Review of the manuscript "Sensitivity analysis and calibration of a dynamic physically-based slope stability model"

submitted by Thomas Zieher et al.

This paper describes an interesting approach to analyze the sensitivity of, and to optimize the key input parameters needed for coupled hydraulic-slope stability modelling with the software TRIGRS. The contribution is generally well written, logically structured, and very nicely illustrated. The methods used are reproducible, and the results are described and discussed in some detail. In my opinion, this contribution is definitely worth to be published in NHESS. As usual, I have identified a number of issues which could be optimized. These issues are addressed in detail below. **All in all, I suggest moderate revisions**.

Even though the paper is well written in general, there are several minor mistakes of grammar and style. It would be out of scope to address these shortcomings in detail, therefore I recommend careful copy editing. In the following, I focus on issues concerning the scientific content of the manuscript. Where numbers are given they refer to the manuscript page, line.

In 8, 20 it is mentioned that each pixel with FOS<1 is considered a single shallow landslide. In 16, 17f you mention that a landslide is considered predicted correctly if at least one pixel with FoS<1 coincides spatially with an observed landslide release area. This seems somewhat inconsistent to me and leads to two questions that have to be clarified:

(1) Do you perform the validation on the basis of correctly/incorrectly predicted landslide release polygons or on the basis of correctly/incorrectly predicted landslide release pixels?
(2) If the first possibility applies, how do you get your true negatives and false positives?

I really like that way of regular sampling of parameter combinations, which is a very efficient method of parameter optimization. However, as I understand it, each AUC value is derived from one single computation (i.e. from one point in the ROC diagram). Even though this is not wrong in principle, the idea of ROC is rather to consider curves instead of single points. There are more appropriate performance indicators than the AUC for single values, for example the CSI, HSS, D2PC, or FoC (see., e.g., Formetta et al., 2016; de Lima Neves Seefelder et al., 2016; Mergili et al., 2017). Please either clarify why you use the AUC, or use other performance indicators instead. For assessing the performance of the model ensemble (with 25 values; Fig. 11ca and f), AUC is perfectly suitable.

Looking at Table 5, the maximum AUC is lower with the best 25 runs than with all runs. This means that the best AUC value is associated with a parameter combination not satisfying the other criteria. In general, the improvement of AUC with a more constrained set of parameter combinations is very minor. This shows two issues:

(1) The AUC might be inappropriate, as mentioned above.
(2) More importantly, the results seem to confirm the findings of de Lima Neves Seefelder et al. (2016) that model performance in terms of AUC (or similar measures) may react quite insensitive to the variation of the input parameters. In this specific case, the other criteria (those leading to the constrained set of 25 parameter combinations) appear much more important to me. This is something that should be addressed adequately in the discussion.

The labelling of Fig. 7 is unclear to me: how can particular values of FOS be associated to a position along the ROC curve? E.g., with FOS=2, there are no true positives and 100% true negatives. Please explain or redraw the figure.

Fig. 8b looks like that the polygons are not drawn in a clean way. Figs. 9 and 12 are very well designed and informative. I also like the concept of the Figs. 11d and 11e.

I hope that my comments will help to further improve the quality of the manuscript.

Suggested references:

de Lima Neves Seefelder, C., Koide, S., Mergili, M. (2016). Does parameterization influence the performance of slope stability model results? A case study in Rio de Janeiro, Brazil. Landslides. doi:10.1007/s10346-016-0783-6

Formetta, G., Capparelli, G., Versace, P. (2016). Evaluating performance of simplified physically based models for shallow landslide susceptibility. Hydrology and Earth System Sciences, 20(11): 4585-4603. doi:10.5194/hess-20-4585-2016

Mergili, M., Fischer, J.-T., Krenn, J., Pudasaini, S.P. (2017): r.avaflow v1, an advanced open source computational framework for the propagation and interaction of two-phase mass flows. Geoscientific Model Development 10: 553-569. doi:10.5194/gmd-10-553-2017