Reply to referees

On behalf of all co-authors I would like to thank the referees for spending more time in this extensive revision of our manuscript. In the following text, you could find the point-by-point responses to every individual comment, in which you will find your comments in regular style and our responses in italic.

Anonymous referee #1

Dear Authors,

Thanks for answering with such precision the reviewers report and having propose a new version of the article.

Thank you for spending more time reading our replies and the revised version.

General comments

In response to the reviewers concerns, you successfully made clearer, in the introduction, that the objective was in fact to identify the most relevant type of field monitoring when dealing with landslide in a sloping environment, and that the models (1D Richards and rainfall generation) were only supporting tools used to enlarge the dataset on which you applied your method (K-mean clustering + random forest).

Thank you for appreciating the changes to the manuscript, that were motivated by your comments.

That being said, some remarks initiated by the reading of the first version remain, especially about the methodology.

Indeed, if the « aim of this study is to identify the major hydrological processes controlling the response of the slope soil mantle […] through suitable measurable variables » (l.99-102), how can you justify the augmentation of your datapool with such a large synthetic dataset (1000 years)? Is it consistent with the practical and operational purpose of the method?

As you mention, lack of data is a very common, not to say a generalized concern: « However, a complete field monitored dataset is not always possible to be analyzed and, when it exists, it is commonly available for short periods, granting a relatively low measurement density » (l.269-272). If data are almost never available in a sufficient amount, the use of a model for synthetic generation is not only supporting your method, but becomes inherent to it.

Therefore, you should explain more precisely the requirements of your methods in terms of data density, and maybe that would shed some light on your choice of generating so much synthetic data. Is 1000 years of data a requirement to ensure statistical consistency? Would a 20 years chronicle (a somehow more realistic perspective) with reinforced data in terms of resolution (ensuring a hourly time step thanks to simulations when needed) be enough? This point is crucial to position your proposed method in an operational and realistic scope.

Thank you for raising this interesting issue. Indeed, long-lasting continuous field monitoring records at hillslopes are rare, and especially as regards critical conditions, such as landslides, they always contain too few data to allow significant statistical analyses. In our case, the clustering analysis of the soil mantle response shows that what we define "effective drainage conditions" (cluster 4 in Figures 10 and 11), quite interesting for understanding the physical processes active in the slope, consist of about 2% of the dataset. Hence, a dataset of only 20 years (containing about 1000 rainfall events) would hardly allow identifying this slope condition. In this respect, increasing the temporal resolution of the dataset would not provide useful information. In fact, the antecedent conditions controlling the response of the soil mantle are linked to

relatively slow processes (i.e., water accumulation in/drainage from the uppermost meter of the soil mantle, and variations of the water table of the shallow aquifer), which would be captured also with daily resolution. You have been probably misled by our unfortunate word choice « low measurement density », which should be much better written as « small number of measurements » (we have indeed rephrased it in the revised manuscript: lines 274-275). By the way, again referring to the conditions corresponding to cluster 4 (very high groundwater level and uppermost part of the soil mantle wetter than field capacity), they occur after an exceptionally long period of continuous rainfall, which would be unlikely observed in a 20-years long rainfall record. Of course, this does not imply that a dataset of 1000 years at hourly resolution is strictly necessary for our analysis, but ML techniques easily handle big amounts of data, and the simulation of 1000 years with the 1D Richards' equation model had to be carried out only once. Hence, we went for 1000 years, but some hundreds would likely have been enough. To quantify this aspect, we have repeated the ML analyses for shorter durations of the synthetic record.

