Comment on nhess-2021-283
Anonymous Referee #2

The authors propose a method and some best practices / strategies for landslide detection and/or landslide density headmaps estimation using multitemporal SAR images pre-processed in google Earth Engine. The method combines quite traditional approaches including change detection thresholding and geomorphological filters. The landslide density headmap based workflow is tested for five different events.

Despite quite clear writing, the manuscript is not, in my opinion, very readable. This is probably a consequence of a research framework with some ambiguities and inconsistencies, and the use of a bit too rough of some concepts. This makes it difficult to say whether the conclusions are really supported by evidence.

In the end, I can't find in this work any substantial scientific significance and it seems to me too hasty to assume that it contributes to defining new strategies or best practices for landslide detection.

My main concerns are referred to:

1) an unclear (probably inconsistent) use of the Hiroshima test to define best practices based on algorithms and products (ROIs, and detection) that are not used in the following test cases. There is no evidence that those best practices are methods invariant.

2) a lack of quantitative validation when the threshold is chosen manually (nothing against expert-driven methods, but they have to be supported by evidence and validation. Any consideration on FN is missing.

3) some decisions taken on an apparently superficial knowledge of the data and algorithms used (see some comments later)

the three points are addressed in my following comments to the manuscript.

I'm sorry to suggest rejecting the paper but I encourage the authors to go through a more correct definition of the purposes and of the research framework.
Introduction

89 experienced landslide activity: correct, but I’d like to suggest not to use activity here because it can be confused with the activity as defined in https://doi.org/10.1007/s10346-012-0335-7 (same at line 95)

2 Methods

2.1 SAR backscatter in Google Earth Engine

126 For this study we only used SAR data in the VH polarization: in a number of papers cited in the manuscript also VV is used with a few advantages compared to the depolarised signal in particular when roughness does not matter much. I think the choice of providing best practices just using a single channel should be better introduced.

2.2 Landslide Detection Approach

132 The AOI can be a single landslide or a mountain range: it sounds a bit weird because it seems to assume that the landslide is already known. Also, the density map seems to me to lose/change meaning

135 full event inventories: which does not seem the real purpose of the procedure (density)

137 - 142 We also ... stacks: this part needs some more clarifications, in fact, stacking per se does not reduce noise. I guess the authors assume that the mean of the images should be less noisy and more representative of the ideal surface backscatter of the different land covers in the area. Since here we are dealing with random processes, this should be true/acceptable when the series is stationary. Is this true in all the test areas? I suggest to explicit when this is not applicable. furthermore, the atmospheric delay is not canceled by averaging the pixel values but it is smashed in the series.

155 -159 Since ... deposits. It is very/too generic. I suggest linking this to the very wide literature on the topic (i.e. susceptibility)

2.3 Change detection performance and determination of most effective detection strategies

164 – 165 quantitatively evaluated our results with a previously published landslide inventory for the 2018 Hiroshima landslide event using Receiver Operating Characteristic curves (ROC) (Fan et al., 2006): It seems to me that this is not evaluation but calibration: usually, thresholds are found in a part of the study area, and then applied to the remaining part of the area for independent evaluation. Here a step is missing.

171 – 174 We quantified ... pixels: see my previous comment: also this sentence can sound ambiguous. ROCs are used here to find the best threshold which makes the benchmark and the new product as ‘closest as possible.

178 10 m x 10 m: this is the pixel size and not the spatial resolution (should be something like 20 m but it really depends on the filtering), so is 100m2 enough? As far as I can understand, images were not filtered, so, dealing with single/very few pixels is quite risky (speckling, multiplicative noise, and so on...)
204 – 205 We manually explored I TR,H using I ratio percentiles to find the threshold value that visually highlights true landslides and reduces noise: in my opinion, a qualitative analysis here is not enough, furthermore a single test case is too dependent on the event. More events should be compared a priori (before the density map).

