Significant uncertainty exists about how increasing temperatures and changing precipitation patterns will affect Arctic hydrological systems and, in turn, freshwater exports and associated biogeochemical fluxes to the oceans. Arctic hydrology is characterized by strong coupling between flow and thermal processes and is generally not well represented in Earth System Models. Rawlins and Karmalkar use the process-based model PWBM and two climate projections in a strong-warming scenario to assess changes in river flows across the Arctic. The study is well-designed and carefully executed, the manuscript is clear and well written, and the results will be of interest to the readership of TC. However, I have one concern/question that needs to be addressed.

Main question/concerns:

In Figure 1, which addresses model confirmation/evaluation for ALT, we can see good agreement in the mean between TPDC and PWBM forced by W5E5 (Figure 1d) but comparing 1a and 1b visually, it looks like the PWBM model is predicting significantly shallower active layer over the northernmost permafrost zone and deeper active layer in the southern parts of permafrost zone. In other words, TPDC and PWBM are producing very different trends in ALT with latitude. The fact that the two produce similar mean values is not an adequate criterion for judging the reliability of the model. An image showing the spatial distribution of the differences is needed here (e.g. like 1a and b, but for differences between TPDC and PWBM). In addition, a better metric would be the something like root mean-difference or similar metric that integrates differences across the permafrost zone. An explanation for the differences and the different trend is needed. If there is independent information available that could lend further support to the model result, that would help build confidence in PWBM's ALT calculation.

The authors thank the reviewer for their time and effort spent in evaluating the manuscript for publication in The Cryosphere. Several problems are inherent when evaluating regional or pan-Arctic distributions of simulated active-layer thickness (ALT) against observed ALT obtained from sparse in situ networks. First, in situ ALT is obtained at a point location that may not be representative of the region in which it is location. Second, observed ALT networks are very sparse across the terrestrial Arctic. Ran et al. (2022) presented an evaluation of the TPDC dataset and its new high-resolution estimates of the permafrost thermal state. The authors opine that TPDC dataset is appropriate for use in this study. Figure 2 in Ran et al. illustrates the dearth in in situ ALT, particularly across the cold mountainous areas of western Siberia and over the northern Canadian archipelago. Moreover, they stated:

"The ALT represents the hydrothermal state near the ground surface with more spatiotemporal heterogeneity than the MAGT, which represents the thermal state of the relatively deeper ground. The vulnerability of the near-surface ground to external disturbances associated with the inconsistency of the ALT measurement method may be one of the reasons for the large uncertainty in the prediction of the ALT. ***Of course, the uncertainty of ALT is considerable, especially in the vast area of western Siberia where the training data are sparse. The low spatial representativeness of training data may lead to an overestimation in several Siberian mountain regions and underestimation near the lower boundary of permafrost.*** *This highlights the importance and urgency of state."*

In summary, Ran et al. clearly stated that the distribution of ALT in the TPDC dataset is constrained.

Rawlins et al. (2013) examined simulated ALT against observations along a transect through central Alaska. Results confirmed that simulated ALT was unbiased. The PWBM has a rich history of use in characterizing land surface hydrology across the northern high latitudes. No evidence has suggested that the model is biased. The authors are confident in the quality of the simulation of soil freeze -thaw, particularly given inherent challenges in modeling pan-Arctic hydrology, and the few number of studies of this type. Recent large-scale modeling studies of coupled permafrost and hydrology reveal challenges in ALT simulation. Lawrence et al. (2019) found strong connections between snow density, forcing data, and ALT. They noted large differences in simulated permafrost distribution and ALT between two different forcing data sets used, which they suggested reveals an important aspect of uncertainty in permafrost modeling. They also found that CLM4.5, with its low-density snow, exhibited ALT that was unrealistically deep ALT (>1 m deep) across nearly the entire permafrost domain. Paquin et al. (2014) noted a tendency for the Canadian Regional Climate Model (CRCM5) to overestimate ALT compared to observed values at Circumpolar Active Layer Monitoring program (CALM) sites. They noted simulated ALTs were overestimated moderately in very cold climate of the Canadian Arctic Archipelago, with larger overestimation of ALT for CALM sites located inland, mostly along the Mackenzie River and Alaska.

Lawrence, D.M., Fisher, R.A., Koven, C.D., Oleson, K.W., Swenson, S.C., Bonan, G., Collier, N., Ghimire, B., van Kampenhout, L., Kennedy, D. and Kluzek, E., 2019. The Community Land Model version 5: Description of new features, benchmarking, and impact of forcing uncertainty. Journal of Advances in Modeling Earth Systems, 11(12), pp.4245-4287.