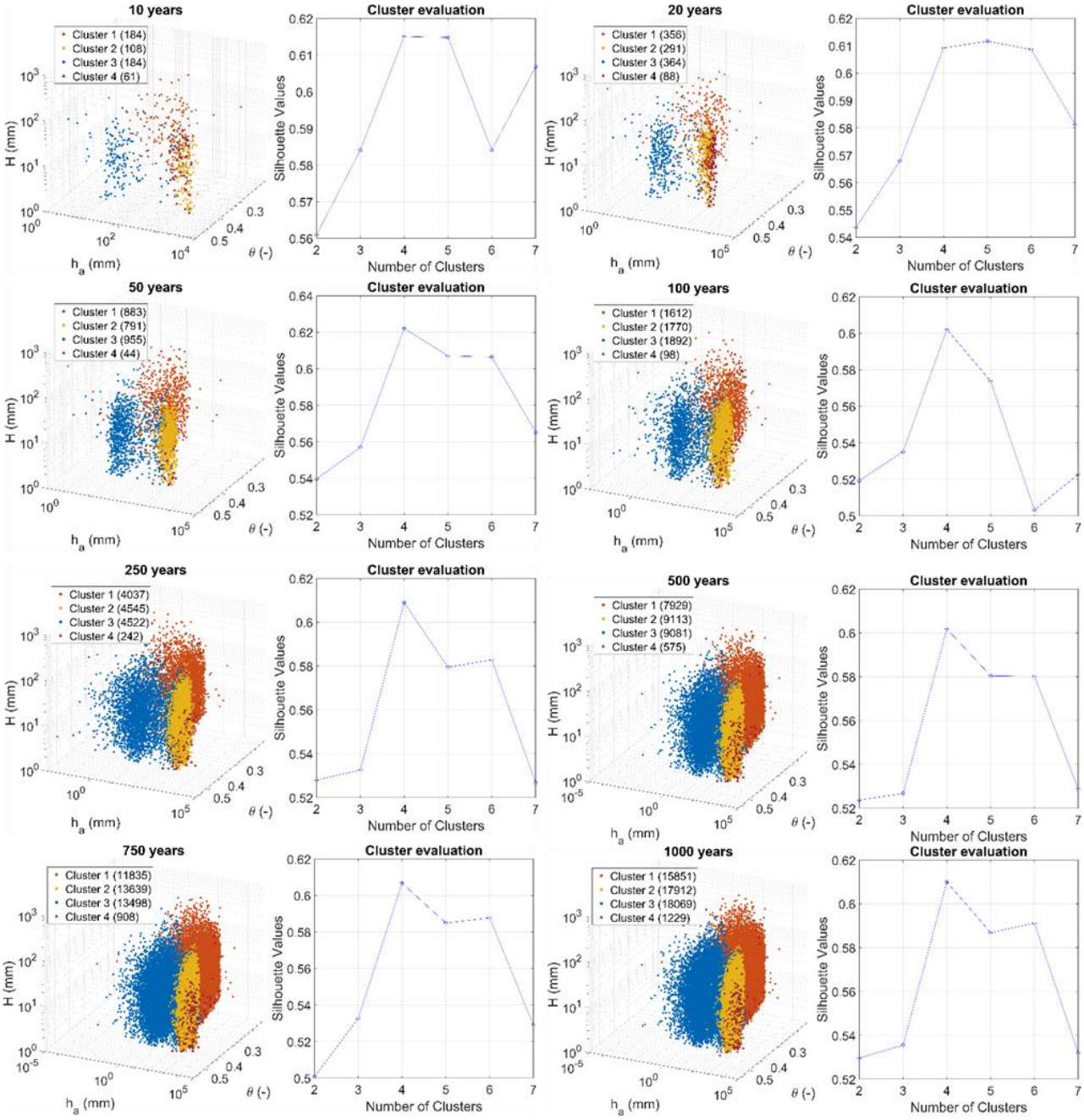

Figure R1-1. Clustering results of the synthetic data triplets $(\theta_{100}, h_a, \Delta S/H)$ represented in the space (θ_{100}, h_a, H) for various durations of the synthetic dataset. The right panels show the silhouette score, indicating the optimal number of clusters.

The above graphs show the spatial arrangements of the data in the space (θ_{100}, h_a, H) for different record durations (i.e., from 10 years to 1000 years), and the corresponding graphs with the values of the silhouette score for different choices of the number of clusters, defined in the space ($\theta_{100}, h_a, \Delta S/H$). Interestingly, the silhouette graphs tend to the same shape for long durations, and at least 50 years are required to obtain that the optimal number of clusters should be four, as for the 1000-year long record. The percentages of the elements belonging to each cluster also become stable for series durations longer than 100 years (e.g., the size of the fourth cluster -very wet conditions- varies between 1.8% and 2.3% of the total number of data considering durations between 100 years and 1000 years; the size of the first cluster -dry conditions- becomes stable around 30% of the data for durations longer than 100 years).

Looking at the results of the RF analysis, aiming at identifying the role of the three tested variables (i.e., θ_{100}, h_a and H) on the possible prediction of the response of the soil mantle to precipitation, the following graphs show how the estimated importance features of the three variables change, when the RF model is trained with different synthetic record durations (between 20 and 1000 years). For each considered duration, the RF model has been trained ten times with the cross-validation technique described in appendix B. The randomness of the choice of the training and validation sets implies that the estimated importance features are different for each trained RF model. However, if the dataset contains enough information, the estimated importance features of the variables should show only small changes. The two graphs show the changes of the mean value and of the standard deviation of the importance features, estimated from the 10 trained RF models, with different durations of the synthetic dataset. It looks clear that, for record durations longer than about 200 years, the estimated means become stable, and the standard deviations become small.

