).

Caption of figure 2, and elsewhere: does T_surf refer to the temperature of the soil surface, or the surface interacting with the atmosphere (i.e. surface of the snowpack if present and soil surface T otherwise)?

Because snow is part of the soil in DynSoM T_surf refers to the surface that is directly interacting with the atmosphere. We will clarify this.

Caption of figure 2: "T_soil, all depths" — does it mean averaged over entire soil column?

Averaged until the depth of 1m. We will clarify this.