General comments

This study has trained a neural network to automatically detect landslides part of multiple landslide events using satellite synthetic aperture radar data. The paper describes the processing of the satellite data and the setup of the neural network. Evaluation of the model is done through a qualitative analysis of 2 unseen events.

The idea and setup of the research is interesting and fits within the scope of GMD. However, the implementation and evaluation of the machine learning model requires major revisions, as described below.

The language and structure of the paper should also be improved: Many sentences are unclear; paragraphs are too long; and the information is not clearly structured.

Specific comments

Line 6: If you hide 2 events for validation, the model is only informed by 9 events, not 11?

Line 6-7: I cannot find any F1 scores, or any other statistics, for the unseen events in the paper? The F1 score given here seems to be from the test dataset? This dataset does not contain any unseen events, and thus this statistic should not be used to substantiate claims on model transferability.

Line 67, 72: full transferability requires meeting a very high bar. Based on your discussion I feel you do not clear this bar. You (rightly) mention events where the model would perform poorly and types of events for which you have no data. The performance of the model on the Haiti events show that the learned patterns are not fully transferable.

Google earth engine (GEE): Google earth engine is a very powerful tool for accessing and combining datasets. However, it comes with challenges for future reproducibility of research. The platform may be discontinued, or datasets could be removed or changed. Relevant for this research in particular is that is likely the pre-processing done by GEE will be changed in the future. Furthermore, code syntax may change which could break scripts. For future reproducibility it is important to describe the entire pipeline in a way that could be reproduced without access to any of the cloud tools you used during this research. I do not feel like I would be able to do this with the information provided in this publication.

Section 2.1 could probably be replaced by extending table 1. this would make the paper more concise and makes it easier to compare events.

Line 128-130: No need for a reference if you explain the mechanism behind the revisit frequencies.

Line 134-137: If I understand correctly from the provided link, the pre-processing is part of the GEE dataset. You may present it as such.

Section 2.2: There should be a clear separation between the description of the Sentinel-1 data/satellite and the dataset you used (GEE). 134-140 should probably have its own paragraph, or you may even add a section on preprocessing.

Section 2.2: pre-processing should be part of the method.

Line 163: To me it is unclear which combinations result in 28 datasets.

179-184: You mention the percentage of pixels classified as landslides within your image influences model behaviour and performance. Why have you chosen 5%? And why not include this variable within hyperparameter optimization?

Another question I have here is if and how you include this category in the test data? Some of the false positives you find may actually be true positives.

Line 187: Note here that a larger bounding box increases the baseline chance of the box containing a landslide. Looking at Figure 3, randomly drawing a 64x64 box already gives decent odds of "finding" a landslide. A larger box makes it easier for the model to make correct prediction.

Section 3.1, dataset design: In the results, discussion and supplement sections you show the model can have same difficulties based on environmental variables such as the slope, aspect and land cover. What happens if you give this information to the model?

Section 3.3: It would be very helpful to have a figure that represent the architecture of the neural network. The figure from the supplement should be in the paper. The model description in the text is not entirely clear.

Section 3.3: The architecture and reasoning behind the model may be of particular interest to the readers of the GMD journal. You can elaborate more on your choices in choosing this architecture, and why you decided to modify the model from your previous publication. Why not use a simpler or more complex setup? Especially because your model architecture was not part of your hyperparameter optimization.

Line 200: Because this is the only place where you describe an activation function it seems like there is only one activation function within the network. I assume there are more.

Training validation and test split (line 208-214): The main goal of this publication is to create a generalizable neural network to detect MLEs. However, the setup of the training, validation and test data is unfit for generalizability, because you validate, test and optimize your model with events the model has seen during training.

The validation partition is used to determine when to stop training the model to prevent overfitting. But because the validation partition is sampled from the same events as your training partition you will keep training for a long time, seemingly without overfitting. However, the unknown events you are trying to predict are different from the known events in your training and validation partition; as a result, you have (significantly) overfitted your model to the training and validation and test data. A similar problem occurs with the hyperparameter optimization, where you are overfitting the hyperparameters to the known events.

You have recognized this problem, and you have hidden 2 events from the model to evaluate the final performance. But by this point the model is already overfitted and overoptimized. This means there is performance left on the table. The paper does not contain metrics for the predictions on the unknown events, these should be included. Traditionally metrics on the unseen data are used to evaluate the generalizability of the model. If there is a large difference in model performance between the seen and unseen events the model generalizes poorly and vice versa. Such metrics allow the reader to judge the generalizability of the model.

To train a generalized model the training, validation and test should consist of different events, otherwise you are unable to recognize when the model or hyperparameter optimization is overfitting to your known events. The validation dataset can even be rotated such that every non-testing event has been used for validation.

