Review of: "How size and trigger matter: analyzing rainfall-and earthquake-triggered landslide inventories 1 and their causal relation in the Koshi River basin, Central Himalaya" for NHESS.

Summary

The paper addresses an interesting problem – the differing controls on earthquake versus rainfall triggered landslides. It provides a novel contribution by taking advantage of the co-location of many rainfall and earthquake triggered landslides in the Khosi catchment. In this respect the study is extremely exciting, offering the possibility to examine the influence of the different triggers while controlling for landscape properties. Given this, I think considerably more could have been made of the results both by examining the parameters used in the logistic regression in more detail, and by examining spatial susceptibility patterns at a finer scale. Finally, the examination of how landslide susceptibility varies with size is an interesting idea but I have major reservations (detailed in MC4) about how robust these findings are at present.

Major comments

MC1. One of the major difficulties for this study is comparing an event inventory (associated with an earthquake and its aftershocks) with a long-term 'historical' inventory that combines landslides triggered by multiple storms. This is unavoidable and does not prevent the study from having value, but it does introduce comparability problems between the inventories and the susceptibility maps that they generate. These problems need to be treated much more explicitly in the paper.

MC2. More discussion (and perhaps more analysis) is needed on how and why the earthquake and rainfall triggered inventories differ (e.g. L404). This is picked up by R1 in Detailed7 but has not been fully addressed. The key question is whether the difference is simply a result of the different spatial distributions of triggering intensity or of different triggering processes. The earthquake inventory is limited to a fairly small part of the study area where shaking intensities in the Gorkha earthquake were high. The ETL susceptibility map would perform poorly for almost any other earthquake you could choose. Differences related to the different processes should be visible in the relative importance of different predictor variables. You don't really examine this in your results (other than one sentence L276-8) but I think this is important to do. You should then discuss the relative importance of different predictors and how these compare to findings from other rainfall and earthquake triggered landslide studies (e.g. proximity to ridge crests, the relative importance of slope and aspect etc). This is particularly important because these parameters are the best and perhaps only tool that you have to address your aim of understanding how susceptibility differs with different types of trigger. The susceptibility maps themselves are in my opinion too strongly influenced by the particular trigger patterns (especially in the earthquake case).

MC3. Landslide susceptibility features throughout the paper and is central to your conclusions. It would be useful to include: a definition of landslide susceptibility in the introduction (see also R1Detailed3, which is only partially addressed); and a paragraph in the results that helps the reader to interpret the landslide susceptibility classes. Are these purely relative classes or is there some absolute interpretation to the values? If so what is the basis for it, if not how should we compare RTL and ETL susceptibilities or those for different size classes? What are the landslide susceptibility maps showing susceptibility to? I think this is to landslide initiation, in which case the step to hazard (being hit by a landslide) involves an assumption that locations with higher susceptibility have higher hazard. This is likely to be true at some spatial scale of aggregation but may not be true at the finest ~12 m resolution of your input data.

MC4. I think you are conflating the effect of landslide size and the effect of sample size in your analysis. To properly test whether large landslide predictors are particularly dissimilar from small landslide predictors you should also examine the extent to which small landslide predictors differ from one another. You could do this by taking multiple (m) subsamples of the small landslide dataset (with n=355 landslides per sample i.e. equal to the number of large landslides), then repeat your susceptibility analysis m times using each subsample. The resultant distribution of landslide predictors (i.e. the distribution of m values for each logistic regression parameter) could be compared to the large landslide predictors (i.e. the value of each logistic regression parameter for large landslides) using standard statistical testing. Without doing this I don't think you can make your key concluding claim that: "the resulting susceptibility patterns are quite different, and it is therefore questionable whether landslide susceptibility maps that are generated for all landslide size would be able to accurately predict the large landslides." (L424).

Detailed comments

L40, it would be useful if you could define 'size', for some it may be synonymous with 'volume'

L54, these explanations of differing size-frequency distributions are methodological but there are have also been attempts at mechanistic explanations. R1 points this out in Detailed Comment 20 but this part of the comment has not been addressed.

L144, "from which the post 1992 and pre-2015 144 landslides" there is a word missing at the end of this sentence.

L151, what is the resolution of the inventory (i.e. minimum mapable landslide size) and what are the planimetric errors for the boundaries of landslide polygons?

L159, as R1 points out (R1Detailed13), taking a single pixel as source or scar zone may bias your statistics. Why not considering a scar surface in the upper part of the polygon (as others have)?

L164, what algorithms did you use to generate these different metrics? Add references.

L164, "Streams and gullies were obtained through DEM processing" add details of the processing including parameter choices e.g. channel initiation threshold. R1 made this point in their previous review (R1Detailed15) but it has not been addressed.

L189, as R1 previously asked: "What is relative relief, computed at which scale? Same drainage density? Distance to fault, which faults?" (R1Detailed17), these questions have not been addressed.

L178, how did you allocate landslides to training or validation sets? This point was also raised by Dr Scaringi in comments on a previous draft but has not been fully addressed. Are these random samples? What additional constraints are you putting on the sampling (e.g. retain the original size distribution)? Did you test the sensitivity of your results to the particular random sample? E.g. by repeating the analysis for a second realisation of the training-validation partition? Given the problems I suggest either demonstrating that several realisations of your cross-validation give the same result or using a more robust approach (e.g. k-fold cross-validation or similar).

L183: How and why were these size groups chosen and what were the boundaries? Why did you choose only two groups? I don't necessarily disagree with the choice but I think that it needs to be explained. You go on to do this in section 5.1 so you just need to point the reader to that section here.

L184: a little more information is needed, what types of bivariate analysis?

L190, I don't think this sentence is correct: "For the susceptibility modeling of ETL, precipitation during monsoon(x10) was instead of peak ground acceleration (x10)."

L244, Frequency ratio estimates can be very sensitive to sample size within each conditioning factor group. I suggest looking at the approach of Rault et al. (2018) as a way to identify bins where sample size restricts your confidence in the frequency ratio estimate (details are in Supp info).

L245, how did you choose the number (and boundaries) of groups associated with each conditioning factor?

L245, I don't understand how you get: E, the area of landslides in the conditioning factor group and F, the area of landslides if your landslide dataset has been transformed into a set of points for the highest part of each landslide (L159).

L253, the comment re L244 is particularly true for multi-dimensional frequency ratio analysis so I would suggest the same approach here.

L268, what is the resolution of the susceptibility maps that you generate and how do you handle differences in the resolution of your predictor variables?

L286, fig 6, what data did you use to define the fault traces that you use in your distance to fault map fig6f? What choices did you make about the types of fault (e.g. age, slip history, size) that should be included?

L294, table 2, it would be useful to include the description of the predictor (e.g. slope) in the column headings.

L363, I don't think it is a good idea to combine discussion and conclusions. A summary of your main findings would considerably improve the paper.

L406, I don't think this is strong enough, 'may not be likely' the probability of the next earthquake producing the same shaking pattern is tiny!

L409, "However, using PGA values based on probabilistic seismic hazard assessment might result is relatively poor statistical correlations with event-based inventories." This needs explaining and or supporting with a reference.

References

Claire Rault, Alexandra Robert, Odin Marc, Niels Hovius, and Patrick Meunier, Seismic and geologic controls on spatial clustering of landslides in three large earthquakes, 2018, https://www.earth-surf-dynam-discuss.net/esurf-2018-82/