Nat. Hazards Earth Syst. Sci. Discuss.,
doi:10.5194/nhess-2017-73-AC1, 2017

Natural Hazards and Earth System Sciences
Discussions

Interactive comment on "Sensitivity analysis and calibration of a dynamic physically-based slope stability model" *by* Thomas Zieher et al.

Thomas Zieher et al.

thomas.zieher@uibk.ac.at

Received and published: 8 May 2017

We thank the referee for the positive evaluation of our manuscript and the provided feedback. Please find our responses below, with referee comments in italics, and the authors' responses in blue.

General comments

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.

The manuscript was proofread by a native speaker of the research field (see supplement for all changes made).

Specific comments

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?

In the validation strategy, rasterized scar area polygons are considered as observed landslide release areas. However, every landslide can be predicted only once. This strategy was chosen because of the discrepancy of the regular raster environment (input and output maps) and the mapped shallow landslide scar area polygons. The spatial resolution of 10 m results from a compromise between the size of most shallow

landslide scar areas, the constraints of the infinite slope stability model and the representation of the topography (which should be detailed enough). However, it remains unknown which pixel represents an actually observed shallow landslide (see example in Fig. 1 at the end of the document). This results from positional uncertainties of the involved data sets, but also from the smoothed representation of the topography associated with the coarse raster resolution. It is therefore assumed, that the pixel with the lowest FOS intersecting the scar area polygon represents the respective landslide (e.g., Montgomery and Dietrich 1995, Casadei et al. 2003, Keijsers et al. 2011). If the lowest FOS value falls below 1.0, the landslide is considered predicted correctly. The remaining scar area pixels are omitted from the validation (neither as false negatives nor as true positives). True negatives are all pixels outside the rasterized scar areas with a $FOS >= 1.0$. False positives are pixels outside the rasterized scar areas with a $FOS < 1.0$.

Additional information on the validation strategy is now provided in Section 3.5.

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.

This comment refers to the original Table 5, page 19. The AUC values in this table are given for the whole curves as stated on page 17, line 30. Generally, the AUC values are not used for the calibration procedure. Instead, the sum of the true positive rate and

true negative rate was optimized (maximized). So, strictly speaking none of the above mentioned performance indicators were applied. In the present case study, minimizing the distance to the perfect classification (D2PC) instead of maximizing the sum of the true positive rate and true negative rate leads to the exact same results (although, considering the theoretical background, the results could differ). However, we think that the D2PC indeed may be a more general and more reliable performance indicator for the parameter optimization. The manuscript was adopted accordingly and in the revised version and the ranges of the D2PC were added in Table 6, Table 7 and Table 8. In the tables of the revised paper, the resulting AUC values are still presented to provide insight into its ranges. As requested in the following comment, the applicability of the AUC as performance indicator is now discussed in Section 5.

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.

We agree that the AUC is of limited value for validating physically-based slope stability models. For the tested parameter value ranges, the resulting AUC is definitively less sensitive than the position of the FOS falling below 1.0. This is now addressed 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.

The receiver operating characteristic (ROC) principle allows evaluating the performance of a binary classifier while varying its discrimination threshold. In case of FOS-maps it's the FOS which is varied (shown in the new Fig. 7; see Fig. 2 at the end of the document). Therefore each unique value of the FOS can be associated with a position along the ROC curve, along with the respective predictive rates. However, only the position of $FOS = 0.\dot{9}$ is relevant, since this value differentiates predicted landslides from stable slopes.

Changes in the manuscript: Figure 7 was replaced by Fig. 2 (at the end of the document).

Fig. 8b looks like that the polygons are not drawn in a clean way.

The figure was revised.

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

References cited in the answers:

Casadei, M., Dietrich, W. E. & Miller, N. L. 2003: Testing a model for predicting the timing and location of shallow landslide initiation in soil-mantled landscapes, Earth Surface Processes and Landforms 28(9), 925-950.

Keijsers, J. G. S., Schoorl, J. M., Chang, K. T., Chiang, S. H., Claessens, L. & Veldkamp, A. 2011: Calibration and resolution effects on model performance for predicting shallow landslide locations in Taiwan, Geomorphology 133(3-4), 168-177.

Metz, C. E. 1978: Basic principles of ROC analysis, Seminars in Nuclear Medicine 8(4), 283-298.

Montgomery, D. R. & Dietrich, W. E. 1994: A physically-based model for the topographic control on shallow landsliding, Water Resources Research 30(4), 1153-1171.