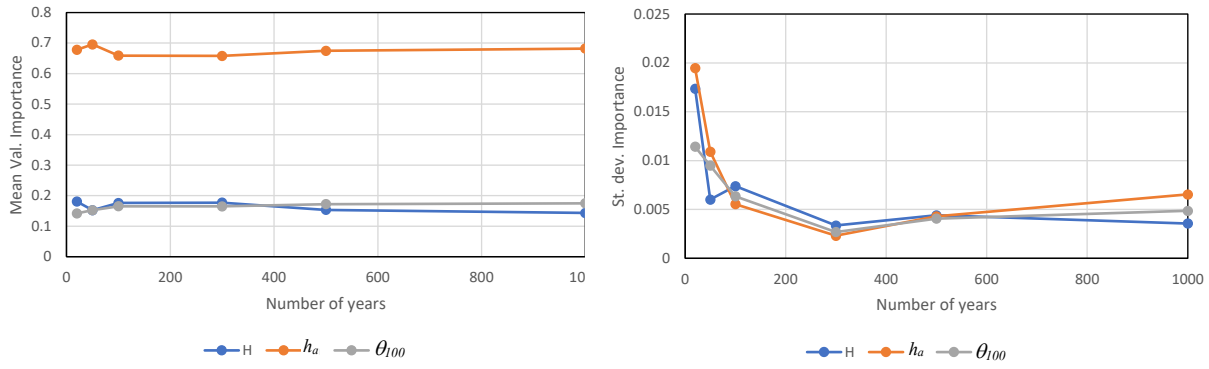

Figure R1-2. Random Forest feature importance for various durations of the synthetic data set: mean value (left panel); standard deviation (right panel).

These results seem to indicate that, in the studied case, a record of about 100-200 years could be long enough to convey the information needed to identify the major hydrological controls of the response of the soil mantle. In the revised manuscript, we have added a short sentence to inform the reader about the reasons for the choice of 1000 years (lines 280-285: "The choice of such a long synthetic series has been made to obtain an amount of data, representative also of conditions rarely occurring at the slope, large enough to ensure significance of the analyses carried out with ML techniques. In this respect, it is worth noting that the adopted clustering and Random Forest techniques allow easily handling big amounts of data without unaffordable computational burden.").

Stepping a bit further in this intricate issue between need of data and use of models, we come back to the topic of the information already carried by the models: through conceptualization, calibration and sensitivity analysis, a model (especially a physically-based one) informs about the relationship between your input (rainfall, initial or antecedent conditions) and your output (here variation in water storage).

In your case, by applying your method on such a large synthetic dataset (that with no doubt utterly occults the field data), more that directly analysing the relationship between actual rainfalls, antecedent conditions and the

underground response (pore pressure in the soil, groundwater level etc.), you statistically sort the relationship between your model input and output.

We are not sure that we fully understand the Reviewer's point here. The following figure shows the (few) experimental triplets, clustered in the same way as we have made with the synthetic data.

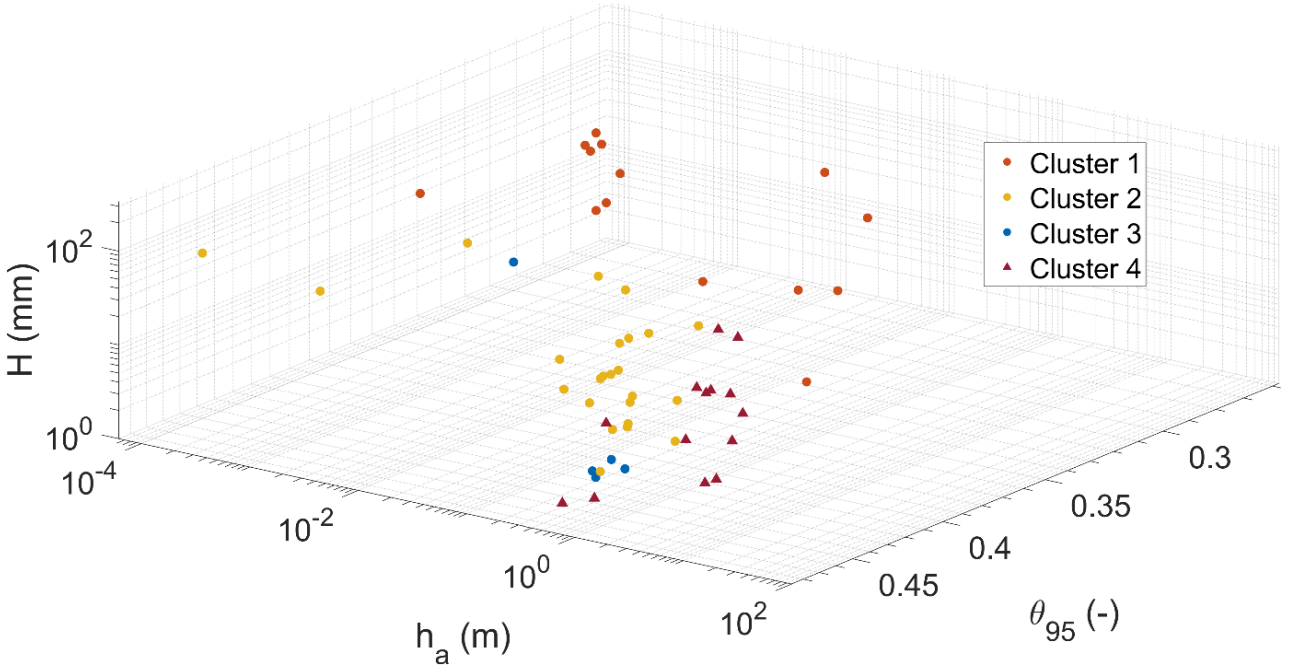

Figure R1-3. Clustering results of the measured data triplets ($\theta_{100}, h_a, \Delta S/H$) represented in the space (θ_{100}, h_a, H) for various durations of the synthetic dataset.