208 – 209 must be defined without the use of an external landslide inventory: then with what? By using an optical image? But this is against the starting presuppositions. Furthermore, this invalidates the fact that the best practices found with ROC can be directly used here because the approach used in an intermediate step is different...

3 Test sites

3.1 2018 Rainfall-triggered Landslides, Hiroshima, Japan

3.2 2018 Earthquake-triggered Landslides, Hokkaido, Japan

3.3 2020 Rainfall-triggered landslides, Huong Phung and Quang Nam, Vietnam

3.4 2021 Earthquake-triggered landslides in Haiti

4 Results

4.1 Determining Effective Strategies for Detecting Landslide

The use of the Hiroshima event and ROIs to define the most effective strategies is really confusing me, in fact, the results are dependent on the approach used and the final product which here are ROIs and pixels potentially inside landslides, while in the main part of the paper other strategies to find thresholds are used, and the final product itself is different: density. How can you make sure that what is eventually true for one method and one type of product are still true when other methods are applied, and furthermore to obtain different products?

311 -312 (i.e., all available pre-event and post-event data) and found the slope and curvature thresholds that maximized the AUC: since this is event dependent, how much this result can be generalised?

326 – 328 lastly ... values: as far as I know, the ratio (of two gammas) should be beta-1 and the log of beta-1 is not in the list of the known pdf. I suggest not to include this analysis or to delve into more.

4.2 Determining Effective Strategies for Rapid Response

338 making comparison: it is not a comparison but a tuning

344 -345 low signal to noise ratio of the post-event images: due to problems in the acquisition? Processing?

4.3 Landslide Detection with SAR-based Change Landslide Density Heatmaps

352 - 353 the true landslides and removes most of the false positives: this is to me a weak point of the work, a quantitative analysis should be carried out, showing for example also false negatives.

5 Discussion

5.1 Landslide Detection using SAR Backscatter Change
This decrease in SAR backscatter intensity occurs because the landslide scar and damage act to decrease backscattering reflectance to the satellite relative to a pre-failure ground surface: the sentence is messed up, just say that landslide cause a decrease (in this particular case) of the radar backscatter.

We found ... layover): these are all quite standard results: the use of multitemporal, geomorphological filters, multi geometry solution is quite present in the current literature, so, I suggest to say that the study confirms that...

The combined effect of stacking hundreds of images with both geometries is an improvement from previous SAR backscatter intensity change studies that have focused on individual acquisition geometries and a relatively small number of SAR images: this is eventually true when the method proposed here is applied (and actually one test case is not significant from a statistical point of view), there is no evidence that other methods proposed in literature would have worked better with long temporal series. furthermore, as far as I can understand, in most of the scenarios only 2 weeks acquisitions, pre-and post images were used.

pixel resolution: I guess spatial resolution. I suggest taking into account that in GRD products multilooking is present...

This resolution limits our ability to detect small landslides with lengths or widths < 20 m and as a result larger landslides are more likely to be detected: consider that the use of pixel-based applications is not recommended in particular when filtering is not applied, so the assumption of selecting such a small limit for landslide detection is very risky.

5.3 Challenges with Rapid Response Landslide Detection

5.3.1 Identifying the AOI and EOI

This paragraph seems to at least partially contradict the premises of the strategies: how does, in the end, the method help?

5.3.2 Challenges using the Landslide Heatmap

In this paragraph and in general, in the paper, the importance of false negatives is really underestimated.

5.3.3 Computation Limits in GEE

5.4 Satellite Acquisition Frequency and Landslide Detection

Our findings suggest that if the satellite revisit was twice the current revisit time, the modeled AUC score would be ~0.7 just 1 week after the EOI, while if the revisit time was half the current revisit the modeled AUC score would be ~0.65 (red lines in Fig. 11b): quite arguable: the processes intercepted by changing the frequency sampling and temporal windows can be different so as the final results.

5.5 Future Work

6 Conclusions