Anonymous Referee #1

The manuscript presents a process of new model development/improvement in frozen ground process. In general, it is clear and reasonable to me. However, the presentation needs proper revisions before it can be considered for publication in HESS. Please clarfiy the merits and limitations of the newly developed frozen ground model, since it is not clear that if this new empirical model can be applied to other cold regions or not.

Reply: We thank Referee #1 for the endorsement of the novelty of this manuscript. For the applicability of this model to other cold regions, we believe that the model has great potential to be applied in other cold regions. There are mainly three reasons, which we shall mention in our revised paper.

Firstly, our study site, the Hulu catchment, although small (23 km^2), has a large elevation gradient (from 2968m to 4955m), diverse landscapes (hillslope vegetation, riparian area, alpine desert, and glaciers), snowfall and snowmelt, and both permafrost and seasonal frozen-soil. Our newly developed model explicitly considered all these spatial and temporal heterogeneities, and eventually achieved excellent performance. With such a comprehensive modeling toolkit, the model has potential to be upscaled or transfer to other cold regions.

Secondly, we obtained the perceptual model from not only the observations and our expert knowledge at the Hulu catchment itself, but also widely considered the impact of frozen-soil on hydrological processes in other catchments, including the Zhamashike and Qilian (two nested sub-catchments of the upper Heihe), the headwater of Yellow River, and the Cape Bounty Arctic Watershed Observatory in Canada. Thus, we developed the model for the Hulu catchment in the context of larger scale observations.

Thirdly, the realism of the conceptual model was confirmed not only by streamflow measurement, but also by multi-source and multi-scale observations, particularly the freezing and thawing front in the soil, the lower limit of permafrost, and the trends in groundwater level variation.

Although our new model generally has great potential to be used in other cold regions, we should be cautious to arbitrarily use the model without any prior understanding of the modeling system. Since frozen-soil is merely one influential factor for cold region hydrology, there are other factors having notable impacts, which are intertwined with frozen-soil. This relates especially to the geology condition, which can have considerable impact on frozen-soil, but has large spatial heterogeneity, and where it is difficult to take measurements. Hence, before upscaling to other cold regions, we recommend to follow a vigorous modeling procedure, i.e expert-driven data analysis → qualitative perceptual model → quantitative conceptual model → testing of model realism.

Major comments:

(1) It is not really clear if this developed model can be applicable to other river basins. It seems to us that the model development is just from a very small cold river basin of northeast Tibetan Plateau, and the model contains quite a few empirical parameters that needs calibrations with in-situ observations. What is the evidence that this model can be suitable for other basins of Tibetan Plateau?

Reply: We thank Referee #1for this question, allowing us to further clarify the modeling details. On model's applicability in other cold regions, which is the same question as your previous one, we refer to the reply above. For the model empirical parameters, most of them are related to the freeze-thaw processes related to Stefan equation, including the thermal conductivity k (W/(m/K)), the water content as a decimal fraction of the dry soil weight ω, the bulk density of the soil ρ (kg/m^3), and the multiplier from air temperature to ground temperature during the freezing season. All these parameters have clear physical meanings, confirmed by previous studies with a good spatial distribution over the entire Tibetan Plateau (e.g. Zou et al., 2017; Ran et al., 2022). Due to the extreme complexity of soils in mountainous catchment, we still need to recalibrate their values while modeling other basins on the Tibetan Plateau. So, in general, our answer is Yes. The model has great potential to be used for other cold basins, but we should be cautious to transfer to other catchments, requiring a vigorous test, especially for specific geology conditions.

(2) Figure 1: please clearly describe the time periods of these observations in the figure caption.
Reply: We shall do this in the revised manuscript.

(3) Figure 4: Why the runoff is so small in the Arctic watershed? Please also explain the difference between the disturbed and undisturbed.
Reply: That is because the runoff was measured in a small Arctic watershed. We will give more explanation in the revised manuscript.

(4) Figure 5: The empirical parameters Ks changes at different periods from 80d to 60d. Is it diffcult to apply/transfer the model to other regions, since you always needs calibrations at different periods?
Reply: This is a very good question. Estimating the value of Ks in different catchments and periods is an intriguing scientific question for hydrologists. Brutsaert and Hiyama (2012) studied several Siberian basins ranging from 1,000 km2 to 100,000 km2, and found the Ks was surprisingly in the range of 45 ± 15 days. Brutsaert and Hiyama explained this phenomenon with the theory of catchment self-similarity and co-evolution of geomorphology and hydrology. The discontinuous recession we found in this study is novel. By comparing with other paired catchments, we identified and attributed the discontinuous recession to two periods recession. The first recession period was contributed by the groundwater in both permafrost and seasonal frozen-soil areas, and the second recession period was only contributed by the seasonal frozen-soil area. To our best

knowledge, even without the impacts from frozen-soil, it is difficult to give accurate Ks estimation in small catchments (less than 1,000 km^2). The calibration-free Ks in different periods in frozen-soil catchments are more difficult. Thus more studies are still required to further understand the spatial and temporal variation of Ks, not only for frozen-soil basins but also in other climate regions.

