The authors thank the reviewers for their constructive comments. The comments are shown in regular fonts (we added numbers), ***our responses are in bold italic, blue fonts***. *Changes made in the manuscript are printed in italic, underlined, blue fonts.* ***Our line references refer to the updated manuscript with track-changes.***

Reviewer #2: Abancó, Clàudia

1) General Comments:

The main topic discussed in this manuscript is the uncertainty on the hazard estimation of hydrologically-triggered landslides. It presents a new model at global scale (PHELS), that estimates the daily hazard of hydrologically-triggered landslides at a coarse resolution (36 km) at the same time that it estimates its uncertainity by generating ensemble simulations. The paper is focused on the analysis of the performance of the temporal component (hydrological predictors) of the hazard estimation, as the static part (landslide susceptibility) is based in a paper already published (Felsberg et al., 2022).

The manuscript analyses the potential of three main hydrological predictor variables: the daily rainfall, the 7-day antecedent rainfall index (ARI7) and the root-zone soil moisture content (rzmc), although it does not go into detail on the uncertainty on the obtention of these values. Specially the rzmc is a factor that is very sensitive to the input parametres of the Catchment Land Surface Model (CLSM), as for exemple the soil porosity. Although my expertise is not in data-driven models, I assume that the results could also be affected by this sensitivity, therefore I'd recomment that authors acknowledge that some uncertainty could be induced by the source of the hydrological predictors.

The topic is interesting and novel, since as the authors point out, the literature on quantification of hazard uncertainty is very scarce, and only some attempts to quantify uncertainty of susceptibility or rainfall thresholds uncertainty have been published. In general, I think the authors should further emphasize the main advantages and limitations of PHELS compared to other models, such as the ones that don't provide uncertainty.

The paper is very well written, clear and, even if in some parts some clarification may help (at least for the non-experts in the topic), it is in general easy to read.

The conclusions are consistent with the evidence and arguments presented. They address the main questions proposed.

The Figures and Tables are in general clear, and helpful to follow the paper.

We thank the reviewer for this positive feedback and the constructive comments. We clarified the manuscript following your specific comments, and extended the discussion to include a paragraph on uncertainty sources that are and are not taken into account, as well as to the applicability, advantages and limitations of the PHELS model framework. Since both were also suggested by reviewer #1, Ben Mirus, the new paragraphs are a combination of suggested discussion points:

This known incompleteness of the inventory not only influences the performance evaluation, but also adds to the uncertainty of the model fitting process (i.e. Equations 5-6). While the goodness of fit can be quantified (see Table 1) and theoretically propagated, it is still relative to the available inventory. Quantification of inventory-induced uncertainty requires very detailed or synthetic landslide inventories and has been subject of many studies for LSS (Steger et al., 2017; Lin et al., 2021) but less so for hazard assessment. PHELS does not account for such inventory-induced uncertainty, but it does include the uncertainties and within-grid-cell heterogeneity of input variables by using an ensemble approach. The latter allows to easily account for, e.g., the uncertainty in rainfall, which is directly available from ensemble weather prediction systems. Or to account for modeled soil moisture uncertainties, which can be obtained from ensemble land surface model simulations that are usually optimized to match the variations in observations. Nevertheless, models are always a simplification of real world conditions and the downscaling of coarse-scale model estimates to fine scale applications remains a challenge.

PHELS provides reliable insights into spatio-temporal patterns of landslide hazard but has limitations in the context of actual early warning systems. These usually require higher spatial resolution and temporal accuracy. The coarse spatial resolution would hence call for downscaling methods to obtain within grid-box distributions. And although we use the evaluation approach LSE3 because of time shifts and possible observation errors, the fact that peaks of hazard are often simulated within a 3-day window around a recorded LSE may also indicate a low temporal accuracy. For early warning systems the question moreover remains how to interpret or use the hazard uncertainty. Low enough uncertainty could be used as a secondary condition before warnings are issued to the public or the uncertainty could be directly communicated as is. However, ensemble measures such as the maximum predicted hazard ("worst case scenario") or the 90th quantile of ensemble hazard prediction might be easier to understand. While this study used PHELS with specific spatio-temporal resolution and input data, its adjustable, modular character makes PHELS a general framework for hazard estimation that can be tailored to specific purposes. If adequate landslide and hydrological data are available, it would therefore be possible to create a PHELS setup that is more suitable for local to regional landslide early warning systems. (Lines 376-398)

Specific comments:

2) L29: binary approach: for the landslide hazard assessment or for the empirical temporal probability?

Here, the binary approach refers to the temporal probability, thanks for pointing out that this was not clear. We merged the two paragraphs to emphasize that the second one still refers to the temporal probability. *(Line 32)*

3) L34-40: do all these refer to the root zone or some include also lower layers of the soil?

