Response to Reviewer 1 (Dr Odin Marc) by Milledge et al.
We thank the Dr Marc for his careful and helpful review which we feel has considerably improved the article. In our response below reviewer comments are in normal text, and our replies are in bold.

Summary

Milledge et al., present a thorough statistical analysis of six coseismic landslides inventories to relate landslide hazard to landscape properties such as slope and contributing areas, but also more specific variables such as skyline angle and a hazard area integrating the probability of initiation and propagation of landslides. They found that the two latter metrics explain best the location of the inventories and may allow to be converted into simple rules useful for hazard management. The paper is well written, with a straight forward structure and informative. It will make a nice contribution for NHESS both for its systematic analysis and its recommendations. I have two major comments that I think could improve the results and the discussion, and then give a number of minor Line by Line comments with potential clarification or additional small analysis.

Major Comments

My first comment is about the normalization of several of the hazard metrics : I am convinced that a substantial part of the difference between the hazard curves could be removed by plotting the hazard against a landscape metric : For example for slope, each landscape as likely a modal slope, that may be interpreted as the result of geolechanical difference (for steady state landscape at least). Thus curves may be plotted against S-mode(S) , somewhat normalizing for difference between two landscape. I can understand the author may still want to express their rules in terms of absolute values of slope or other variables, but I suspect this normalization would clarify and strengthen the result and their analysis (as this did in other studies). I make suggestion for the other variables in my inline comments.

This is an excellent suggestion, and indeed we found that normalization collapses the hazard curves to some extent. This is very satisfying in terms of explaining our observations. We include these new results in our revised manuscript though they do not alter our conclusions since normalization does not alter the rank order or improve predictive skill of any metric.

My second concern is that their maybe some over-interpretation of the data scatter towards the extremity of the hazard curves. And the author do not provide clear metrics or indication of the validity of individual datapoint. This is not an easy task but the work of Rault et al., which I co-authored, recently proposed a method to do exactly that. I would suggest the author to apply these criterium and check.
In this work we consider: the probabilty p of the whole topography, and the one resulting from the landslides affected area only p_L.
To assess whether p_L is significantly different from p we compute the confidence interval Ip associated to the random drawing of n (n the number of landslides) pixels out of the landscape distribution. If p_L belongs to [p-Ip : p+Ip] then we cannot exclude that the difference between p and p_L just comes from random fluctuations and it is likely not significant. Given landslides remain rare in the whole topography, the drawing can be assumed independent, and similar to a Bernoulli sampling. Provided the central limit theorem is respected (i.e. n>30, np>5 and n(1-p)>5) the 90% confidence interval can be estimated as: Ip = p – 1.96 (p(1-p)/n)^0.5 ; p + 1.96 (p(1-p)/n)^0.5. Some additional details can be found in the supplementary methods of Rault et al., 2018. Basically n is large (n>1000-10,000) so the authors should obtain very narrow Ip until they reach p<0.001 – 0.0001 but I expect these low probability to be reached in the tail of the distribution (Fig 3,4, 6) and the cut off will vary for the different landscape with higher or lower p or n. The authors could compute Ip as well as the convergence criterium and show the points which may be insignificant in shaded / transparent ?

Thank you for pointing us to this approach. We had struggled to find a way to account for sample sizes in our analysis but the Rault et al. approach is extremely well suited to the problem! We have now implemented this method in all cases where we generate hazard curves (i.e. conditional probability curves). In each case (Figs 3-5) we show both those data points that show a significant difference and those that don't, and we explain this distinction in the text.

Line By line comments:

L123 : I could not find Milledge 2018 in the reference list... please check. **Added.**

L133: Add couple of reference for shaking: e.g. Khazai and Sitar 2004, Meunier 2007. **Added on L141**

L138: I think you should also cite Meunier 2008 here, and probably the recent analysis discussion for an extended number of earthquakes in Rault et al., 2018. **Added on L145.**

L142:152 : A couple of references on the suspected effects would be relevant. Especially the ones cited elsewhere in the text: Parise and Jibson 2000, for lithology, Maufroy et al., 2015 for curvature and ridge amplification. **Added on L154-159.**

L155: True they pertain to initiation, but vast majority of studies highlighting their role or quantifying statistical relations between these predictors and landslide use total area and therefore are combining both initiation and runout.
We agree. However, the mechanistic justification for the factors is almost always initiation based, as are the GIS approaches that are typically applied to assess landslide susceptibility. Our point here is that when these variables are used for landslide hazard prediction they are used to represent controls on landslide initiation. We have modified the sentence to say:
"The potential predictors described above are primarily chosen in hazard models for their perceived link to the probability of coseismic landslide initiation." (L163).

