This manuscript concisely presents a new approach to probabilistic landslide hazard modeling, which leverages a susceptibility model published previously in NHESS (Felsberg et al., 2022) with ensemble modeling to evaluate the suitability of a few alternative predictor variables for hydrologically triggered landslides. This PHELS model is applied at the global scale with coarse spatial and temporal resolution. Of the hydrologic predictors considered, rainfall with root zone soil moisture performed better than either of those variables alone and slightly better than an antecedent rainfall index used in the LHASA model (Stanley et al., *Frontiers*, 2021). When using a sparse global landslide catalogue, the output compares favorably to this existing model and the spatial and temporal variability in performance is shown, which reveals that uncertainty is lower during wetter seasons than drier ones. Furthermore, the probabilistic analysis reveals that very high and very low hazard predictions are well constrained, whereas the combinations of moderate susceptibility and moderate triggering conditions exhibit far greater uncertainties in landslide hazard predictions.

Overall, the topic is of considerable interest to NHESS readers, and this is a nice piece of work that presents several notable contributions, which ultimately warrants publication. Specifically, the approach for incorporating uncertainty in spatial and temporal probability of landsliding is novel and broadly applicable to hazard modeling, the evaluation of multiple predictor variables for hydrologic triggering is interesting, and the seasonal and spatial analysis as well as comparison of results to other global-scale analyses is useful. While these results are not particularly surprising and follow somewhat logically from the methods and input data, the work is technically sound and a useful reference. Methods and data are clearly described, such that results should be readily reproducible and the methods applied successfully with other data. The primary areas for improvement are largely editorial and include adding a more involved interpretation and context for the results, as well as some minor details in the presentation and figures. Therefore, the paper should be published after undergoing some minor revisions to address the following general and specific comments.

In general, the authors should make a little more effort to discuss what is interesting and useful about the results in the context of other studies. Readers should be confronted with both the advantages of the approach and its application, as well as its limitations, so that the utility and value of the resulting PHELS model is more apparent. For example, the framework for uncertainty assessment provides a robust approach to integrate forecast uncertainty that would be an important methodological advance for regional-scale landslide warning, but at the same time the global application seems of limited value for practical implementation locally and the reasonably strong performance is likely linked to the coarse spatial and temporal resolution and the decision to exclude a 6-day window for selecting non-triggering rainfall conditions. At the same time, the finding that the combination of daily rainfall with root zone soil moisture is more effective than seven-day rainfall index used in the LHASA model is consistent with recent advances in local landslide warning that leverage in-situ monitoring in favor of antecedent rainfall for reducing failed and false alarms, so it is useful to know that this potentially applies globally. Lastly, the revised version should include some discussion of why the model at such coarse resolution is useful, how that affects performance, and the limits of that utility.

In terms of presentation, the writing is clear and most of the figures are great. To address these and the general comments above, I have included the following specific comments and suggestions by line number:

L18. Specify: "… a two-step approach that separately evaluates where and when landsliding will occur." Also, why are you referencing the first approach (Stanley et al papers) and not a long list of others using the two-step, particularly since that is your approach here?

L23. It is misleading to imply that LHASA is specifically for landslide early warning since the rainfall data includes some latency and can at best be considered a "now-cast" of potential landslide conditions (i.e., see title of Stanley et al., 2021). At this point assessments that combine when and where landslides are likely cannot be used in a predictive mode for warning and their value for real-time hazard assessments is unclear. Furthermore, these are not typically ideal hazard assessments, in that they do not explicitly account for the magnitude and mobility of the potential landsliding, which is a significant consideration.

L25. Maybe it's worth mentioning that in addition to relying on a single threshold for landslide initiation (e.g., the 95%-ile of the ARI7 predictor at each grid cell), these also employ some susceptibility threshold that also amounts to a simple binary yes/no for landslide occurrence (though different thresholds of moderate or high have been used for different applications of the model). The value of your study is combining both within a probabilistic framework for both potential and initiation, so highlighting that here is worthwhile even if it's stated elsewhere.

L27. Whiteley et al (2019) is a very nice review of geophysical methods for landslide monitoring, but it's not exactly an appropriate reference for physically based modeling of landslide initiation potential. If I understand correctly, there are a lot of more appropriate examples (TRIGRS, Baum et al., *JGR-ES*, 2010; SCOOPS3D, Brien and Reid, *Rev. Eng. Geol*, 2008; SHALSTAB, Montgomery and Dietrich, *WRR*, 1994; etc., or review of this in the textbook Lu and Godt, 2013).

L45. Ok, here you define your limits of hazard modeling, but potentially should acknowledge that an ideal hazard assessment also includes the magnitude and mobility to determine where hazards are present, but in my view that's acceptable since at such coarse resolution that's not really as relevant.

L60. Here the use of "magnitude" is misleading. Be more specific and careful to clarify that you are evaluating the degree to which the spatiotemporal potential for landsliding is related to the uncertainty. There are lots of different ways you could consider "magnitude" of the event, from the size and velocity of the landslides to the number or extent of landslides or even their severity.

L71. Isn't there also a bias towards landslides in areas where they have an impact (i.e., roads, developed areas), as well as for landslide types that have a more notable impact (e.g., debris flows and shallow mass movements that affect infrastructure)?

L100. This seems to imply that most landslides (in the inventory?) fail at <1m depth. Is that supported by analysis of the GLC? Again, see comment earlier about dataset bias: is the GLC indeed mostly shallow landslides?

L191. This is a major assumption that warrants further discussion. Even if others have used this +/- 3days method before, I'm not convinced it is appropriate since you don't need a model to predict landslides when it's not raining hard. It seems to undermine the value of the model predictions by weighting the non-predictions to times when it's not really raining. A landslide hazard assessment tool really needs to be able to distinguish between triggering and non-triggering conditions *when it's actually raining hard* which your uncertainty assessments show it struggles most with in all but the most extremely high and low hazard levels.

