Comment on nhess-2021-299

Anonymous Referee #1

I have revised the manuscript "Assessing the importance of feature selection in Landslide Susceptibility for Belluno province (Veneto Region, NE Italy)", submitted by Sansar Raj Meena, Silvia Puliero, Kushanav Bhuyan, Mario Floris, Filippo Catani, focused on the the importance of feature selection for landslide susceptibility zonation. The manuscript could be interesting for the journal but requires a strong revision. Data are not clearly presented, some definition are confusing, the structure needs to be reorganized. Moreover, the manuscript is not written in a good English and some sentences are difficult to understand. I recommend the authors to submit a revised version of the manuscript after a revision by an English-speaking person. Detailed comments are throughout the text.

Dear Reviewer, thank you for reviewing the manuscript. We have marked the changes using track changes manuscript (Line numbers are referred to track changed file).

Commentato [r11]: Do you mean “thematic data”? Feature is confusing (conditioning factors

(**features**) in the overall prediction)

Ans: Yes, the word “features” seemed quite confusing so we hereby change it to “conditioning factors” throughout the manuscript. Please refer to the first change in line number 20.

Commentato [r12]: Interesting statement that should be explained better in the text

Ans: We have explained and discuses this statement later in the text.

Commentato [r13]: Not clear (eliminating the least available ones in most of the use cases due to data scarcity)

Ans: Yes, we agree with your comment, and we hence removed the sentence to avoid confusion.

Commentato [r14]: Do you mean recover? (post-event relief measures)

Ans: Yes, we agree with your comment, we meant post event recovery measures.

Commentato [r15]: Not sure this is the correct word (activities such as **informal** settlement development and cutting of roads along the slopes).

Ans: We have replaced it with **unplanned** settlement. We refer you to line 38.

Commentato [r16]: Not sure it is the correct word (it is crucial to **realize the significance** of landslide studies)

Ans: We have rephrased the sentence and refer you to line 51-52 in manuscript.

Commentato [r17]: Please rephrase (**Examples of such approaches is given in the study area, by Floris et al. (2011) which combined traditional LSM methods with an updated online landslide database in the Veneto Region, Italy, where they used online spatial data from Italian portals for mapping landslide susceptibility at medium and large scales).**

Ans: We have rephrased the sentence and refer you to line 70-72 in the manuscript.

Commentato [r18]: Not the correct word (However, despite advances in LSM, the **advent** of feature importance)

Ans: We changed the word accordingly and amended the sentence. We refer you to line 88-89.

Commentato [r19]: It would be interesting to see where is the area affected by the VAIA windstorm

Ans: We refer you to updated figure 1.

Commentato [r110]: Why cost?

Ans: We have replaced cost with computational time, please see line 99.

Commentato [r111]: What do you mean with postprediction and pre-prediction?

Ans: We refer you to line 109-111 for the rectified sentences.

Commentato [r112]: Not clear. Please rephrase

Ans: This sentence was removed entirely as we felt that it did not serve much purpose.

Commentato [r113]: Not clear feature vs factors

Ans: We have changed it to "conditioning factors" throughout manuscript.

Commentato [r114]: This means that your analysis cannot be exported to other test sites?

Ans: This the statement we meant that due to data scarcity in other regions of the world, it can be possible that same conditioning factors and data may differ. We want to point out that our analysis can be performed in other regions around the world by utilising the available datasets for that area.

Commentato [r115]: The most important locations should be shown in the map

Ans: We refer you to updated Figure 1.

Commentato [r116]: IN the table you should provide information referred to your test area

Ans: we have included the information related to our test area in table 1.

Commentato [r117]: What is the scale of the maps?

Ans: Scale of the maps varies for landcover, lithology maps and road and drainage data were in vector format, However, we resampled them to 25 meters for our analysis.

Commentato [r118]: Some of these justifications are very well and the description are not useful

Ans: Based on your suggestion we have improved the justifications and reduced the descriptions. We refer you to Table 1.

Commentato [r119]: Which is the radius in your study?

Ans: Please refer to the previous answer. We have improved the justifications.

Commentato [r120]: This is not the case if you use average monthly data. What is the resolution of this information?

Ans: We used daily hourly rainfall data.

Commentato [r121]: Low permeability is not always true

Ans: We changed the sentence accordingly.

Commentato [r122]: What do you mean?

Ans: Please have rephrased the sentence. Kindly, refer you to justification of the “Landcover” in table 1.

Commentato [r123]: The dimension of these maps can be reduced

Ans: We have improved the figure 2.

Commentato [r124]: The legend is in Italian

Ans: We have changed the lithology Legend to English.

Commentato [r125]: The entire chapter should be better organized. You should start with the description of the flowchart and then describe the single models

Ans: We have changed the position the framework diagram and organised as suggested by following the description of the single models after the flowchart.

Commentato [r126]: Explain better what do you mean

Ans: We have changed the sentence. Please refer to line 199-202.