L180: I have the impression it should be the minimum skyline angle, not intersecting topography, …. Indeed a maximum reach angle. Cf comment on Fig 1
We could phrase this either as: 'the maximum angle from horizontal to the skyline' or 'the minimum angle from the horizontal that does not intersect the skyline'. We have chosen the former because it is shorter and because we are concerned that the latter is more open to misinterpretation. In particular, people may not think of cones at increasing angles and thus may misunderstand or ignore the second clause.

L200 – 300: This is certainly at the appreciation of the authors, but I have the impression the earthquake environment (tectonic, climatic, vegetation) is over described. Given you never re-refer to this context later, you may shrink those description and end this section with a sentence like : "these epicentral areas encompasses a large diversity of tectonic (X to Z) , climatic (X to Y) and vegetation cover (X to Y) contexts, but we assume landslides in all of them should be at first order driven by topographic parameters in the same way".
Thank you, this is a useful suggestion. We have considerably shortened this section, compressed the information into a table in the Supplementary Information, and added a summary paragraph in line with your suggestion.

In contrast, some aspects may be missing or insufficiently discussed:
1/ I think the number of landslide polygons used in Chi-Chi is missing. **Agreed, added on L229**
2/ The fact you used a sub-inventories in Wenchuan may mean you artificially limit your analysis to a range of shaking quite different from the other cases . This should be mentionned.
We agree that this is an important issue and needs clarifying, as also mentioned in our response to Dr Scaringi's comments above. We have added maps of the study areas in the Supplementary Information to help readers interpret our results. In the Wenchuan case we show both the full inventory of Li et al. (2014) and the subset that we use. We have added a statement in the paper itself (L239-242) to show that the range of PGA values experienced in our Wenchuan study area (0.16-1.3 g) is similar to those for the whole inventory (0.12-1.3 g).

3/ A few words on the Implications of polygon mapping quality on your analysis may be given here (or in discussion), such as the affect of amalgamation ; inclusion or not of debris flow propagation within the river network ? (I know it is a difficult distinction, often ignored but it should impact the statistics, especially of contributing area for example). Implictions of the different resolution limit (for Chi Chi or FInisterre compared to Northridge) or of the location accuracy ?
Agreed, we have now added a paragraph on mapping quality in the discussion (L840-853)

Figure 1: Caption: a cone projected from P no ? **Agreed, modified.**
If your cone as angle define from horizontal upward, are you looking for the minimum skyline angle, not intersecting the topography? That is what I get from your sketch in c). See previous comment about skyline angle definition.
Both are possible definitions of the angle we are seeking, but we believe our current definition is less open to misinterpretation (see response to comment on line 180 above).

L372-380: < 10 observations ? Why ? And this is only for the 2 ariable case (Fig 4). For the single variable case it is not clear what is noisy data and where to really set the boundary or unsignificant datapoints. Rault et al., 2018 propose an extension of Meunier et al., 2008 studies with an estimation of the uncertainty of observations based on both the number of observations and the probability. See major comments.
We have now applied the Rault et al approach, so thank you for pointing it out to us.

L458: Shalrun-EQ= Probability of mobilization convolved with connection probability. Average in the above area. So hazard area is basically the number of pixel where debris flow can occur and reach the interest cell (say Nhaz)... in the contributing area, times pixel area, and divided by the contour length, i.e. the square root of contributing area. Although I am confused because in Fig 2 : Hazard Area seem to be Nhaz times Pixel resolution (or Nhaz.a/sqrt(a)). But then smallest vales should be 0 and 30 (as it does in Fig 2). But in Fig 6 it goes from 0.1 to 1e3… So there seems to be a problem between the 2 definitions. Please check.
Your interpretation of SHALRUN-EQ is correct. The reason for sub-pixel hazard areas is due to multiple flow path routing. We now explain this more clearly (see L432).