The results have been plotted in the space (θ_{100}, h_a, H), by replacing the available data of h_s (i.e., stream water level) with the estimated corresponding aquifer level h_a (indeed, we have only recently installed two piezometers in the field, but we do not have available aquifer level data so far). To do this, we used the conceptual relationships, used in the physically based model, to link the aquifer level to the estimated stream discharge and, in turn, to the stream water level. This is only a rescaling of the h_s axis, which does not significantly affect the spatial arrangement of the experimental dots. The total number of field experimental data is too small to draw quantitative interpretations, but the obtained four clusters resemble, to a reasonable extent, those of the synthetic dataset.

Looking at what was the response of the soil mantle, in terms of the distribution of $\Delta S/H$ in the four clusters, plotted as boxplots in the following figure, it looks clear that the four clusters, identified from the experimental data, correspond to similar responses as those from the synthetic dataset (plotted in Figure 12 of the paper, here reproduced for your convenience): impeded drainage for cluster 3; strong drainage for cluster 4; similar response for clusters 1 and 2, with "slightly smaller $\Delta S/H$ for cluster 1" (this is what we wrote at lines 624-625 of the previous version of the manuscript, commenting the synthetic responses).

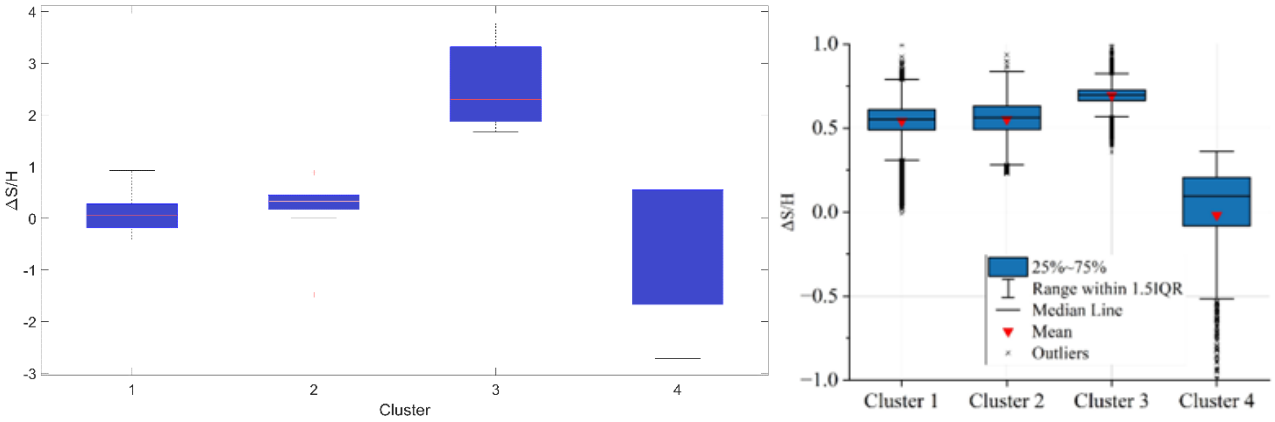

Figure R1-4. Distribution of the slope response ΔS/H for the data in each cluster: field data (left panel); synthetic data (right panel).

*Hence, we don't understand why the Reviewer claims that synthetic dataset, being too numerous, "*with no doubt utterly occults the field data*". Maybe, here we misunderstand something of the Reviewer's comment, but we believe that the behavior of the synthetic data is a good reproduction of the reality, which does not seem to introduce any biased sorting of the data.*

Therefore, sorry to insist, but I feel obligated to advise a sensitivity analysis. In particular, Sobol indices can be used to untangle the relative influence of a parameter (or initial condition in your case) or their combined influence on a variable of interest, which is specifically what you also aimed to do with machine learning techniques. The advantages of the sensitivity analysis are that it doesn't require much data, you can set the range of values over which you want to explore the sensitivity of the model as you wish (focusing on specific parameters or input), and discriminate different sets to mimic seasonality (benefiting from the statistical clustering analysis, but that can also be set based on less profuse climatic data). The drawbacks are a quite demanding number of runs (with a 1D model, it shouldn't be an issue) and a complete subjugation to the model, which was not the stated as the initial scope of your paper, I agree.