Please also note the supplement to this comment:
http://www.nat-hazards-earth-syst-sci-discuss.net/nhess-2017-73/nhess-2017-73-AC1-supplement.pdf

---

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., doi:10.5194/nhess-2017-73, 2017.

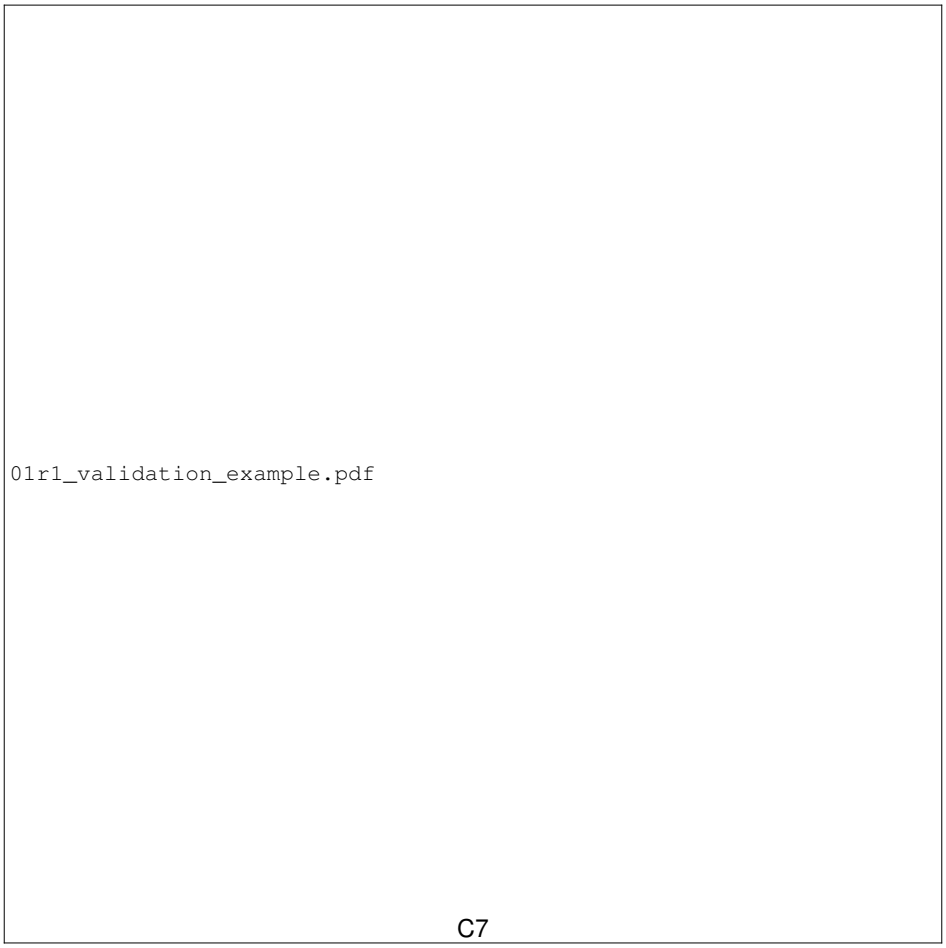

Fig. 1. Example for the validation strategy.

