Nat. Hazards Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/nhess-2020-200-AC1, 2020

Natural Hazards and Earth System Sciences
Open Access
Discussions

# *Interactive comment on* “Near Real-Time Automated Classification of Seismic Signals of Slope Failures with Continuous Random Forests” *by* Michaela Wenner et al.

**Michaela Wenner et al.**

wenner@vaw.baug.ethz.ch

Received and published: 11 September 2020

Dear editor and reviewers,

Thank you for thorough reading and revisions of our manuscript "Near Real-Time Automated Classification of Seismic Signals of Slope Failures with Continuous Random Forests". Enclosed you will find a response to all reviewer comments on the manuscript. The most important changes to the manuscript are the following. i) We will add additional data from Illgraben, Switzerland and test the proposed method on both the Illgraben data set as well as the Bondo data set. ii) The final classifier will be trained on both data sets, which tends to improve the classifier. iii) More information on

the random forest parameters will be given. iv) The results and discussion section will be reorganized according to the reviewer comments. v) The discussion will be more thorough.

On the following pages, we provide a detailed point-by-point response to the reviewer’s comments. Our replies are in blue. Most minor comments which are straightforward to implement (such as typos and rephrasing of sentences) are simply ticked off (using the ✓sign) without providing a response.

If you have any questions, we would be happy to answer them. We are looking forward to hearing from you about your decision.

Best regards,

Michaela Wenner

Comments of reviewer 1

**Comments to the Author**
**General questions:**

I read a study that has explored the potential of a machine learning algorithm to jointly detect and classify mass wasting and earthquake events from a small linear geophone array along a channel in the Swiss Alps. The study opens a new and timely avenue of "close to real time" hazard event warning by combining state of the art approaches in an arguably not optimally suited experimental setup. It discusses these drawbacks as well as different ways to account for them. The document is mostly well structured, provides adequate background, justification and motivation of the study. The applied/developed methodology is clearly described and can be digested without major ambiguities. The study is well placed in the scope of the journal and I am confident that after some modifications, it will be a valuable addition to the journal’s portfolio.

Indeed, as the authors point out, the study is faced with suboptimal boundary conditions. The most important drawbacks are i) network geometry (linear array with 20 m station spacing), ii) a lack of independent control on the hillslope events and iii), a striking event type imbalance (10ËĘ1 hillslope events, 10ËĘ2 earthquakes, 10ËĘ3 noise cases). All these drawbacks are transparently mentioned, and their impact and counter measures are discussed in the text. Consequently, from a technical perspective, there is no reason to worry. However, it strikes my why this study design has been chosen to work with from the beginning. Why has this timely, rigorous and relevant study not been set up at a more suitable study site? There are many examples (cited in the text) where the network geometry is better (perhaps even in including a section of linear and densely spaced sensors to test the impact of such conditions, e.g. at the Sechilienne landslide), where there is excellent independent control on location, magnitude and to some degree the timing of hillslope activity, and where overall there are significantly more hillslope failure events that would lead to a less imbalanced data set? Somewhat, this excellent idea and study approach is vastly undersold due to the quality of the data.

The performance of RF or other ML techniques (e.g. HMMs) to distinguish between different seismogenic events has already been proven for ideal archived data (e.g. Provost et al., 2017). The original point of this study was to show that even though the network geometry and data availability are not ideal for the site, this method still gives valuable information on the occurrence of slope failure events. However, we understand that other data sets might have given better results in terms of the classification score. For this reason, we decided to include an event catalogue of seismic signals recorded with an array of eight stations at Illgraben, Switzerland. We will explore the performance of our proposed method on this data set, as well as have a look at how the classifier transfers from one site to the other. First tests have shown, that the performance of the classifier at Illgraben is similar to the performance shown for the Bondo site. Additionally, we found that a combination of both catalogues slightly improves the classification

results on the 2019 test data set of the Bondo test site.

