Natural Hazards and Earth System Sciences Discussions

Nat. Hazards Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/nhess-2021-148-RC1, 2021

Comment on nhess-2021-148

Anonymous Referee #1

Referee comment on "Integrating empirical models and satellite radar can improve landslide detection for emergency response" by Katy Burrows et al., Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2021-148-RC1, 2021

The Authors investigate the advantages of adding coherence-based landslide features detection in earthquake-induced landslide empirical modelling.

The paper is well written but difficult to review because this research makes use of multiple expertise including machine learning, remote sensing, and geomorphology. As an average reader, I acknowledge that some of my notes/questions are more like clarification requests than criticisms.

The topics (there are many) are interesting and quite current, in particular the use of SAR for rapid detection/mapping, and earthquake-induced landslide susceptibility modelling (probably the use of the term susceptibility here is 'controversial' because these models exploit 'dynamic variables'. I'm not particularly expert in this so I will not delve into). Some of the results are interesting and encouraging.

My concerns include:

Data re-sampling consequences seem to be underestimated or not fully considered: re-sampling is here used to overlap layers with different resolutions but, when data are re-sampled, at least two points/questions must be taken into account: (I) is my process still represented/sampled at the correct scale? (II) Does re-sampling introduce artifacts (e.g. aliasing) that in quantitative analysis can impact the results?

Apart from the re-sampling, the original scales/resolutions of the data are not reported but they look to be quite different. Are they all adequate to sample the different processes and to be combined in a machine learning model (no ghost relationships)?

The results are (quite well) numerically discussed but the geomorphological part is missing. So we know where and when there are numerical advantages in the combination of the two products, but we don’t know why (e.g. geomorphological characterization of the true/false positives/negatives), this makes it difficult to evaluate the real possibility to export the framework in different contexts (e.g. Par 4.4)

There is a point that I admit, I still have problems to unravel. I feel that the input variable ICF is somehow a proxy of what the models want to find (the output variable). Furthermore, it sounds to me as if the ground truth was used twice: the first time when landslides (Y - independent variable) are exploited to label the units, and then, when the ICFs say (again, but as feature variable - X) to the model where landslides occurred. Can a regression/classification model make use of variables that somehow are what the model wants to predict/classify (no spurious correlations?)? If yes, can their importance be evaluated in the same way the importance of the other variables is evaluated? My doubt can be originated by the difficulties I'm having to interpret the output of the model as landslide mapping instead of landslide (spatial) forecasting (see my several comments later.

It would also be helpful to evaluate how ICFs false positive and negative (random, systematic, reproducible?) impact the model.

More details can be found later. I hope this helps.

Title: not so sure the title is correct: the empirical models don’t detect landslides but they model the spatial probability of occurrence.

1 Introduction

20-25: “generating information on the intensity and spatial

the extent of landslides in the immediate aftermath of a large earthquake”: what does intensity mean here?

25-44: The discussion about whether to consider driving factors or not into susceptibility

maps is quite hot and open. I suggest anticipating here the comment at par 2.3.2 to avoid possible misunderstandings.

45: maybe helpful: https://doi.org/10.1016/j.earscirev.2021.103574

60-65 ‘Here we aim to establish which of these three options is best’: susceptibility and detection provide different types of results (LAD probability, and LAD measure), and what is best is not absolute but it depends on how they are used. I suggest explaining what ‘best’ means here.

70-75 “We chose to model LAD rather than individual landslide locations as both empirical models and SAR-based methods perform best at relatively coarse spatial resolutions”: It does not sound very clear. I suggest rephrasing the sentence and try to explicit more the relationship between spatial resolution and the models (I guess susceptibility and LAD from coherence and not random forest). Another possibility can be to demand this concept to the discussion.

2 Data and Methods

2.1 Landslide data sets

2.2 Training and test data sets

109-115: 2 questions:

1: ‘a small number of landslides’, small, but I guess the number must be representative from a statistical point of view and a geomorphological point of view, in the sense that it should be statistically significant and well representing the density of landslides in the area, correct?

2: ‘we randomly selected 250 landslides...’: if I understand well, you labeled as yes, those pixels in which LAD > 0.01, and no elsewhere, correct?

115-120 ‘we randomly undersampled’: I guess you mean that you randomly selected the pixels, correct?

115-120 ‘resulting model was therefore trained on equal numbers of pixels from the two training events’: interesting point. The choice is probably dictated by the model, I’m wondering what implications this choice can bring. Don’t you ‘de-correlate’ the strength of the earthquake (represented somehow by the ground shaking data) and the number of landslides (the consequence of the strength) (and other covariates)? Forgive me if I’m using some terms inappropriately.

120-125 ‘each model within the ensemble is trained on a different set of cells’: do you estimate/suppose that the inventories have locally the same quality?

120-125 ‘This process reduces variability’: in what sense?

125-150 ‘Masato...’: I suggest to better contextualise: the model tuned using rainfall does not use (a polite guess…) shaking data...

130-135 ‘Instead, a high performance on at least one predicted event would be considered a success’: I would be a bit more cautious, one result is not statistically significant, I suggest to cancel the sentence and say (correctly) that would encourage further investigation.

2.3 Input Features

Two general comments related to the fact that this is a quantitative (and not qualitative or based on interpretation) analysis:

1) resampling to a higher resolution can be problematic, what type of resampling did you use?

2) how did you aggregate here? In this case, have you considered aliasing problems? Is the result still sampling the variable at the right scale?