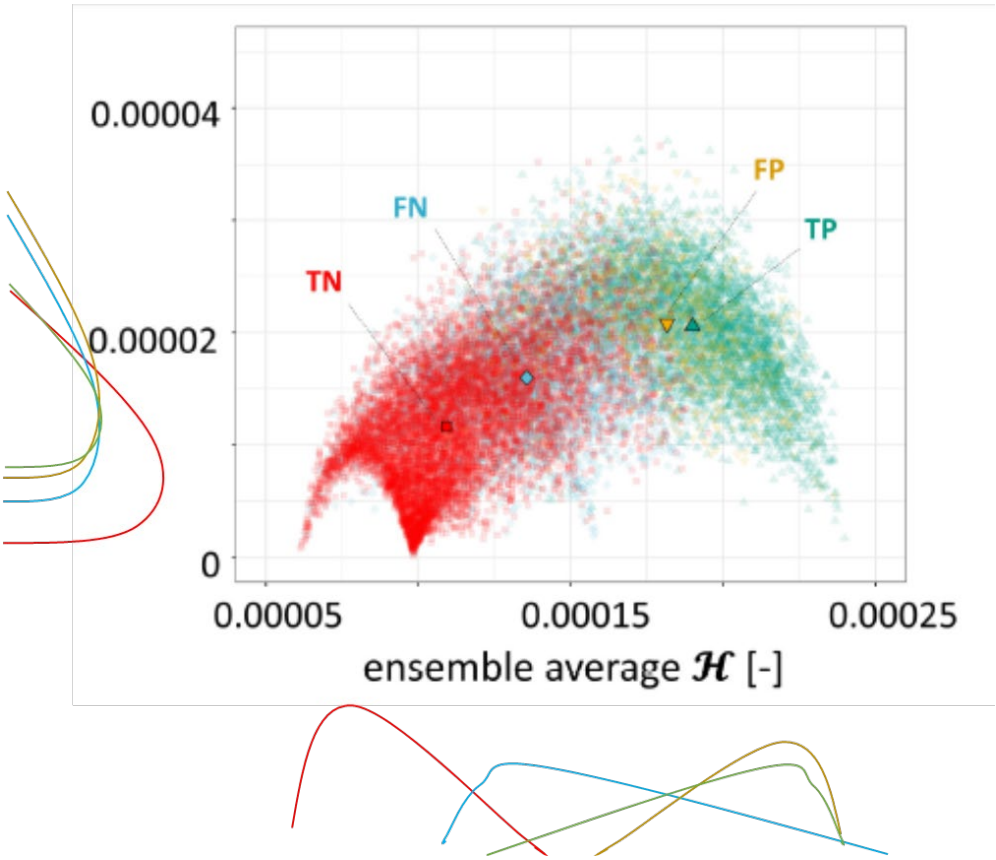
Figure 4. The greater Seattle Area experienced widespread landsliding in mid-late January 2016, including several mentioned explicitly in our paper (Mirus et al., *Landslides*, 2018) and I think also elsewhere (see Luna and Korup, *GRL*, 2022). This is an interesting opportunity to show an example where the inventory is not just incomplete in space, but also in time and how that influences results.

Separately, this figure does a nice job of visualizing why a combination of RZSM and daily rainfall is a valuable combination.

L238. This is more interesting in that it shows that the cumulative rainfall variables are not necessarily the best predictors for landslide triggering, instead there are likely some sub-daily rainfall characteristics (e.g., three-hour rainfall intensity, see Patton et al., *NHESS,* 2023). In contrast, the predictors that integrate the hydrologic variable, RZSM, may better capture those, albeit not explicitly.
Still, relying on a +/- 3-day window to obtain great model performance really undermines some of the potential utility of such a tool. Emergency planners and the general public don't really want to be on high alert ready to take action for 6 days.

L265. Can the discussion include any conjecture on why these spatial variability in uncertainty? Is it all due to the triggering not capturing the type of landsides, is it greater combined uncertainty in both the susceptibility and initiation? Is it data limitations?

Figure 10b. This figure is visually very challenging and not particularly accessible for individuals with red-green color blindness. Can you further reinforce this with a better color scheme and show TN, FP, FN, and FP with distribution curves of H and standard deviation on the x and y axes, respectively?



L285. Is this accounting for errors the +/- 3 day window? Again, this undermines the practical value of this type of model if it can only perform well at capturing a landslide event within a 6-day window, particularly given the coarse spatial resolution.

L330. Yes, we also found replacing antecedent rainfall with antecedent soil moisture decreased failed and false alarms (Mirus et al., *Landslides*, 2018).

L331. I don't quite follow this logic. ARI7 uses a weighted averaging daily rainfall over 7 days… how does that capture sub-daily bursts in rainfall intensity that may trigger landslides?

L334-345. Interesting, but I would think that with the same dataset a coarser resolution model would likely perform better, particularly for FPR and particularly for isolated landslides.

L346-348. Yes, the GLC is incomplete both spatially and temporally, see previous comment about Seattle-area landslides in January 2016. Are there areas where it is more or less complete that you could compare to assess this?

L349. The discussion section would benefit from presenting the sources of uncertainty are not considered and which of those potentially could be included. You are able to consider the uncertainty in susceptibility and triggering conditions since they can be quantified. However, the uncertainty due to incomplete inventories in space and time is not considered and could only be done if there were appropriate data to support this. Conversely, could the framework integrate weather forecasts and incorporate uncertainty in those forecasts relative to triggering conditions identified with the MERRA2 data?

P.S. The animation is a nice bonus of this paper.