Commentato [r127]: Not clear. Why space-time?

Ans: We have changed the sentence. Please refer to first paragraph in methodology .

Commentato [r128]: Rephrase because it’s not clear.

Ans: We amended by removing the sentence altogether due to the confusion it caused. We improved the first part of the methodology.

Commentato [r129]: Do you mean “for each factor”?

Ans: We have changed the text to “landslide conditioning factor”.

Commentato [r130]: So far, you didn’t explain the methodology but the FR. The flowchart should go later

Ans: We have put the methodology flowchart before section 3.1 for better organisation of the text followed by it’s description.

Commentato [r131]: A bit confused

Ans: We have rephrased the sentence. We refer you to line 235-239.

Commentato [r132]: The flowchart should show the steps described below. The “Feature selection algorithms” is for example missing

Ans: We have improved the flowchart based on the suggestions. Please refer to figure 3.

Commentato [r133]: Equation 1 is missing Where is this step in the flowchart? What is the mapping unit of your analysis?

Ans: We have renamed equation 2 as 1. We have updated the flowchart, we have carried out the analysis at pixel level.

Commentato [r134]: Do you sum weights?

Ans: We have updated the Equation 1, please check for clarification.

Commentato [r135]: What do you mean?

Ans: We have rephrased the sentence. We refer you line 255-257.

Commentato [r136]: This means that the average value of LSI is 1?

Ans: We have rephrased the sentence. We refer you line 255-257.

Commentato [r137]: Something wrong… If LSI is the landslide susceptibility Index what do you mean with correlation? Correlation between what?

Ans: We have rephrased the sentence.

Commentato [r138]: ?????

Ans: We have rephrased the sentence.

Commentato [r139]: You didn't mention the training data set before. How did you define it?

Ans: We have remove the word to reduce confusion in text.

Commentato [r140]: Not clear. Rephrase

Ans: We have rephrased the sentence.

Commentato [r141]: Explain better how the model works in the case of LSM

Ans: We have addressed this issue by explaining how the random forest model is used for modelling landslide susceptibility. Please refer to line 292-297.

Commentato [r142]: Explain what is the meaning in case of LSM

Ans: We have explained better how the core idea of XG-Boost helps in the case of LSM. We refer you to line 311-315.

Commentato [r143]: Explain better how the model works in the case of LSM

Ans: We have addressed this issue by explaining how the XG-Boost model is used for modelling landslide susceptibility. Please refer to line 311-315 .

Commentato [r144]: Where is this step in the flowchart?

Ans: We have added this step in the flowchart. Please refer to figure 3.

Commentato [r145]: Too general... what do you mean?

Ans: We have edited the sentence. We refer you to line 329-330 .

Commentato [r146]: Not shown in the flowchart

Ans: We have added this in the flowchart. Please refer to figure 3.

Commentato [r147]: Not clear

Ans: We have edited the sentence. We refer you to line 331-335.

Commentato [r148]: Equation 1 is missing

Ans: We mistook the equation numbering for this section and have amended them.

Commentato [r149]: How did you utilize them?

Ans: We used class weights of each factor derived from LSI values and factor weights from predictor rate.

Commentato [r150]: This should be figure 6

Ans: Thank you for pointing it out, we have written it as figure 6.

Commentato [r151]: Why 0.30?

Ans: Based on trial and error approach, we find out that coefficient values lower than 0.30 were not contributing a lot to landslide susceptibility results.

Commentato [r152]: What is the value of the y axes?

Ans: Value of y axis is coefficient values

Commentato [r153]: What is?

Ans: we have removed the word to avoid confusion.

Commentato [r154]: IN the map you show validation landslides. These are failures not used to prepare the model.

Ans: We agree with you in this study we did testing so, we have updated the figures with testing landslides.

Commentato [r155]: I think this is the model skill not the prediction capability

Ans: We changed “prediction capability” to “predictive skills”. We still consider that the “prediction ability” of the two models remain similar and hence choose the words to “predictive skills”. Refer to line 377-378.

Commentato [r156]: Not clear how it’s possible to use table 2 to see the number of pixel

Ans: we have removed the sentence to avoid confusion.

Commentato [r157]: This should be explained better because it’s not clear

Ans: We have explained this in the text but to recall, we tried countless values as a cut-off or threshold value to see which of the conditioning factors gave the best accuracy for the susceptibility after removal of the factors based on the cut-off value. Refer to line 388-393.

Commentato [r158]: Why? (0.03)

Ans: Please refer the answer to comment for r157.

Commentato [r159]: Why 10 and not 9?

Ans: This is because using 9 layers gives the best accuracy, not 10. And since we are interested in the highest accuracy that we can get after factor removal, we choose 9 as the layers.

Commentato [r160]: IN the map you show validation landslides. These are failures not used to prepare the model?

Ans: We agree with you in this study we did testing so, we have updated the figures with testing landslides.