Currently, a wider impact is impeded by the big question marks on the representativeness given that only a handful of slope failure events has been detected and this with a 270 % error (3 seismogram interpreted hillslope events versus 8 random forest-based hillslope events). Regarding the latter, while the abstract sounds quite confident (80 % prediction accuracy), the implementation of the approach does not. And it is a bit contradictory to claim the random forest approach would overcome manual inspection efforts to correctly classify an event, whereas in the discussion it becomes necessary to judge manually, which of the eight detected events is due to hillslope activity and which is an earthquake.

We realize that the representativeness is not ideal with such a small data set. We hope that by including the Illgraben data set, we can make better statements on this. Furthermore, we agree that the false positive rate is quite high, however compared to the more than a million data windows being classified in the 2019 data, we believe that 8 events falsely classified as slope failure events are reasonable. However, we should focus more on the false negatives instead of the false positives, as it is (arguably) more important to catch all events. Compared to other studies (e.g. Dammeier et al. 2016), the number of earthquakes detected slope failures is significantly smaller. We will rephrase statements in the text to highlight this better.

Long story short, I see two points that should receive more attention in the manuscript: i) a robust justification of the study site and experiment setup (Why working with an obviously unsuited network and missing event control?), and ii) a more thorough discussion of the classification errors, with due respect to the very small number of actual events and the resulting implications for the overall uncertainty.

1) See answer above 2) We will deepen the discussion on the above-mentioned points.

Regarding the classification quality part, one way that might be worth to explore is to use the hillslope events from the entire data set, not just the training subset. This of course only in the exploration of the classification quality (sections 4.1 and 4.2). The idea is to reduce the imbalance by increasing the number of hillslope events. In addition, this would shed some light on the actual impact of 5 versus 8 hillslope events.
We do not have a full labeled data set of 2019, but only the slope failure events, and checked the events that were falsely classified as slope failure events. This is why for an accurate testing of all classes, we are using parts of the 2018 data set. We agree however that the accuracy might increase by using the whole 2018 data set as training data and will test that.

**specific comments:**

The results section partly grades into a discussion. I recommend keeping these things separated, especially since there is a dedicated discussion section. Examples are l221-224, l249-254, l261-262, l268-269, l271-272.
We agree that the mentioned parts of the results section should be moved to the discussion section.

l20, I do not think it is necessary to use climate change as driver of this study. As in the abstract, it is sufficient to motivate by the mass movements, alone. But this is just a recommendation. No need to stick to that.
We would like to keep climate change as a driver for this study, as the threat for mountain communities will increase in the future, and simple and robust monitoring techniques will be the key for hazard monitoring and mitigation.

l34-35, check journal guidelines about order of references, here and throughout. Commonly, this is by date or author name, rather than apparently random order.
The references will be changed according to the journal guidelines.

l39-40, the larger amplitudes of slope failures must be compared to something. I assume you mean tremors. But the distance to the source will dominate the amplitude discussion. I suggest, to remove this misleading part of the sentence, it is of limited use, here. Overall, I am not sure the comparison of rock avalanches to tremors is a good one, especially in this journal and its readership.
We agree with this point and will remove the comparison to tremors from the sentence.

l48-54, well summarised. I suggest to pick that up in the discussion again, because like your routine the HMM approach also generates near-real time

classification of events. Thus, a verbal comparison of pros and cons of the two approaches is something the reader is interested in, and for good reason. Ideally, one would benchmark both approaches using the same input data, but I fear this is not feasible, here.
We agree that this is an important point, however, a benchmarking of both (or several more) approaches is not in the scope of this study. The pros and cons of both approaches will be added to the discussion.

l 55, the section about STA/LTA picking is a bit unfortunate, here. In the above paragraph you discuss detecting and classifying. Here you go back to just detecting. Would it not be more intuitive to first give a general introduction that defines and distinguishes detection and classification, and then elaborates on the different approaches to these two tasks? I suggest to write such a short introduction prior to l. 48. Then you can list the different approaches.
We agree that changing the order would be more intuitive and will do so in the revised manuscript.