The method should be repeated while training, validating and testing your model with such a train/validation/test split. Consider adding a train + validation loss curve to your appendix or supplement. This allows the reader to see how you prevent overfitting the model and hyperparameters.

Even with this setup full generalizability is difficult to prove as your dataset is missing data for various locations and environments. 11 events just are not enough data to support such a broad claim.

Training data (table 1): How does the different number of landslides per event influence your training? Some landslides have a very low number of events, such as: Capellades, Milin, Mestas. Thus, learning to predict these events will be of little value for the model. Yet, learning these events may be crucial for generalizability. It is like a class imbalance, but for events instead of prediction classes. It may be interesting to experiment with some class imbalance techniques such as oversampling the data from the small events.

Hyperparameter optimization, Line 221: Testing only 2-3 values during hyperparameter optimization does not provide much insight into the effect of the hyperparameter on the model. Also, you are unlikely to find the (near) optimal configuration for any parameter. With the small range of values tested, you may as well have picked a value based on expert judgment. Consider increasing the range for the parameters you optimize, even if you must decrease the number of parameters you optimize.

Currently there is no way for the reader to assess the sensitivity of the model to the choice in hyperparameters, consider adding a table/figure to the appendix or supplement.

Essentially temporal buffer lengths are just another hyperparameter you have optimized. For the structure of the paper, it may be nice to have a section on hyperparameter optimization, including the temporal buffer lengths.

Section 3.4: The title does not provide much information on the contents of the section.
Section 4.1: Please also provide these metrics for the unseen data. Because the test datasets are part of the same events the model was trained on, they are a poor benchmark of model performance.

Line 257: There is no table 3.1, you probably mean table 4?

Table 4: Considering the objective of you model for use in rapid assessment keeping a 12 day post event stack is probably a good decision. Even though model performance improves significantly with a longer post event stack

Section 4.2: The performance of the VV models is significantly lower than the VV_VH models. Can this all be attributed to the inclusions of the extra datasets?

Section 4.2: Please provide metrics for the Sumatra and Haiti events. Also, a quantative analysis in addition to the qualitative analysis would be helpful. This provides a more objective way to assess the performance of the model in unseen situations.

Figure 3&4: The outline of the landslide inventory is difficult to see on the green and grey background.

Figure 4: Why is the model failing to predict so many of the landslides on the eastern part of the island?

Line 266: It is may also be the case that the environment was poorly represented in the training data. Together with the aforementioned overfitting on the training data this can (partly) explain the poor performance of the model in this case.

Line 268: It is also likely these overpredictions result from the model overfitting to the training data, where it has learned to associate some patters with landslides, that shouldn't be.

Section 5: I'd like to see some discussion (and quantification) about the performance of your SAR model against more traditional optical models; How does this relate to their applications?

Figure 6: Looking at the pixel contribution maps even the true positives are missing some of the landslides from the inventory. In the case of 6b the pattern of the pixel contribution is also different from the landslide. How accurate is the inventory?

Figure 6: A True negative and a false positive may also provide some interesting insight into the workings of the model.

Line 327: This statement should be supported and consistent with your results, not by a citation to another paper.

Line 330: The model has difficulties in predicting landslides of the Haiti event. Your statement here seems inconsistent with your results.
Line 331: In your discussion you mention there are various environments without a proper landslide inventory. This seems contradictory to the statement in this line.

Line 349: Why is this pixel threshold not part of hyperparameter optimization?

Line 359: You mention there is no apparent bias given by the landcover. However, figure S2 shows a much higher False negative and False positive rate for herbaceous vegetation and open forest compared to closed forest. For these landcovers the False negative and False positive rate is higher than the True positive and True negative rate; It seems that for herbaceous vegetation and closed forests the model performs worse than chance.

Line 361-370: Interesting section about the impact of slope on the classification
Section 5.3: How about additional / improved landslide inventory datasets?

A2&A3: In the figure it is not clear what configuration you have decided to use.

A2&A3: Separate metrics for final predictions on the unseen data should also be added. (could also be separate, or part of the next suggestion)

A2&A3: Considering this is an imbalanced classification problem, it would be helpful for the reader to show the baseline accuracy.

Appendix: A classification confusion matrix of prediction from the final model on the test data and the unseen data would be interesting addition.

Corrections

Line 16-17: Grammar

Line 19-20: No need to add the names in addition to the reference.

Line 24-26: Grammar

Line 29-30: Grammar

Line 33-34: Grammar

Line12-53: This information should be divided over multiple paragraphs.

49-51: changes in Anomalies?

Line 129: please explain the acronym GRD

Section 3.3: Please split the information over multiple paragraphs with clear subjects.