Commentato [r161]: Please add the legend of the y-axes

Ans: Value of y axis is coefficient values

Commentato [r162]: Is this the model skill or the validation skill?

Ans: We have updated it with accuracy assessment.

Commentato [r163]: Did you define a training set and a validation set? How did you define them?

Ans: We defined 30% testing sets randomly for testing the models.

Commentato [r164]: Need to be rephrased

Ans: we have rephrased the sentence.

Commentato [r165]: Not clear

Ans: we have rephrased the sentence.

Commentato [r166]: ROC curve and success rate curve are two different thinks.

Ans: we have rephrased and improve the caption of the figure 11.

Commentato [r167]: How do you compute the percentage of land using points?

Ans: We mean here percentage of area susceptible to landslides.

Commentato [r168]: Is this index reliable using points to compute ni?

Ans: ni is the percentage of area susceptible to landslides, it means pixels which are labelled as susceptibility class are used. This index does not use point dataset rather uses the overall susceptibility raster and check number of landslides that fall in that particular susceptibility class.

Commentato [r169]: You have mentioned the quantile classification to define the susceptibility classes (the classification separates the values into groups with an equal number of values). Why the number of pixel in each class is so different?

Ans: We checked the classification and it was natural breaks instead. Therefore, we amended the classification in the manuscript as “natural breaks”.

Commentato [r170]: There is a problem with the dimension (km2??)

Ans: we have changed it to m^2.

Commentato [r171]: The chapter should be reorganized and revised by an English speaking person

Ans: We carefully checked the English and fixed the issues accordingly.

Commentato [r172]: Explain better the importance of feature selection in Landslide Susceptibility if it doesn’t affect the results

Ans: We refer you the line 541-544 where we have explained the reasoning to choose the conditioning factors according to the feature importance.

Commentato [r173]: IN the article you do not provide this type of information

Ans: For the analysis, we used resampled maps of 25 metre pixel resolution.

Commentato [r174]: How can you confirm this statement?

Ans: We have removed this statement as it did not play any role in the matter of this methodology/study.

Commentato [r175]: This means that feature selection is not relevant?

Ans: It means that after performing feature selection, we now know which of the conditioning factors to remove and thus still retain high accuracy in the susceptibility prediction.

Commentato [r176]: What do you mean?

Ans: We mean that unlike the ML models which gets introduced to instances of both landslide and non-landslide samples, the statistical model is not trained with non-landslide samples, but simply with the landslide samples for landslide susceptibility.

Commentato [r177]: The chapter should be reorganized and revised by an English speaking person

Ans: We carefully checked the English and fixed the issues accordingly.

Commentato [r178]: Check the list. Some references are not completed

Ans: we have improved the references.

Commentato [r179]: Non completed

Ans: we have improved the referenced.

Commentato [r180]: Non completed

Ans: we have improved the referenced.

Comment on nhess-2021-299

Anonymous Referee #2

I have gone through the manuscript and found that the quality of work is very good and applied. I have some observation needs to be correct before its goes to final publication.

1. The words features and factors are used interchangeably in the paper. Better to stick to one word.

Ans: Thank you for suggestions. We use the words “conditioning factors” throughout the manuscript to avoid further confusion.

2. Definition between pre and post predictions are not clear. Making it difficult to read and understand the context sometimes.

Ans: The “pre-predictions” refer to the moment before actually training the model and thus, we refer to some literature that performs factor importance prior to model training. Similarly, “post-predictions” refer to factor importance after model training, which we perform in our study here. Nonetheless, we have edited these two words in the document to avoid confusion.

3. An image of the affected area for example could be very insightful to comment on the extent of the damage over the area.

Ans: Thank you for your comment. We have added some images in figure 1.

4. Scale of the maps that are taken into the experimentation are missing.

Ans: We have added the scale information in the manuscript.

5. Explanation of the methodology can be better, especially the starting paragraph.

Ans: We have re-arranged the paragraphs along with the conceptual framework diagram to make a more comprehensive and suitable readability experience.

6. The mapping units are not defined as of yet, which must be mentioned.

Ans: We have added this information in the manuscript. We have done the analysis at pixel level for our study area.

7. Better explanation of the models is required, mostly in the case of LSM and how these models learn and predict using the models for LSM.

Ans: We have explained the models better keeping in mind the usage of them in the context of LSM. Please refer to sections 3.2.1 and 3.2.2.

8. The reasoning for 0.3 as the threshold must be reasoned better.

Ans: We have given the reasoning of this in line 388-393. But to recall, we tried countless values as a cut-off or threshold value to see which of the conditioning factors gave the best accuracy for the susceptibility after removal of the factors based on the cut-off value.

9. Graph axes have no labels.

Ans: We have added the y-axes labels in the graphs. We refer you to figures 6 and 10.

10. No definition of training and testing datasets for model prediction. Need a section for that.

Ans: We have added the definition of training and testing datasets for model prediction in section 2.2.