As you suggest, we have carried out a variance-based sensitivity analysis. Before commenting the obtained results, we would like to stress that there are some other drawbacks of the SA, that you do not mention in your comment.

First, differently from when a SA is carried out to evaluate the influence of (the uncertainty of) model parameters on the output, in this case the aim would be to identify the effects of different antecedent conditions on the response of the soil mantle of the slope to precipitation inputs. However, the variables that we use in our ML analysis cannot be either the actual antecedent (initial) condition, nor the actual precipitation input. In fact, for the sake of simplicity (and also to stick to what can be easily measured/handled for practical operational purposes), as antecedent conditions we have considered the mean water content of the uppermost meter of the soil mantle (the initial condition is instead the water content profile throughout the whole mantle) and the water level in the underlying perched aquifer (this latter is indeed an initial condition, but in our model it is physically (and mathematically) linked to the soil moisture at the base of the soil mantle (please, refer to Greco et al. (2018) for details about the equations), so it affects also the initial soil moisture profile).

Hence, to carry out the SA, we have been obliged to simplify the initial soil moisture profile in the bi-linear format shown in the following sketch (we already used a similar format in Marino et al. (2021), where you can see to what extent it resembles, at least in many cases, the actual moisture profiles observed in the field):

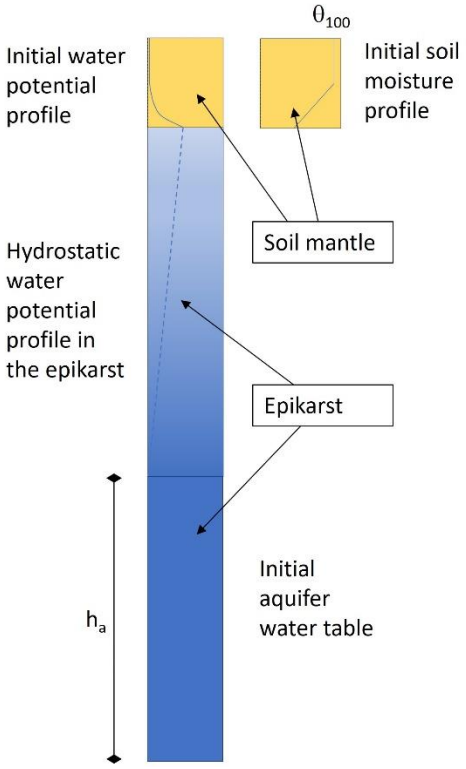

Figure R1-5. Simplified initial moisture profile in the soil mantle, with the corresponding water potential profile, and its relationship with the underlying perched aquifer water level.

The rainfall events that are the input of model simulations are also much more complex than the simple total rain depth that we used for our ML analysis of cause-effect relationships. In this case, we carried out the SA considering rainfall events with constant intensity (i.e., rectangular hyetographs), that could be characterized by means of total duration and depth.

Summarizing, the SA analysis cannot be easily carried out considering the actual initial and boundary conditions. In our opinion, this is a point that moves the needle in favor of the Random Forest analysis.

The second drawback is related to the generation of the set of combinations of input parameters, required to carry out the variance-based SA. In fact, to avoid introducing a bias in the estimated output variance, the set of generated input values should be consistent with their probability distributions, which cannot be simplistically assumed Normal (as it is usually made when the SA aims at quantifying the effects of model parameter uncertainty). In the case of θ_{100} and h_a, we have little information from our field data (about 50 couples, as can be seen in figure 5 of the manuscript, here reproduced for your convenience):

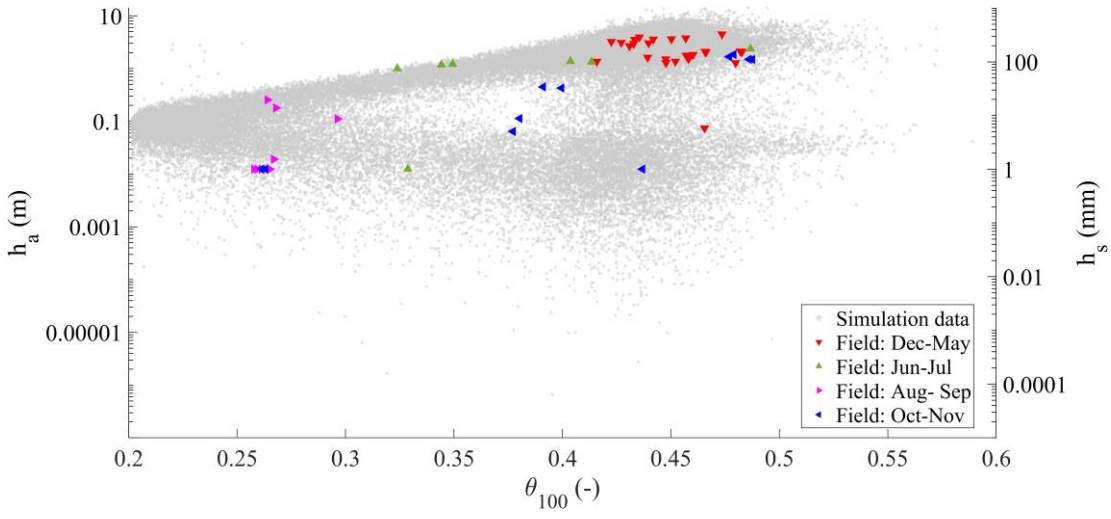

Figure R1-6. Field monitored mean volumetric water content in the upper meter of the soil profile (θ_{100}) and water depth in the Castello stream (h_s) compared with simulated data (the vertical axes are plotted in logarithmic scales to help visualizing small water levels).

