Review of *Machine learning nowcasting of the Vogelsberg deep-seated landslide: why predicting slow deformation is not so easy*

The work presented by Natijne et al. provides an in-depth investigation into why their machine learning nowcasting model did not provide adequate results at the Volgelsberg landslide in Austria. The authors use a state-of-the-art machine learning model that incorporates detailed deformation data and remote sensed environmental data to try and predict deformation responses through time. The article is well-written and does a good job exploring the limitations of machine learning approaches. The article is important as machine learning methods are often heralded as a panacea. However, as this paper outlines, machine learning methods need to be applied to the right types of problems with the right type of data of adequate quality to provide effective solutions. While I believe the paper is well constructed and has valuable findings, I do have a few concerns that need to be addressed before publication. Below I provide a few general comments, followed by a line-by-line issues that should be addressed.

General comments:

1. You provide a table of several research articles that have produced now-cast models. In section 2, you also say that at least some of these articles were for deep-seated landslides. I think it would be valuable for you to explicitly highlight what was different between this study and the ones that seemed to have success with now casting. Do you know exactly why they had success and you didn't? You discuss the reservoir being a factor for some of these studies, which makes sense. However, did any other studies try to nowcast deep-seated landslides with success?

2. I'm a bit concerned by the environment predictor data you use in you model. The spatial resolution of the satellite data is often much coarser than the size of the landslide. I have a hard time seeing how these coarse resolution datasets could provide any meaningful information at the local scale you're looking at.

3. I also have concerns with the deformation data itself. If the data is so noisy, are you sure you can trust it at all? Are you certain that the deformation signal you are trying to replicate isn't an artifact? If you are confident in the deformation signal, please better demonstrate why to the reader. For example, provide the accuracy of the data, the preprocessing step used by the Division of Geoinformation of Tirol, and elaborate on the corrections to the measurements. Based on the smoothing you conducted to get a usable signal, I find it difficult to trust this data. The poor quality deformation data may be partially responsible for your poor model fit.

4. Other studies (e.g., Thomas et al., 2019; Yatheendradas et al., 2019) have assessed the utility of satellite-based weather data for slope stability and found mixed results. I believe its possible that many of your issues could be attributed the poor representation of satellite data for your study site. I think this merits discussion or justification for why you think this is not an issue.

5. I don’t think section 2 in necessary for this paper. I found it to be a bit burdensome and ancillary to the main point of the paper. Consider trimming it down to only describing the points that are pertinent to the model you use and then putting it in the methods and/or discussion where you describe the model(s) you use in this study.

Line by lines:
11: Please clarify what you mean by “standard quality metrics”.

35: “correlate *with* time-delayed”.

50: Please provide citations for this claim.

58: Consider, “…may reveal slope processes *responsible for landsliding*”

68: Consider “…naive modelling. *That is, the model is* unaware…”

76: Please don’t start a paragraph with “therefore”. It is confusing as a reader. Reintroduce the idea you are discussing.

77: This is a sentence fragment. Please fix.

122: This sentence has a typo. I think you need to delete “in less”.

123: Define epoch.

123: The number of data points required largely depends on the type of model used. Neural networks require a lot more than other methods (e.g., logistic regression), both of which fall under the term ‘machine learning’. Consider being more specific by what type of machine learning method you’re referring to.

Figure 1: Please put sub-figure labels on the figure and in the text to help orient the reader.

Put a box around the right sub-figure to show extent of the bottom left sub-figure.

On my computer, the colors of the land slide are not clear. I see red to the north, then yellow, then green going south-east. Also why are there three colors? What is the 'overlapping area'?

151: Please show all the data somewhere (appendix) so that the reader can see this.

153: change ‘till’ to ‘until’.

158: please define ‘operational system’.

161: remove comma and change to “aim *is* to predict”.

162: I'm not sure what you mean by "no precursory deformation data is included". In line 123 you say that you give it 32 days of data and section 4.4 describes the different lengths of time used for training the models. Please clarify your meaning.

164: why 4 days?

176: Provide an overview of the numbers for us (absolute max, absolute mean, etc).

182: I don't see Wattenberg on Figure 1. Please include it.

183: You already describe the moving average. I don't think it needs to be repeated here.

206: I'm not sure what 'support the model' means.

207: Trial and error on what?

214: It also resembles the sm_profile data.

Section 4.3 I think here is where you should explain how neural networks work. Not above. And I suggest only including enough information for the reader to understand your model.

218: What do you mean disturb the model? You do this to keep the data at the same scale as the training data.

229: Please justify why you decided to share the memory rather than develop two individual models for the two benchmarks.

232: times series of deformation data, right? Please clarify.

234: I don't really see the value of this paragraph. Consider omitting.

291: But do all the examples have a strong driver? Why did the ones without a strong driver work well and yours did not? Do their study sites have different scales compared to this one?

Figure 8: please define the shaded background.

320: "*The* major challenge for the model…"

329: Please explain why reducing the number of parameters matters. Preventing over fitting, right. I think you say this at the end of section 2.2 but it was pretty convoluted. See general comment #5.

344: If the traditional least squares isn't tested why do you include it.

347: Can you be more specific on how you created this deformation rate model? Do you describe this somewhere that I missed?

Figure 11: I think you should consider merging figures 2,6, and 11. They show basically the same thing except some annotation.

424: I don’t understand why this is the introductory sentence of this paragraph. Consider rewriting.

462: Just write out the meaning of SNR.

528: What assumptions? Please restate.

Figure 13: I don’t see a reference to this figure anywhere in the text. Only the caption of figure 10. I think you should discuss this somewhere if you’re going to include it in the main text.

558: Amplitude of what?