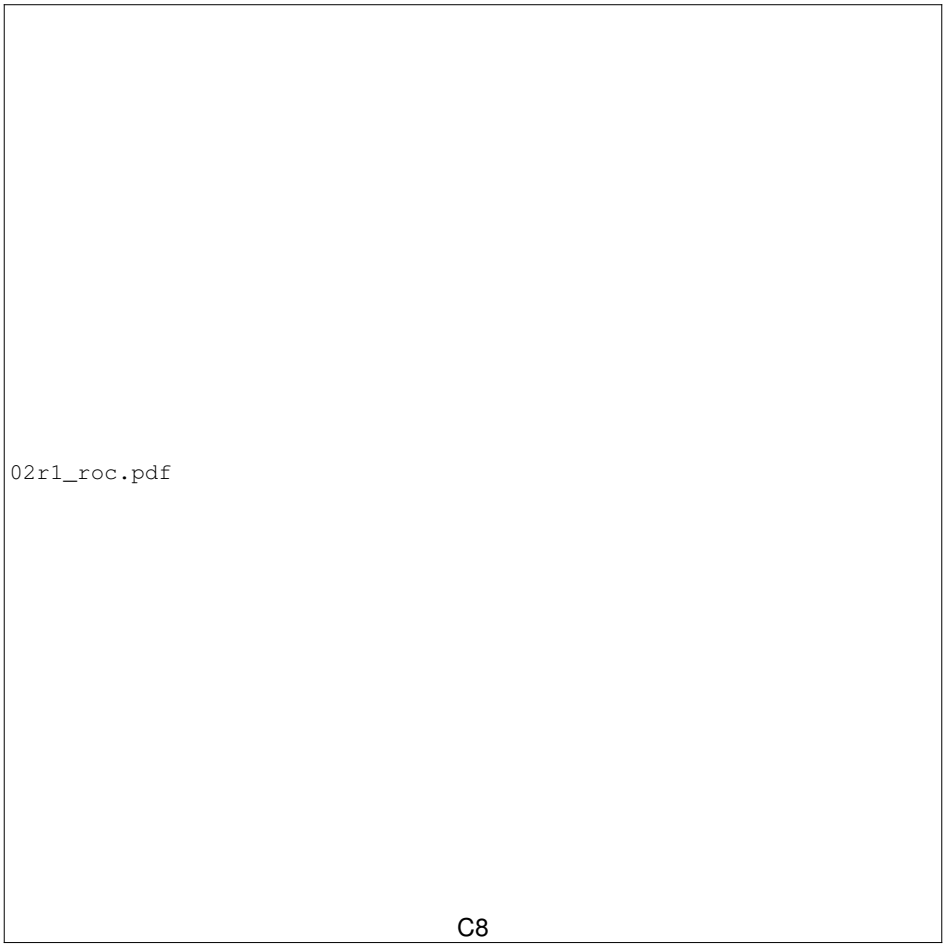

Fig. 2. Principle of the receiver operating characteristic (ex modified after Metz, 1978) and its

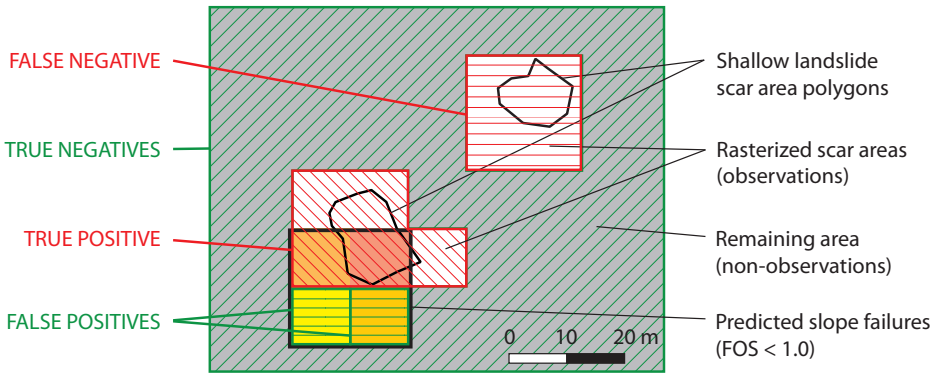

Fig. 3. Example for the validation strategy

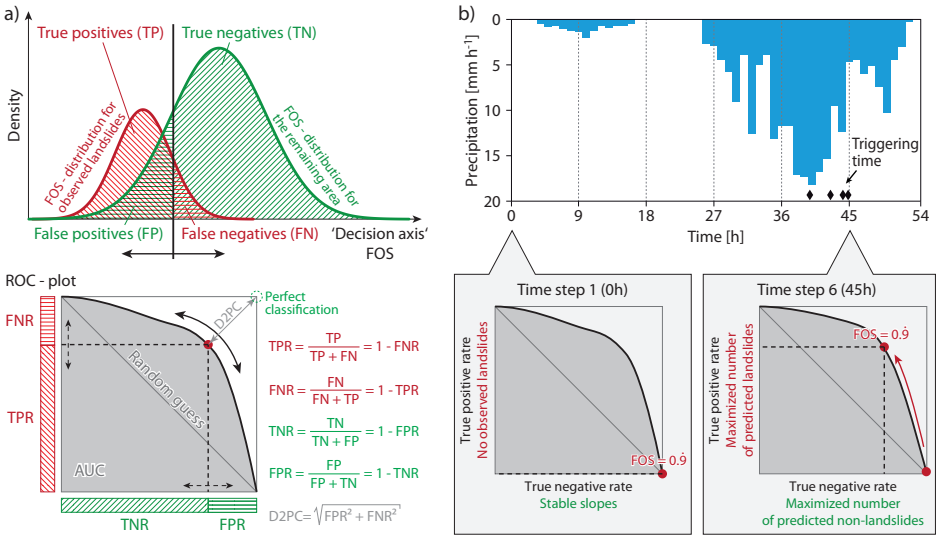

Fig. 4. Principle of the receiver operating characteristic (a; modified after Metz, 1978) and its application in the calibration procedure (b)