L462: repeat from L452-453. Cut or rephrase ? **Rephrased to avoid repetition.**

L478-482 : Steepest descent may be too conservative, even if your rule needs to be simple, maybe you could mention that probability to propagate on non-steepest descent path is probably non-null.
This is an error in our explanation. We in fact use multiple flowpath routing, and amended our description of this in the methods section (see L435).
Also what about landslide large enough to be continuing beyond the first cell with angle below the deposition threshold? I see this is partially acknowledged in the discussion. Maybe you can flag here the fact you discuss such limits later. **We Agree and now point the reader to our discussion of this in section 7.1. (see L450).**

L512-516 : Ok, simplicity is important and it is difficult to integrate other effect mentioned here. But what about checking the actual evolution of probability with slope, for both initiation and stop? A reasonable estimate of scar area can be obtained by selecting the highest elevation pixel in your landslide, and selecting as many as needed to reach a scar area with an aspect-ratio of 1.5 (Domej et al., 2017) and a mean width representing of your polygon (see Marc et al., 2018 for how to do that). Doing so you could check if a plateau develop in your probability ratio after 39Æ or so... Interestingly you could reverse the idea and take the lowest N pixel (N ~ Width / 30 for 30m resolution DEM) of your landslides to obtain a probability of stopping.
We had performed this analysis with interesting results. We found that landslides initiate on a fairly narrow range of slopes but stop on a much broader range (consistent with our observations), with modal values similar to our optimized values. However, these results are telling us the slopes on which landslides initiate and stop rather than the probability of initiation and stopping given slope, so we need to be careful in connecting the two sets of results. A careful examination of this connection is outside the scope of the paper since we focus on developing simple rules.

L536 : Did you check the curve appearance when using gradients, that is tan(Theta) (with theta the slope in degree) ? Because the tan(theta) does appear mostly exponential over a large range of Theta. Thus a linear function of tan(theta) (the relevant parameter for landslide stability) may appear exponential when plotted against theta.
The kink in some curves at low slopes indeed suggested that tan(theta) might be a better predictor. While this is consistent with landslide mechanics, we found that for most inventories this relationship does not provide a good fit to the data at high slopes. Given the additional complexity of the tangent function it is not well suited to a simple rule so we chose not to report it here. However, the misfit between tan(theta) and landslide probability is clearly interesting and merits further examination in the future.

L538-542: Northridge and Haiti are shifted compared to other. They both become > average probability around 20Æ , vs 30 for others. This roughly correspond to modal slopes of these areas. It would be interesting to re-plot all curves not against slope, but slope – Sm the modal slopes. This collapsed curves on a similar analysis for rainfall (cf Marc et al 2018) Similarly is there a large variety of drainage area distribution ? Haiti and Northridge are very peculiar again compared to the other cases. Some normalization by the mode of the landscape drainage area may be important.

This is a good suggestion in terms of improving the explanation of the dataset as a whole but does not alter the simple rules because: 1) they are applied in relative terms (i.e. choose the location with the lower local slope); and 2) the alteration would require knowledge of the slope distribution for a location (which will not be available to most users). Nevertheless, we have now performed normalization on both slope and UCA. We include the figures in the supplementary information and briefly report the findings in the main text (L529).

L542-543: If you consider that Haiti and Northridge are more sensitive because they reach higher ratio it may be a confusion because of the lack of normalization (previous comment). It is plausible the relation between slope – Sm and hazard is similar, only the difference between resolvable slope (with a 30m DEM) and the modal slope is larger, allowing to reach larger relative hazard. I think the effect of normalizing for the landscape must be assessed.
This is exactly what the normalization shows. We have adjusted the text to reflect this observation, added a normalized panel to the figures, and refer to the normalized results in the text (L505).

