Overall the authors present an impressive study, which: 1) compares simulated permafrost dynamics in the CMIP6 models with an extensive collection of observational data sets, 2) assesses differences in model performance between CMIP5 and CMIP6, and 3) uses the CMIP6 models to derive a relationship between future increases to air temperatures and predicted volumetric thaw in the upper 2 m of permafrost-affected soils. The analysis is timely and important, and fits very well within the scope of the journal. The analyses described in the paper are generally rigorous and well-described.

My main suggestion for the paper is that the authors should put more effort into connecting modeling outcomes with real-world consequences. Currently the Discussion section is very short, and ends abruptly, after presenting a calculation related to the mass of frozen carbon that may thaw as air temperatures rise. The authors should describe what this number means—for example, how does 80-100 GtC compare with the size of other reservoirs, and what processes might cause that thawed carbon to leave the soil?

Additionally, the Discussion could do a better job of describing the limitations of using the present ensemble of CMIP models to predict permafrost and carbon dynamics. Currently, the second paragraph cites a lack of simulated thermokarst processes as a limitation to earth system models. I think it would be worth describing in a bit more detail how thermokarst might change climate projections, beyond accelerating permafrost degradation. For example, Liljedahl et al. (2016) describe how thermokarst in terrain with ice wedges often results in greater runoff and drier soils, which in turn may impact carbon fluxes out of the active layer. I would like to see a bit more discussion of how processes like this may contribute to climate model uncertainty.

Even outside of thermokarst-affected terrain, there is evidence that melting of excess ground ice affects permafrost thaw in ways that are difficult to capture in earth system models. For example, Shiklomanov et al. (2013) describe how thawing in the upper permafrost may cause uniform subsidence at the landscape scale, which isn't reflected in traditional ALT measurements such as those provided by CALM. This topic is worth mentioning in the Discussion as well, as it adds uncertainty to assessments of model performance against observational data.

Otherwise, most of my comments are related to presentation. The structure of the paper is logical, but at times the text is somewhat wordy and hard to follow. I make some suggestions for more specific changes in my detailed comments, listed by line number, below.

19-21) Please provide a bit more detail about how warming permafrost will impact each of these processes or systems. For example, it's more descriptive to say that fires in permafrost-affected areas will increase, rather than saying fires will be impacted.

33) Change to "within some of the models".

43) Delete the word "being"

35) This sentence sounds premature for the introduction. I'd change it to "We evaluate whether the sensitivity of permafrost to climate change…"

29) Delete the phrase "In fact"

53-54) The sentence beginning "Under increased global mean temperature…" is a bit hard to follow. Please rephrase.

56-57) You are referring to thermokarst in this sentence, but please be more specific. What are examples of landscapes changing in a hard-to-predict manner? This would be a good place to list a few types of thermokarst, and state that they are all caused by melting of excess ground ice.

63-64) I'm not sure that the sentence beginning with "The assessed soil diagnostics…" adds anything. Please reword it or remove it.

74-75) What does "high end of the range of future pathways" mean? Does this just refer to radiative forcing, or does it mean something more in the context of a Shared Socioeconomic Pathway?

105) Whenever you refer to mean annual ground temperature, please specify at what depth. Is this always at the top of the permafrost?

115-117) The first time I read this sentence it was unclear that you applied to relationship from Chadburn et al. both to observations and individually to each model. Please clarify this.

117-118) I wasn't sure what this sentence means. I think you are saying PFbenchmark is useful because it decouples the performance of the land surface module from biases introduced by the climate module. Please specify this if it is the case.

138) Please make sure MAGST is defined the first time you use the acronym.

156) I think you mean 0.25 m instead of cm.

164) MAGST is defined here, but it should have been earlier, in line 138.

175) Did Slater and Lawrence provide a probability of permafrost being present, if mean soil temperature at the deepest level was below 0?

218-220) Please provide more details about this method. How was the site-specific relationship between D and MAGT derived, and how was it extrapolated across each grid cell with permafrost?

223) Change "is" to "are."

233-252) This paragraph is quite long and hard to follow, especially with all the acronyms in it. I suggest breaking it into two or three paragraphs, and making sure each starts with a clear topic sentence.

279-280) This sentence was confusing. I think you mean that you binned MAAT at 0.5° resolution and calculated the median offset for each bin.

284) Change "to" to "too."

285) Please rewrite the sentence beginning "Figure S1.3 shows the variation…"

292) Delete "to be able" from the phrase "to be able to accurately represent…"

294-296) The sentence beginning with "The offsets are a function of…" is very unclear. Please rewrite it or break it into two or three shorter sentences.

328) Change the comma to a semicolon.

409) At what depth?

458-459) I recommend listing some of the northern high-latitude processes explored in these papers. Also, consider adding recent work by Langer et al. (2016), Aas

et al (2019), and Nitzbon et al (2019) on representing thermokarst lakes and ice wedges in land surface models.

Figure 3) If the orange lines come from the same observational data set, why are they different between the two panels?

Figure 5) Why do you plot the simulated summer and winter offsets separately, but only plot the surface offset for observations?

Figure 6) Why is this figure missing some of the subplots present in Figure 7?

References:

Aas KS, et al. 2019. Thaw processes in ice-rich permafrost landscapes represented with laterally coupled tiles in a land surface model. *The Cryosphere* **13**, 597-609.

Langer M, et al. 2016. Rapid degradation of permafrost underneath waterbodies in tundra landcapes—Toward a representation of thermokarst in land surface models. *Journal of Geophysical Research: Earth Surface* **121**, 2446-2470.

Liljedahl AK, et al. 2016. Pan-Arctic ice wedge degradation in warming permafrost and its influence on tundra hydrology. *Nature Geoscience* **9**, 312-318.

Nitzbon J, et al. 2019. Pathways of ice wedge degradation in polygonal tundra under different hydrologic conditions. *The Cryosphere* **13**, 1089-1123.

Shiklomanov NI, et al. 2013. Isotropic thaw subsidence in undisturbed permafrost landscapes. *Geophysical Research Letters* **40**: 6356-6361.