Based on the small set of experimental data on antecedent conditions, it is possible to roughly visualize the frequency distributions of the observed values of θ_{100} and h_a, here compared with those of θ_{100} and h_s of the synthetic dataset (reproduced from Fig. 7 of the manuscript).

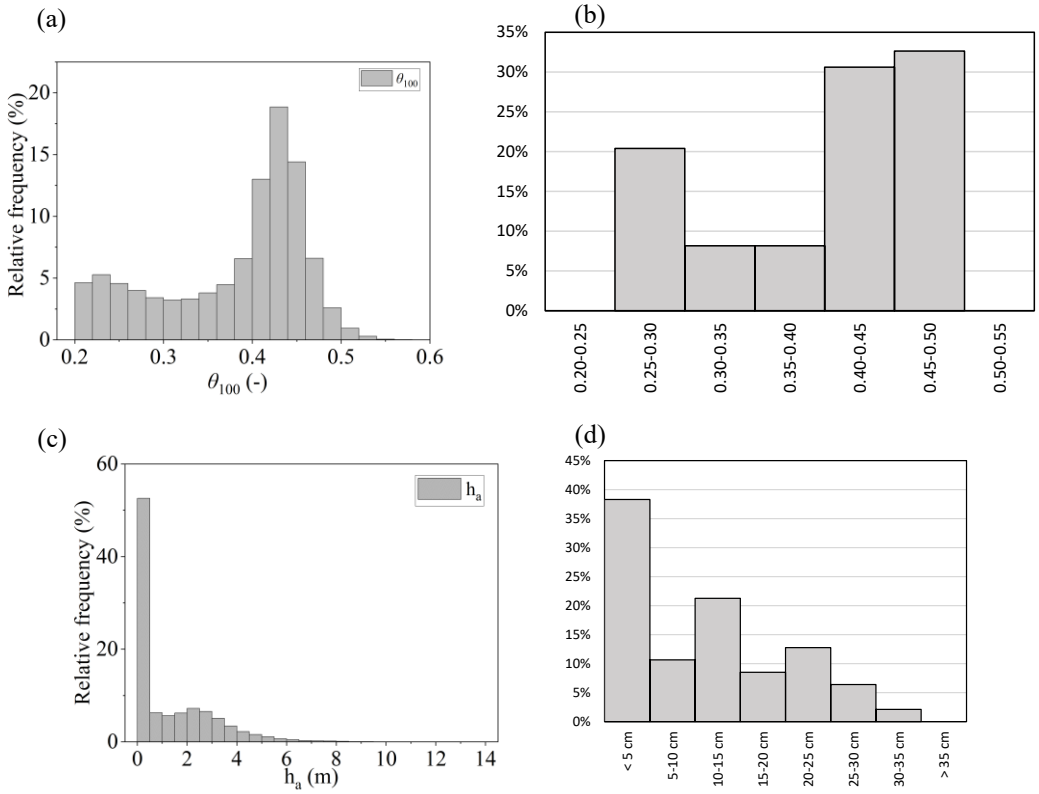

Figure R1-7. Frequency distributions of (a) synthetic and (b) observed initial moisture of the uppermost meter of the soil mantle, θ_{100}; frequency distribution of (c) synthetic antecedent water depth in the Castello stream, h_s, and (d) observed perched aquifer water level, h_a.

The few observed data (by the way, somehow confirming the shape of the distributions obtained with the model chain used for the generation of the synthetic dataset) indicate that the distributions of the antecedent values are neither Normal nor uniform.

About the precipitation input, the available record of 17 years, which had been used for the calibration of the NRSP model (as described in Appendix A), contains enough information about the distributions of the values of rainfall event duration and depth (see figure A4 of appendix A, here reproduced for your convenience).