Paquin, J.P. and Sushama, L., 2015. On the Arctic near-surface permafrost and climate sensitivities to soil and snow model formulations in climate models. Climate Dynamics, 44, pp.203-228.

Rawlins, M.A., Nicolsky, D.J., McDonald, K.C. and Romanovsky, V.E., 2013. Simulating soil freeze/thaw dynamics with an improved pan-Arctic water balance model. Journal of Advances in Modeling Earth Systems, 5(4), pp.659-675.

Regarding model evaluation statistics, metrics which rely on squared differences are known to be problematic (Willmott et al., 2005; Hodson, 2022) Indeed the RMSE in particular is RMSE is inappropriate because it is a function of 3 characteristics of a set of errors, rather than of one (the average error). RMSE varies with the variability within the distribution of error magnitudes and with the square root of the number of errors, as well as with the average-error magnitude (MAE). Interpretation problems can thus arise because sums-of-squares-based statistics do not satisfy the triangle inequality (Willmott et al., 2009). The authors feel strongly that MAE is a more natural measure of average error, and evaluations and inter-comparisons between models and observations should be based upon it.

Hodson, T.O., 2022. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. Geoscientific Model Development, 15(14), pp.5481-5487.

Willmott, C.J. and Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 30(1), pp.79-82.

Willmott, C.J., Matsuura, K. and Robeson, S.M., 2009. Ambiguities inherent in sums-of-squares-based error statistics. Atmospheric Environment, 43(3), pp.749-752.

Other comments:

The manuscript has a good summary of the PWBM at about the right level of detail, but neglects one important piece of information: what is the spatial structure? I presume it's not fully 3D, but a collection of independent columns with parameterized landscape runoff and routing through a river network? A brief description would help.

The spatial domain encompassing the terrestrial pan-Arctic as defined in this study involves 35,693 grid cells of 25x25 km resolution. These are, indeed, 35,693 columns in which water and energy interact with the soil and vegetation. Implementation of a routing routine (not used in this study) would make the model 3D. Results herein are based on the interaction of soil physics and hydrology as it manifest in changes in runoff, both spatially, with depth, and with time, both a seasonal component and difference from recent past to end of century. Methods section will be augmented with additional language to provide additional detail for the interested reader.

It would be useful to know what fraction of the contributing area for the major rivers comes from non-permafrost regions. This information would allow the reader to judge whether the results are coming mostly from trends in precipitation or from deepening of the ALT in a warming climate.

The authors propose to compute the statistics and add statements based on them. However, the close correspondence between changes in simulated net precipitation and simulated runoff suggest that net precipitation, rather than de-watering permafrost, is forcing the changes. As described in the paper, and specifically shown in figure 7, the runoff ***increases*** will arise mostly from colder northern areas, which tend to be underlain by permafrost. Physically, a deepening ALT would tend to store more water that could potentially be evapotranspired, and advected away, at a later time. The author feel that the results clearly illustrate that changes in net precipitation---increases in colder areas where runoff/precipitation rates tend to be high because of frozen ground, and decreases in southerly areas of unfrozen ground, are the dominant factor. Deepening ALT is playing a role in the transition to proportionally more subsurface runoff and increasing flows in autumn. The latter change is supported by many recent studies that are based on in situ observations. The addition of the contributing area statistic may allow readers to gain insights, and so will be added.

I'm not sure what is meant by "seasonally maximum ALT" in Figs 1 and 4 as ALT is already the annual maximum thaw depth. Isn't this just ALT?

The authors appreciate the question regarding seasonally maximum ALT. Actually, ALT is the thickness of the active (thawed) layer at any time. For example, during the early part of the thawed (warm) season, the active layer is typically deepening. ALT in a given area may be, for example, 10 cm in mid

June, 20 cm in mid July, and a maximum of, say, 30 cm in mid to late August. Thus, the authors, as other researcher have done, feel it is important to make clear to the reader that the metric of most relevance for validation is the maximum depth that occurs during the thawed season.

The manuscript correctly notes that subsidence, which is neglected in the model, may result in more discharge. It may also be worth noting that the cited modeling study by Painter et al. (2023) was specific to polygonal tundra so the effect on large river basins will depend on the fraction of those basins that contain polygonal tundra.

This is a valid and important point. The modeling study of Painter et al. (2023) is an important contribution to the growing body of evidence thaw permafrost thaw is impacting Arctic terrestrial hydrology. The originally submitted manuscript did make reference to polygonal tundra (line 574). Additionally mention will be added in the Introduction section.