150-155 ‘a static proxy for soil moisture’: did you find it relevant (from a geomorphological point of view) for earthquake-induced landslides? Or just from a numerical point of view?

2.3.2 Ground shaking estimates

General comment: I suggest adding the scale/resolution of the map, and eventually its uncertainty.

Since this layer is also quite different from the others (sort of causing factor) I also suggest commenting (maybe not here) how this can be an adequate sampling of the measure in relation to the product you want to obtain and its resolution. Is the map enough resolute to influence/characterise the LAD? Is it compatible with the other products?

160-165 'is what distinguishes': so, is there any way to call it differently?

170-176 'Our inital susceptibility model may

therefore perform better than one that uses the data made available within the first few hours of the earthquake...study': this should be part of the discussion. In any case, I still strongly recommend testing and compare the two products, since you are evaluating a progressive improvement of the performances of the system devoted to working in an emergency phase, and SAR images might be available just after the event. (I guess inital is a typo).

2.3.2 Lithology

General comment: I suggest adding the scale/resolution of the map

180-185 ‘One advantage of Random Forests is their ability...’: I suggest moving this elsewhere

2.3.4 Land cover

2.3.5 USGS ground failure products

General comment: I suggest removing/move the first part of the paragraph. Here only the method should be described and not the reasons for which others failed. In the discussion, you can say why you did not use the same model…

And again, I suggest finding another way to say that the model would have probably worked worse if you had used another product… this sounds like ‘incomplete’.

2.3.6 InSAR coherence features (ICFs)

230-235 ‘This volume of SAR data..’ for sure for S1, also true for ALOS-2?

240 – 245 ‘Burrows et al (2020) … and 2019, and the whole paragraph: I had a look at the papers, unfortunately, I’m not sure I got correctly a point: the different methods are based on differences (generally speaking) and to decide what is the right threshold, a ROC analysis is required. A benchmark is then required, so you need to have the landslide map already prepared. I see a kind of loop. Can you please clarify? Or, If I’m wrong, can you please make explicit when ‘lower’ or higher’ is low enough or high enough to say that the changes were caused by landslides?

255 – 260 ‘showed some level of landslide predictive skill’: not sure I would use predictive here, I suggest detection capacity or similar

2.4 Random forest theory and implementation

General comment: not sure that Fig 2 and some parts of the description of RF are really helpful. My suggestion is to remove fig 2, shorten the paragraph, and eventually add some references.

2.4.1 Feature importances

2.5 Performance metrics

325 – 330 ‘by comparing the test areas of these predicted surfaces with the mapped LAD calculated in Section 2.1.’: LAD obtained from the external inventories, correct?

330 – 335: I understand the need of choosing a threshold, in fact, I think the most appropriate sentence to define what is tested here is “The ROC AUC values calculated here, therefore, represent the ability of the model to identify pixels with LAD > 0.1” but, according to the fact that the percentages are so different for Hokkaido and for Lombok, I’m not so sure that the value can have the same weight in the testing phase (different densities related to different events/geosettings).

340 – 345 ‘The second method...’: this sounds to me like an indirect evaluation because in between the real numbers and the results of the RF there is a further model to obtain the interpolation. If correct, fine for me but I suggest to better motivate the choice.

3 results

3.1 Same-event models

3.2 Global models

General comment: I would have preferred to see the original ROCs and not only the differences.

3.3 Do these models outperform individual InSAR coherence methods?

General comment: I'm not so sure that I got the point. Susceptibility and mapping are two different things and when you compare using the LAD benchmark you are comparing different results. In the first case the capacity of predicting spatial landslide occurrence (if this is a real susceptibility … what is more, including a sort of ground truth obtained from the coherence-based LAD), in the second you measure the capacity of mapping using a technique.

3.4 Feature importances

410 ‘where these become the most important feature in the model’: Here there are at least 3 different ‘categories’ of features: (I) those features that take a static picture of the pre-event situation (lithology, slope…) (II) the velocity, which is strongly contingent upon the event, and (III) the Insar features, that map the event. An RF feature importance evaluation is only numerical, I suggest providing a geomorphological interpretation, and considering whether the 3 classes can be really evaluated together using the same criterion.

4 Discussion

General comment: I in principle agree with all the topics chosen for the discussion, but I think that a real discussion on the geomorphological interpretation of the results is missing.

4.1 Model interpretation based on ROC AUC and r2

420 – 425 ‘how the current generation of empirical models at this spatial scale should be

interpreted.?: any hint, personal opinion, on this?

4.2 Selection of the training data for the same-event model type

4.3 training data format

4.4 towards a global landslide prediction model

4.5 Current recommendations for best practice

General comment on L-C bands: probably one of the things in common among the 3 events is the presence of vegetation in the affected areas before the event. Would your suggestion change if lc was different? See recent earthquakes in Iran that triggered landslides.

525 – 535: I reiterate a previous comment: can the two products be directly compared? One is mapping, the other is 'probabilistic spatial forecasting'. I agree with the suggestion about having a single product to update.

4.6 Future possible SAR inputs

545 ‘is thus not usually reliable in landslide detection...’ I suggest relax this sentence, actually, it is, it just needs more controls that are not highlighted in the cited studies.

560 – 565: See my previous comment on C-L bands, I suggest mentioning these parts earlier in the recommendations for best practices.

4.7 Possible applications to rainfall-triggered landslides

General comment, which can be somehow applied to the next paragraph as well: it seems to me too generic, and I’m not so sure that this topic deserves a dedicated paragraph. I suggest shortening this part.

4.8 Possible application in arid environments

see my previous comment