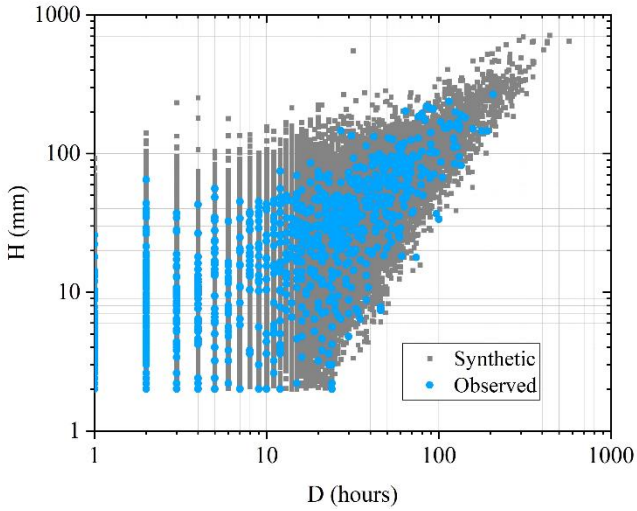

Figure R1-8. Scatterplot of total rainfall event depth (H) vs. rainfall event duration (D). The events have been sorted within the rainfall datasets by considering a separation "dry" interval of 24 hours with less than 2 mm rainfall. The blue dots represent events extracted from the 17 years experimental rainfall dataset, while the grey dots represent events extracted from the 1000 years synthetic rainfall dataset.

The third drawback, also regarding the generation of the set of combinations of the variables (θ_{100}, h_a, H)*, is related to the issue of the existence of correlation between the input variables. For instance, event rainfall depth* H *and duration* D *clearly exhibit significant correlation (see figure R1-8), but also* θ_{100} *and* h_a*, both related to previous precipitation history, may show some degree of correlation, which should be considered when the combination of input variables are generated, in order to carry out rigorously the variance-based sensitivity analysis.*

All these considerations have been made here just to underline that carrying out a SA with the purpose of characterizing the input-output relationships in a system is not an easy task, which implies assumptions and issues which make the interpretation of the results far from being obvious.

This said, the SA has been carried out, based on the methodology outlined by Sobol (2001), which is implemented in the Sensitivity Analysis Library in Python - SALib toolbox (Herman and Usher, 2017; Iwanaga et al., 2022). To reduce the error rates in the calculation of the sensitivity index, the Saltelli's sampling scheme is used to generate the uniform sample set (Saltelli, 2002; Saltelli et al., 2010). Specifically, $N \times (2P + 2) = 65536$ *triplets of* (θ_{100}, h_a, H) *have been generated, with* N*=8192 (a power of 2) and* P*=3 the number of inputs, and each time the physically based model has been run to obtain the corresponding soil mantle response* $\Delta S/H$*. For each sampled value of* H*, the corresponding event duration* D *has been obtained by means of a power-law fitting relationship of the experimental rain events plotted in figure R1-8 (*$D = 0.282 \times H^{1.525}$*). The calculated sensitivity indices are given in the following table.*

Variable	S_{tot}	S_1 *(single parameter variations)*	S_2 *(mutual interactions)*	
θ_{100}	*0.532*	*0.471*	(θ_{100}, h_a)	*0.002*
h_a	*0.058*	*0.058*	(θ_{100}, H)	*0.060*
H	*0.469*	*0.412*	(h_a, H)	*0.000*

The obtained results shed more light on the different meaning of a variance-based SA, compared to the importance feature estimated with the RF model. The SA explains how the variability of the output (here $\Delta S/H$*) is related to the variability of the inputs, and it looks clear that the variability of* θ_{100} *and* H *strongly (almost completely) affects the variability of* $\Delta S/H$*. Differently, the RF analysis aimed at estimating what are the most informative variables, useful to make good predictions of the output. While the SA only looks at the variation of the output, the RF also looks at how well the output is predicted, compared to "real" observations (in this case, synthetic). We have added some paragraphs in section 2.3.1 to highlight the difference between RF*

importance analysis and SA (lines 435-438; lines 443-452; lines 468-474), and the comparison of the results of SA with RF (lines 546-567, and Table 3).

In this respect, while it is physically obvious that, the more variable are the rain and the initial wetness of the soil (i.e., H and θ_{100}), the more variable is the stored precipitation $\Delta S/H$ (i.e., more rain, more storage; dryer soil, more storage possible), the results of the RF analysis are less trivial. The aquifer water level h_a, not affecting so much the variability of $\Delta S/H$, as indicated by the SA, is anyway an extremely informative variable, as it allows separating the initial conditions in two families: low levels (clusters 1 and 3), high levels (clusters 2 and 4). In conjunction with the information about rainfall amount and initial soil moisture, this separation strongly improves the capability of predicting the response of the soil mantle to precipitation in terms of $\Delta S/H$ (e.g., the different response of cluster 2 and cluster 3, that share the same values of antecedent θ_{100}, but exhibit a quite different attitude to retain infiltrating rainwater).

To summarize and conclude, please make clear the reasons that made you create a such long period of synthetic data. If it was not essential to the method, note that it is somehow contradictory with your operational purpose. If it was indeed essential, you cannot rule out both the bias involved in using models to feed your dataset, and the opportunity to compare your own method with a sensitivity analysis.