✓ l76-77, that last sentence of the paragraph is actually results and discussion. I recommend to remove it here.

l78-80, in your scope, points a) and b) are not actually discussed and investigated. You do not write about decreased slope activity as a precursor of larger events or transitions of hillslope to channel activity. In fact, you cannot do this with only a hand full of events in total. I suggest to reword these points, here. Or simply collapse this paragraph with the above one after the corrections have been implemented.
We agree that this paragraph needs rewording, as the scope points in itself are not picked up again in the manuscript. We believe however, that our method

enables us to monitor an increase in slope activity or an early detection of hazardous events.

✓ l86, check SI unit conformity of volume numbers. Also see journal guidelines.

✓ l92, you may want to add more information about the loggers and recording frequencies, as well as on the installation of the sensors (surface, depth, coupling)?

l98, in the methods, I recommend adding the benchmark efforts that you discuss in section 5.2. This is a laudable and insightful test and it must be justified and described in the methods section.
We agree with the comment and will add a description of the STA/LTA benchmark to the methods section

✓ l101, check conformity of closing parenthesis in figure reference. Also, in other parts of the manuscript, this parenthesis is missing, check for correctness and consistency.

l134-135, this sentence kind of glances over a maybe important topic. Is there any way to show this more rigorously? I might suspect that i) local versus teleseismic earthquakes are quite distinct in terms of labeled features and ii) that smaller local quakes might be more similar to slope activity. Thus, could this lumping not be one reason for the result of 5 out of 8 hillslope event classifications being earthquakes? Usually, sentences that start with "After rigorous testing..." tend to hide potentially important subjective decisions instead of transparently showing the foundations of these decisions. Consequently, it would be good to be more transparent here, and show the effect of the lumped case versus for example two or three earthquake classes. Or at least to discuss

why for random forests it may be appropriate to stick to very small numbers of classes.
We do understand this criticism. It is true, that a large number of classes can improve prediction accuracy, due to a more accurate feature selection. However, as we are mostly interested in slope failures, we decided to keep the number of classes and the classification as "simple" as possible. Our first thought was also that keeping local, regional and teleseismic earthquakes separated would increase the prediction accuracy, this turned out to not be the case. We will follow the reviewer's suggestions and add plots showing the accuracy without lumping the classes in the appendix to back up this step.

l187, to account for the bias due to the imbalanced data set, can you not calculate the confusion matrix based on log-scaled numbers? I think in one of the Hammer HMM papers this has been done.
We think this is a good idea and will represent the confusion matrix in log-scaled numbers in the revised manuscript.

l216-217, why different colour schemes for the two matrices? It is not intuitive. No big deal but I may mention that it took me some thought to wonder why these different colours. Unless there is a reason (which should then be mentioned in the text/caption) I suggest to use the same colour scheme.
We used different colour schemes with one of them showing the AUC values and the other one is showing the 95% confidence interval. The different colour schemes were chosen to highlight the different meanings.

l220, reword, currently it reads as if RF and BRF are techniques at the same level as RF with US, OS and SMOTE. From the methods I read that US, OS and SMOTE are data manipulation steps prior to a subsequent RF classification,

no? Also, it would be good to actually discuss these findings later on (section 5). What does it mean that the imbalance countermeasures do not yield any improvement, but rather decrease the quality of the classification? What can we learn from that? What might be the reason?
Good point, we will reword that. Additionally, we understand that it might be misleading in the text, but the countermeasures do actually bring an improvement compared to an ordinary RF if we don't change the prediction threshold for RF. For this specific data set though, it seems that the accuracy for the slope failure class is highest if we do just that. From this we learn, that different techniques or improvements for an algorithm will not necessarily always give the best results for specific data sets. A few sentences on this will be added to the discussion part.

