Changes to *Machine learning based nowcasting of the Vögelsberg deep-seated landslide: why predicting slow deformation is not so easy*. [egusphere-2022-950]

We thank the reviewers for their detailed and thorough comments. We were pleased to read the reviewers appreciation on the quality of the work, as well as the recognition for the need to explore the limitations of machine learning.

Here we would like to summarize the changes made based on the discussion. Where possible the comments by both reviewers are merged and/or intertwined. Comments by Katy Burrows are shown in blue. Comments by the anonymous reviewer are shown in orange.

- **Environmental conditions**

  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.

  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.

  Another thing that could adversely affect your model could be that small spatial scale rainfall events might not be captured by your satellite rainfall products, which have quite a coarse resolution (although unfortunately there would not really be any solution to this)

Satellite precipitation products are notoriously bad at measuring peak precipitation. However, unlike shallow landslides, deep-seated landslides are less sensitive to short periods of intense precipitation, illustrated by the 20-60 day lag time between precipitation and accelerated deformation (l. 147, now l. 152). A bias in the precipitation data, a consistent underestimation, for example, would be counteracted by the scaling in the neural network.

Thanks to the study by Pfeiffer et al. (2021) we know that the Vögelsberg landslide has this slow response (l. 147, now l. 152) and precipitation higher up in the catchment is relevant to the landslide as well (l. 526, now l. 536). Therefore, infiltration is a catchment wide rather than a slope limited process, approaching the lower resolution of the satellite precipitation measurements.

We agree that the spatial-temporal resolution of the satellite observations is not optimal. However, these products are operational globally, and the necessary foundation for an operational nowcasting system.

- **Deformation data**

  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 Divi-

In May 2016 a permanent automatic tracking total station (ATS; Leica TC1800) was installed on the opposite valley flank. Within the active landslide and in the surrounding area assumed stable, 53 retroreflecting prisms were installed on buildings and other man-made structures in a range between 600 and 1700 m. With this setup hourly measurements were conducted automatically, aggregated to daily means. However, the resulting displacement time series showed a drift towards north-west, owing to a minor movement of the installed ATS itself. This drift was removed assuming that reflectors at locations around the actively moving landslide did not move (no signs of damage owing to ground movement were observed). For each epoch a transformation matrix was derived by matching the coordinates of the respective measurements at stable locations to their initial positions at the beginning of the monitoring (arithmetic mean of first five measurements) based on the closed-form solution for rigid body transformation provided by Horn et al. (1988). By applying these transformation matrices to all measurements of each epoch, including the reflectors in the active landslide area, the observed pseudo-movements at stable reflectors were corrected in all displacement time series. The finally achieved accuracy of the displacement time series is in the order of ±0.54 cm/a (Pfeiffer et al. 2021).

The text of §3 (l. 148) has been expanded to reflect this.

- **Previous work (Table C.1)**

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.

Table C1 This is a comprehensive table, but since not all the studies were included in the original table from Van Natijne (2020) there is no definition of some of the acronyms e.g. GRNNS.

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?

In section 2 we provided a concise introduction in now casting of deep-seated landslide focused on machine learning. This is, in our opinion, important to provide the reader a broader perspective of our work on landslide nowcasting and explain our findings later on. Except for one, all methods in the table are some form of machine learning. We did not look into the model or optimizer that was used as we did not compare results. Therefore, we have removed the column with methods. We prefer to keep the table, as it contains the expected references to earlier work. Most of these studies

2

were on deep-seated landslides, that did not undergo catastrophic collapse. This has been added to the text and table caption.

All the cited papers that claim to be successful either had a strong driver (l. 72, now l. 73), or split the signal in some way. In that case they typically subtract a trend from the data and apply a complex machine learning model to capture the smaller deviations from that trend. In Figure 10 we show that the mean/trend is already a very good predictor.

- **Minor comments**

The title has been changed to "Machine learning *based* nowcasting […]".

To support colorblind readers symbols were added to the color-coded lines in Figures 10 and 12.

> 50: Please provide citations for this claim.

This statement was not intended as a claim, but as a 'vision' for the paper. It was reformulated as such.

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

The data of all benchmarks is featured in Pfeiffer et al. (2021), Figure 3. A more explicit reference has been added to the text for the curious reader.

> **Line 157-158** "Furthermore the amplitude of the filtered signal lags behind the original deformation signal" Why is this? Is it a side-effect of the filtering or have you done it deliberately?

This is an unintended consequence of the filtering. This line has been removed to emphasize the previous, more important, statement that the onset of acceleration is severely dampened by this filter.

> 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.

Our model was designed to work in the absence of recent deformation measurements, and work on environmental conditions only. This sentence has been rewritten to clarify this.

> 164: why 4 days?

A four day prediction would demonstrate the model's ability to predict a tipping point based on the environmental conditions (acceleration, peak, deceleration). Furthermore, a four day prediction would give sufficient time for further investigation as part of an early warning system. This has been clarified in the text.

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

The mean and maximum correlation have been added to the text.

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

Agreed. However, in our perception the redundancy is desirable for consistency.

The deformation nowcasting model. This has been clarified in the text.

Line 209 Is API calculated from the ERA5 or GPM datasets? Does it make any difference which one you use?

Although better performance may be expected from the re-analysis in ERA5, the API time series is based on the GPM data, as this is an operational system that has data available with limited latency. This has been added to the text.

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

Indeed, the trained model will be sensitive to uncontrolled scaling of the input data. It has been clarified in text that the scaling of the data should be consistent with the scaling used during training.

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.