We use "soil moisture" in a general sense. The references use soil moisture at different levels and also from different data sources. Soil moisture information from satellites informs about the surface soil moisture (upper 5 cm of soil), in-situ observations reach

depths up to 1.40m, and studies that applied modelling use layers from surface to groundwater. We added this information in the manuscript. *The measures of soil moisture range from antecedent soil moisture (Mirus et al., 2018; Wicki et al., 2020) and increase in soil saturation (Wicki et al., 2020) to soil moisture of the day (Bordoni et al., 2020), and refer to different soil layers (surface: Ponziani et al. (2012); Brocca et al. (2016); Thomas et al. (2019); Bordoni et al. (2020), root-zone: Brocca et al. (2016); Mirus et al. (2018); Thomas et al. (2019); Wicki et al. (2020), groundwater: Uwihirwe et al. (2022)). (Lines 39-43)*

4) L99: CLSM- what is the source of the inputs of the model (e.g.: soil porosity)? Also, in what units is rzmc?

The input for CLSM-parameters of e.g. soil porosity comes from the U.S. General Soil Map (STATSGO2) and the Harmonized World Soil Database version 1.21, further details can be found in De Lannoy et al. 2014, and an overview is provided in Table 1 in Felsberg et al. 2022. The unit of rzmc is [m3/m3] as introduced in the introduction. To avoid future confusion, we also added the unit here in the methodology section. *[…] to simulate rzmc [m3/m3] (0-100 cm) […] (Line 106)*

5) L125: as the temporally dynamic soil moisture...or also ARI7?

Yes, indeed. We added this to the sentence. *While LSS could conceptually be considered a prior probability we opt to use it as a temporally static (but spatially varying) variable and implement it as a condition in a similar way as the temporally dynamic soil moisture, ARI7 and rainfall. (Lines 131-133)*

6) Table 1: I am not sure if ARI7&rzmc were not tested because conceptually they both represent the same parameter (soil moisture)? If this is the case, I am not sure this is correct, as ARI7 may imply infiltration of water to lower levels (and consequent instabilization of the slope due to the water table rise) while rmzc is only for the upper layer of the soil.

Indeed, this was our reasoning. Antecedent rainfall has in the literature often been used as a proxy for soil moisture and we expected the information content to be redundant. For land surface models it is known that lower level soil moisture and rzmc strongly correlate and the inclusion of total water storage in addition to rzmc could hence be expected to have negligible amount of new information. Since ARI however includes additional information on rainfall, it might be worthwhile to investigate the combination of ARI and rzmc. We added a sentence on this in the discussion: *Furthermore different combinations of the predictors could be tested, e.g. rzmc and ARI7. (Lines 309-310)*

7) L186: As I understand, PHELS is trained with all the LSE for 2007-2020, and tested with the same data? Do you think this could induce some misleading results in the evaluation?

Yes, we evaluate against the set of LSE that was also used to derive the parameter values for the PHELS equations. In the discussion we already have a paragraph explaining why we do not expect misleading results from this approach, and also give an outlook about possible other approaches *(Line 287)*.

8) Figure 4: What about the high H values in 2016? All the predictors show high H around 2016-02 but no LSE is recorded. Would they be false alarms or a limitation of the GLC?

Theoretically, both are possible. Thanks to reviewer #1, Ben Mirus (see his comment 11), we however got reminded that the "Seattle Area experienced widespread landsliding in mid-late January 2016" (Mirus et al. 2018), i.e. it is a limitation of the GLC. We now mention this fact in the discussion: *Figure 4 for example misses mid-late January events of 2016 in the Seattle area that were reported by Mirus et al. (2018). (Lines 373-374)*

9) Figure 6: This is an interesting Figure!

Thanks for this feedback!

10) Figure 8a: Again, going back to the rzmc absolute values. I can see values around 0.5 m3/m3 in SE Asia, that would correspond to soil porosity of 0.5 (considering that the soil is fully saturated). These are very high values for porosity, only typical for some sort of coarse sand or silt, but not common. I have noticed this in SMAP-L4 products derived from CLSM, and in my opinion is a limitation that the use of global models have. I would only make a point here to say that absolute values of rzmc may be overestimated due to this

Thank you for sharing this insight. Indeed, global models lack accuracy, especially when it comes to soil characteristics and geology. With our approach to transform the absolute values of rzmc into percentiles according to the long-term climatology, we hopefully reduce the effect that a positive bias, i.e. a consistent overestimation, would have on landslide hazard estimation.

11) L332: I think this could be because ARI7 is actually giving a larger picture of the mechanical process going on in the slopes and closely related to the instability process. So, I agree that soil moisture of the upper layer is not always a good indicator.

This is a good point and we rephrased the sentence as follows: *While the number of missed alarms is lower for PHELS based on rainfall&rzmc than for PHELS based on rainfall alone, it is even lower for PHELS based on ARI7 . A possible reason for this can be that (intensive) antecedent rainfall prepares failure by progressively destabilizing the slope. (Lines 354-357)*

12) Conclusions: as I said earlier, in the discussion/conclusions I miss some emphases on the advantages and limitations of PHELS over other models, i.e: the applicability of such a model

We now included a paragraph on the advantages, limitations and applicability of PHELS in the discussion (see second paragraph in reply to comment 1)