L545: combined or merged PDF rather than amalgamated (that sounds negative an unusual to me but I may be wrong). **Altered to combined (L506).**

L555: You say you observe contributing area, but you have normalized by contour length. In the paragraph about hazard (L489), you say contour length is a^0.5, but it is not so clear what is a (the area of a cell, which cell ?) On Fig 2, contributing area seems to be the square root of a. It would be consistent with the contour length estimated as sqrt(a) but then why not say straight you look at the sqrt of drainage area ? Maybe I missed something, or it is worth clarifying a bit.
We now clarify this on first introducing upslope contributing area "and normalising by the grid cell width to minimise grid resolution biases" (L371), and in our definition of l_j, as the cell width (L461).

L562: This was somehow my expectation, so why not normalizing the contributing area and thus analyzing a/a_rc , with a_rc the channel ridge transition area? Like this the relative decrease or increase away from this objective characterization of the landscape could be analyzed (and the plot in Fig 3,4 would compare hazard curve shape only, not locations). This seems like an important improvement even if I understand that you may point to the fact a layman user of an hazard rule may not guess the modal slope of its landscape or the value of a_rc. After some analysis in the normalized domain general rules for the natural domain may be derived.
We agree, and now include a normalized plot showing the Northridge curve partially collapsing onto the other curves. Given the generally poor performance of upslope contributing area we choose not to come up with a new rule based around it, nor adjust the other rules in light of these results.

Fig 4 is very interesting and make a lot of sense after Fig 3. However, I am wondering about two things…
1/ Would all the plot look the same if you use normalized area and slope ? Maybe not given that it was not expected from Fig 3 that Finisterre would be different, but it seems worth and easy to check.
The differences that result from normalization are largely in the steepness of the surface rather than the way that slope and upslope contributing area interact. As a result there is little obvious change in Fig 4 as a result of normalization. However, we show the normalized results in the supplementary information for completeness.

2/ You work with 100 log-bins of a and it seems 1degree bins of slope. So I wonder what is your typical number of DEM cells in each of your bins, and thus how statistically significant bins are… This is a detail as pattern are very consistent and a larger bin size would rapidly increase the amount of data.
We have applied the approach of Rault et al., extended to 2D, to indicate bins where the hazard is significantly different from the study area average. We flag that in the figure caption.

Fig 5 : Skyline angle is strongly uni-modal. So I would study all areas with a relative skyline hazard: Sky-Modal(Sky). The modal will account for difference in incision/relief between landscape. A potential outcome of such normalization may be that your case have all similar behavior for high skyline angle (increase and then plateau) but that Gorkha, Haiti, Northridge have a steep decrease below a certain angle while not the three others.
We tested the effects of normalization and as you suggest it does collapse the data to some extent. We find that normalization is particularly effective at aligning the Gorkha, Haiti and Northridge hazard curves with those from the other sites. We now describe this normalisation in the text.

The definition you take for hazard area gives 0 hazard area for the reference in all cases and then a decrease. It does not seem that shift in the horizontal direction would do any good, and the vertical shift seems due to the proportion of zero hazard area in the landscape, so maybe computing a landscape PDF ignoring the zero would be insightful?
We suspect that normalisation may not be particularly informative in this case. We could normalise the initiation and stopping angles but we are then farther from a simple rule and, particularly in the case of stopping angle, it is not entirely clear what the appropriate property to normalize by would be. As a result we do not pursue normalization for hazard area.

L645: Ah<1 mÇ/m . I am surprised by this threshold, but maybe it is a typo. I would have say in Fig 5b the curves steepens most in all case around 20. It is true that for Haiti, Gorkha and Northridge there is a slight increase in the trend after Ah~1, but minor compare to the later steepening.
This was a typo, and has now been fixed.

Also to be sure that the difference between a peak or a plateau is a real result it would be important to check the evolution of the uncertainty in your last bins, where certainly few data are available (even if we cannot read the probability of Ah>1e2 or 1e3).
A peak followed by a decline in hazard with increasing hazard area is retained within that part of the data where there are sufficient observations to allow confident hazard identification only for Haiti. However, we have adjusted our plots to indicate which observations are more or less certain, using the approach of Rault et al. as described above, and discuss this in the modified text at L630.