✓ l230, this number of 2 RF in the test data set comes out of the blue. Please revise and mention this at an appropriate place.

l258, the manual classification parameters must be defined in the methods (What are your classification judgements based on?). The image and radar methodology must be mentioned, as well. Also, since the catalogue is a key feature to validate your approach, I recommend to spend significantly more than just one short sentence on this topic, both in the methods description and the presentation of the resulting catalogue (a table or in the text).
Indeed, the catalogue is crucial for the testing. However, unfortunately not more information is available on that. We are in contact with the local stake holders who informed us on any reported events. We crosschecked radar and images, however not all events were caught, unfortunately, but no other events were caught either. We will elaborate on this in 1-2 additional sentences.

l262-263, I suggest you give more details here, in terms of description of the

events. It is only three failures, so there is space for that and it is important as the main goal of your study is to work out such events. Based on which criteria did you define these signals as hillslope failures? What are the event's properties? Also, in fig. 6, I only see one event and not all three. I suggest to plot the PSDs and seismograms also for the two other events, as in fig.6 c-e-g.
We will add the waveforms and spectrograms of the other events to the figure (or in a new figure) and describe the slope failures in the text.

l268-269, this is an unsupported statement. How are we to judge that this was an earthquake without seeing any data of it? Why do you think it is no hillslope event? Please present a PSD and seismogram as well as a more detailed description of the properties. This is the results section and it should present results sufficiently clear and exhaustive to allow you to draw conclusions from it.
We labeled this event as an earthquake as it shows a clear P and S-Wave arrival. However, we agree that a figure showing the event would be beneficial for the reader and will include one in the revised manuscript.

✓ l302, can you quantify this statement? What means high SNR, compared to what?

✓ l310, as mentioned above, this section should be motivated and described in the methods section, already. And its outcomes should be described in the results section, so that you can focus on the implications, here. Please revise.

✓ l311, delete comma after "shown".

l333, this is a valuable finding but strikingly out of context. Either include the runoff classification part from the beginning or leave it out (I recommend the

latter). Also, runoff appears to be a continuous feature rather than a comparably short lived event. In fact all PSDs of the manuscript show the seismic signature of water runoff. So why classifying it and how handling the case of two "events" occurring at the same time, such as runoff and rockfall?
We agree with the reviewer that discussing both runoff classification and “short” event detection in the same context seems contradictory. Accordingly, we will reword this discussion. Nevertheless, we prefer to leave the runoff part here, because though preliminary, our results suggest that the continuous RF classification can be applied beyond our study's scope.

✓ l342, revise this first sentence. Yes it is feasible, but with an error of 230 % (3 times right, 5 times wrong).

l345, rewrite "is a challenge that an imbalanced training data set enhances". Do you mean a challenge that is due to an imbalanced training data set? Or a challenge that may be solved by a less imbalanced training data set?
We understand this confusion and will reword accordingly.

✓ l349, manual inspection is not just advisable but crucial to account for the issue of misclassification, see comment two above. In the same line, replace "then" by "than" and "monitoring" by "inspecting".

l350-353, these are arm waving sentences. Either expand on this topic or leave it out. Currently this does not help the reader much. What is behind semi- and unsupervised ML algorithms, more specifically? Which specific drawbacks of the current approach would they solve? What are "unseen patterns"? I summary, I suggest not to mention this part, unless you find a way to explain its value in more detail.

We agree and will leave this part out.

Fig2c, value of that sketch is very limited. You may consider removing this panel. Fig.3, check font sizes, this is a really small font, hard to read. See journal guidelines on minimum size.
Fig 4, a and b homogenise colour schemes.
Fig 6, as mentioned above, also show other hillslope events, as well. Font on legend colour bar is too small.
All of these comments will be addressed (see comment above for color scale homogenization).

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2020-200, 2020.