For the sake of brevity, we prefer to limit ourselves to giving a suitable reference here. We have added a reference to Jian et al. (1996) as it is easily readable and complete on the workings and possibilities of such networks.

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

The text has been expanded to clarify that this was done to reduce the number of parameters. The later model lstm3-32 uses three memory cells, increasing the number of parameters required by a factor of three for the simplest combination of input parameters.

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

The reviewer is right that this paragraph is not clear. The conclusion, that from a parameter perspective a more information rich variable should be preferred over a more complex model, was missing and has been added to the text.

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?

All studies in Table C.1 were conducted at slope level. Studies featuring multiple slopes often apply the same methodology to multiple slopes within the same catchment or along the same reservoir. We are unaware of regional studies that feature an automated analysis of many slopes with diverse drivers, as we mention in our outlook (l. 543, now l. 555).

Figure 6 The shaded areas for lines 1-4 begin before the deformation measurements. Is this a mistake?

Figure 7 Like my comment for figure 6, I wonder why the shaded area starts before your deformation dataset. Is this the "warm-up time" you describe in your figure caption? Or is that the

> part starting from early June 2016 where you have deformation data but no prediction? Maybe you could label this warm-up time on the time series with a box or shaded section?

This is not a mistake. The deformation training data is only available after the first set of 32-days of total station observations is available to the moving average filter. After that the model needs another 32-days to warm-up the memory in the LSTM-nodes. Theoretically these periods could have overlapped, at the cost of complicating the model.

This explanation has been added to the captions of figures 6 and 7, and the period has been marked in the figures as well.

> **Section 5**, **line 274-292** In the end, your model does not contain any snowmelt input, although you expected this to be relevant for the landslide. Could this be why your model only predicts the training data well in Summer and Autumn? Does a model including one of your snowmelt inputs (V3 or V4) predict Spring and Winter better (even if it's worse over all 4 seasons combined?)

This is an interesting hypothesis, that touches upon the information content of the variables, as well as the performance metric. However, evaluation of the performance over individual seasons would further reduce the length of the already short validation data set. We have added this suggestion to §6.1.1.

> 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.

Agreed, the goal of the reduction of the number of parameters have been briefly restated in this section of the text.

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

Traditional least squares would be the obvious solution. Given the resemblance of the da-32 model to the least squares solution, it was mentioned for reference.

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

The average deformation rate was calculated over the training period, and used as a prediction for the remainder of the time series. This mean squared error of this parameterless 'model' is shown in Figure 10. The text has been expanded to clarify this.

> **Section 6.1.1 Lines 360-365** Your R^2 value (0.31) seems low, but interpreting a single R^2 value is difficult. I'm not sure it's useful to include this metric when you have nothing to compare it to.

Agreed. We have removed this whole paragraph.

> **Section 6.1.2** I think I would have put this subsection in your methods section as it contains similar information to Sections 4.1 and 4.2

Indeed, the first paragraph replicates most of the considerations in §4.2 and has limited value for the discussion. We have removed this first paragraph of §6.1.2 and renamed section §6.1.2 to 'derived variables'.

> **Lines 409-413** If you separate the two benchmarks in the model, would this result in them no longer being connected in space? (So one could accelerate independently of the other). I would have thought that since one part of a landslide moving is likely to destabilise another part, separating them would be a disadvantage.
>
> Actually, I think I have misunderstood what you mean to say in these lines, can you find another way to write this?

Although no spatial relationship is provided to our model, the relationship is found during training as two realizations of the same slope process, represented by the shared LSTM-node. This has been clarified in the text.

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

Figure 2.6 shows the training periods used by the machine learning model. Figure 2.11 shows the accelerations periods as previously identified by Pfeiffer et al. (2021). The annotation is the main purpose of the figure, not the deformation time series. Therefore, no amplitude (y-axis) is shown in Figure 2.6. Figure 2.11 could be integrated into Figure 2.6, although the acceleration periods by Pfeiffer et al. (2021) are not introduce until §6.1. Therefore, we prefer not to merge the two figures.

> **Lines 480-486** I would specify Sentinel-1 since the temporal resolution of SAR satellites varies. Also, it is not clear here whether you are suggesting the use of InSAR as an input variable for the model, or if you are suggesting that maybe for other landslides where you don't have such detailed deformation data, deformation time series derived from InSAR could be used to train a similar model.

The first part of this paragraph is generic: InSAR may be used to train to train a similar model. The second part is specific to Vögelsberg and Sentinel-1. This has been clarified in the text.

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

This sentence was a remnant of an old paragraph and has been removed.

> **Lines 522** For landcover changes as an input, won't you run into the same problem of temporal resolution as you found in the SAR data? And if your landcover product was derived from e.g. Sentinel-2, it could actually be worse because of cloud cover.

Indeed, the timing of fast changes can be difficult to capture by (optical) remote sensing. This reservation has been added in text.

> 528: What assumptions? Please restate.

The assumption that, given the low spatial resolution of the remote sensing data, a single pixel represents the conditions of the higher part of the catchment as well. This has been clarified in the text.

> **Lines 543-547** Here, with the EGMS product, it is based on Sentinel-1 data so you would only have a 12-day temporal resolution (Especially following the failure of Sentinel-1B), which would result in the same temporal resolution problem you discussed in Section 6.3.1

The temporal resolution between April 2016 and December 2021 was at the original 6-day interval. The current EGMS data set runs till the end of 2021, the same time as the failure of Sentinel-1B, and still has an approximately weekly frequency.

> 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.

Indeed, the figure is superfluous in relation to Figure 10 and has been removed.

- **Textual corrections**

We thank both reviewers for their detailed textual corrections. These textual comments and minor comments have been addressed without further discussion.