We also do not see the difference in availability of such high hazard area in the different areas, so could a very low availability of such hazard areas in Haiti and Northridge (that have less steep slopes) caused a scattered behavior for Ah>100 instead of Ah>1000. A quantification of uncertainty may clarify that. See major comment. **Your suggestion that the earlier onset of scatter in Northridge and Haiti hazard curves is likely to reflect a lower availability of such steep slopes is supported by the Rault et al. analysis, which clearly identifies the point beyond which the curves become more scattered as the point beyond which hazard cannot be confidently resolved. We make this point in the main text at line 636.**

L672 : This sentence confused me. Do you mean each of the three parameters, may be better than the skyline angle for at least one event ? **Yes, your interpretation is correct. We think the confusion was due to a punctuation error (full stop should have been comma), and we have now fixed this.**

L674: These values do not match Table 1 with 0.72, 0.69, 0.74.... Please correct one or the other. **Thanks for spotting this typo, we have now fixed it.**

L694: I am a bit surprise by the term of channel inside this rule. I guess it derives from the fact that the hazard consider upslope contributing areas defined from flow algorithm. But the hazard area at many intermediate locations on hillslopes may be a channel for your analysis but not for the resident and deciders of the area. Because a channel is defined on finer scale than the DEM. You already say that this metrics is anyway difficult to estimate and handle for application, but this terminology would also complexify the problem for deciders or policy makers.

We can understand your concern here and have considered alternatives. However, we have chosen to retain the word channel within the rule for two reasons. First, because we feel that it is important to capture the notion of convergence and we are unable to find an alternative wording that can do so. Second, because we expect that if SHALRUN-EQ is calculating convergence using a 30 m DEM, it is extremely unlikely that the real topography does not have some sort of channel or gully within that area. It’s hard to imagine a topography that would be convergent at 30 m scale but not obviously channelised or gullied at finer scales.

L699 : This is fortunate indeed, almost surprising.
Agreed - we expected more sensitivity to the parameters here.

L711: Interesting. Do you think this could be somewhat validated by making skyline and hazard graph for landslide above and below a certain threshold (say 5e3 m2 or even better above a certain width...) ?

This is an interesting idea and something that we will investigate in future but we feel that it is outside the scope of the current paper.

L739 : You certainly mean Meunier 2008 here. However, note that the new study from Rault et al., 2018 is considerably nuancing these past studies.
Modified to add citation and account for Rault et al's work (L727).

L820: And even for a trained observer.
Agreed but given our simple rules focus we choose to retain the focus on untrained observers here.

L822-23: I do not understand what you mean by "we expect the length scales over which this occurs to be long (order kilometres) relative to the other factors examined here" Do you mean that main lithological units are usually big (regional scales) and thus significant part of a landscape will have homogeneous lithology, whereas topographic attribute change at the scales of 10s of meter ? Then it is the length scale for the variability of lithology that you want to mention. Anyway please clarify.
Your interpretation is correct but we have clarified our point in line with your suggestion (see L812 in revised manuscript).

On a side comment, normalizing each landscape slope by their modal slope would be somehow a step toward normalizing difference in landscape that can be due to major lithological or geomechanical attributes (Korup 2008).
Agreed, but as discussed above including this in a simple rule would be problematic.

L824-826: This is an important and natural point to make but I would mention rainfall induced landslides straight here, as area affected by coseismic landslides are often even more often affected by rainfall induced landslides (at least for wet climate Nepal,Finisterre, Taiwan).
Agreed, and we have modified this text to: 'such as flooding or even rainfall induced landsliding'. (L815).

L830: And they likely do, given that large landslide (likely to travel further away as you recall in the introduction) are usually reported closer of the fault or at larger shaking values (Khazai and Sitar (2004), for the Chi-Chi earthquake (1999), Massey et al. (2018) for Kaikoura or Valagussa et al 2019 for systematic evaluation of PGA and landslide size distribution. So future exploration of the behavior of your hazard curve split for specific lithology of different area class should be done.
Agreed, both landslide size and lithology are interesting topics for future work but are outside the scope of this paper.

L834 : I would say we can reasonably expect strong differences : given that hazard increase strongly with local slope for EQ (Fig 4) but not for the rainfall induced landslides : as shown by the anaysis similar to your Fig 3 in Marc et al., 2018. Further, the longer runout (due to lower stopping angles) and stronger dependence on contributing areas are additional changes.
We agree that large differences are possible, but we think it is fair to say that the strength of these differences is not yet clear.