(5) Figure 6: Why are the simulated results only shown in one winter at 1974? How about the results in other years? You have never mentioned 1974 in the observations.
Reply: Figure 6 shows that discontinuous baseflow recession happened almost every year at the Zhamashike station. We only highlight the hydrograph in 1974 to illustrate the same phenomenon in Zhamashike station as Hulu catchment, since both basins have both permafrost and seasonal frozen-soil, but with very different size (5526 km^2 vs 23 km^2).

(6) Figure 8: It is not clear here. please explain the FLEX-w, FLEX-h, FLEX-d, FLEX-g in the figure caption.
Reply: We will do this in the revised manuscript.

(7) Figure 9, Figure 11 and Figure 10 can be merged. Particularly, Figure 9 and 11 should be put together for direct comparison. Similarly, a new result by FLEX-Topo-FS should be added to compare with Figure 10 that is the simulated result of freeze/thaw depth by FLEX-Topo. For the results in Figure 11, it seems that the new model simulates much more fluctuations that observed. What is the problem about the new model?
Reply: This is a good point. We will merge Figure 9 and 11 for direct comparison. For Figure 10, FLEX-Topo did not consider freeze/thaw processes, thus cannot be involved into comparison.

Related to your question about the larger number of fluctuations, we showed the detailed model evaluation metrics in Section 5.2.2: "The KGE of FLEX-Topo-FS in calibration was 0.78, which was the same as FLEX-Topo. But in validation, the performance was significantly improved, the KGE improved from 0.58 to 0.66, KGL was from 0.36 to 0.72, NSE was from 0.41 to 0.60, R^2 from 0.82 to 0.83, and RMSE was reduced from 0.95mm/d to 0.79mm/d. All the model evaluation criteria were improved. The most significant improvement was the baseflow simulation, and KGL was increased from 0.36 to 0.72. We also noted that the FLEX-Topo-FS model reproduced the spikes during the thawing in Figure 11. This confirms our conceptual model of a sequence of thawing breakthroughs, which trigger the sudden release of groundwater starting at lower elevations and progressing to higher landscape elements."

(8) Figure 12: the comparison (only showing the periodical variation) is not meaningful, since the time periods for the two graphs are different.
Reply: Groundwater fluctuation in natural catchments has strong periodicity, which can be observed in Figure 12. Groundwater variation does not show significant difference

among different years, which is especially true for the 25m well. Due to the extreme difficulty of continuous observation in this region, there was no groundwater measurement in 2011-2014. But due to the strong repeated temporal variation of groundwater level, we have good reason to believe the trends happened in 2016-2019 will also happen in 2011-2014. Moreover, this is a qualitative comparison, rather than a quantitative one, which we don't think is a big problem.

(9) Figure 13: Future projection on 2 degrees warmer is just too simple. Please use IPCC outputs or more scientific designs.

Reply: We thank Referee #1 for this suggestion. We also believe it is worthwhile to use IPCC future climate prediction for more robust projection for future changes. In this study, we used 2 degrees warmer more like using a toy model to illustrate how warming will impact on baseflow, to further verify the capability and robustness of the model itself. We agree that using future climate prediction to force model prediction deserves further study.

(10) Reference: This reference list has missed a lot of recent literatures in frozen ground modeling studies at Tibetan Plateau. Regarding this topic, there have been quite a few studies in past five years, at the headwaters of Yangtze, Yellow, Heihe, and other rivers.

Reply: We will add more relevant publications in revised manuscript.

References:

Brutsaert, W., and T. Hiyama (2012), The determination of permafrost thawing trends from long-term streamflow measurements with an application in eastern Siberia, J. Geophys. Res., 117, D22110, doi:10.1029/2012JD018344.

Ran, Y. Li, X., Cheng, G., Che, J., Aalto, J. Karjalainen, O. Hjort, J., Luoto, M., Jin, H., Obu, J., Hori, M., Yu, Q., Chang, X. (2022) New high-resolution estimates of the permafrost thermal state and hydrothermal conditions over the Northern Hemisphere. Earth System Science Data, 14, 865–884. Doi: 10.5194/essd-14-865-2022.

Zou, D., Zhao, L., Sheng, Y., Chen, J., Hu, G., Wu, T., et al. (2017). A new map of permafrost distribution on the Tibetan Plateau. Cryosphere. 11, 2527–2542. doi:10.5194/tc-11-2527-2017