We believe that the results and considerations, developed in response to the issues raised by this Referee, have satisfactorily clarified the doubtful points, and have allowed improving the quality of the manuscript.

Specific comments

Figure 1: Thanks for having taken my advice.

Indeed, it was a good advice, and the Figure has been improved.

L.375-378: « It is important to note that […] would lead to difficult application of the model at less detailed scales such as regional and catchment scales ». If the model is only supporting your method in this particular case, the issue of using it to larger scale wouldn't be an issue worth mentioning in the scope of this paper. Here again, some ambiguity between the objective and the method remains.

You are right, the paragraph between lines 376 and 381 is truly unclear. The model aims at representing a geomorphological setting, air-fall pyroclastic deposits overlying calcareous bedrock resting on steeply inclined slopes, which is found in large areas of Campania (southern Italy), and not only in the Partenio Massif, where the studied slopes belong to. The following figure, adapted from the cited paper by Cascini et al. (2008), sketches the areas where slopes share such characteristics.

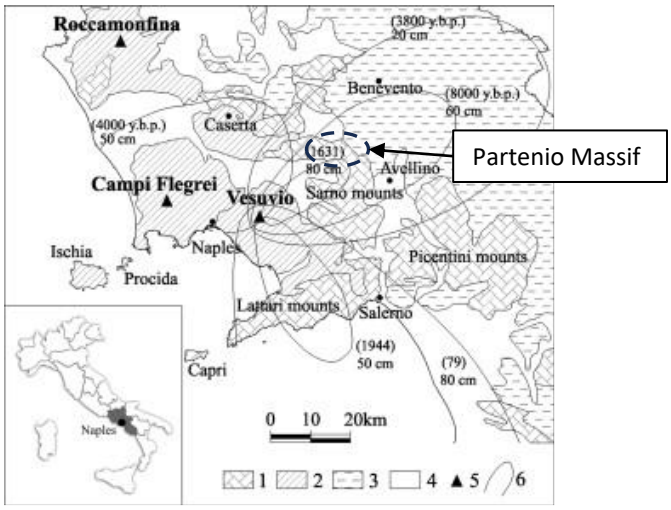

Figure R1-5 (adapted from Cascini et al., 2008). Air-fall pyroclastic deposits in the Campania region: 1) carbonate bedrock; 2) tuff and lava deposits; 3) flysch and terrigenous bedrock; 4) alluvial and continental deposits; 5) volcanic complexes; 6) isopachous lines of the pyroclastic products from the main eruptions (in brackets eruption data).

The slopes of Sarno mounts, Picentini mounts, Lattari mounts, as well as those of Partenio Massif, present pyroclastic mantles with similar physical and geometric characteristics, laying upon bedrock with similar geological features, covered with similar vegetation, and they share similar climate. Consequently, all the slopes of this whole large area (i.e., few thousands of square kilometers) are frequently subjected to rainfall-induced shallow landslides, triggered by similar rain events in similar antecedent conditions (e.g., Di Crescenzo and Santo, 2005; Cascini et al., 2008; Greco et al., 2021). We have modified the Introduction to better highlight that the study area is representative of a geomorphological context quite common in the region region (lines 74-81:" This research focuses on a case study of a slope located in Campania (southern Italy), representative of a wide area frequently hit by destructive rainfall-triggered shallow landslides (e.g., Fiorillo et al., 2001; Revellino et al., 2013). In fact, such geohazards are recurrent along the carbonate slopes covered with unsaturated air-fall pyroclastic deposits, diffuse over an area of few thousand square kilometres around the two major volcanic complexes of the region, the Somma-Vesuvius and the Phlaegrean Fields (Di Crescenzo and Santo, 2005; Cascini et al., 2008)"), and that the aim of the study refers to slopes of the area with characteristics similar as the modelled one (lines 102-105).

Obviously, every slope in the area has its own specific features, but the hydrological processes controlling the response of the soil mantle to precipitation events can be considered similar over wide areas, as they are related to large scale (in time and space) processes such as long-term cumulated rainfall and evapotranspiration, and perched aquifer recharge. In this sense, we believe that our simplified 1D slope model (with a homogeneous soil mantle of constant thickness, constant slope inclination, and homogeneous epikarst) could be useful to approximately assess the antecedent conditions that control the response of the soil mantle to precipitation in a wide area. This is the meaning of the fateful expression at lines 376-379 (and also at lines 379-381). To make them clearer, we have modified the paragraph, which now reads (lines 389-399): "The model assumes a homogeneous soil profile and a simplified slope geometry, and indeed it is not aimed at reproducing the details of local flow processes through the unsaturated soil mantle. Consequently, the hydraulic properties of the homogeneous soil layer should be considered as effective properties, useful to reproduce the major features of the infiltration and drainage phenomena. The model is rather used to assess how large-scale (in time and space) hydrological processes, such as long-term cumulated rainfall and evapotranspiration and perched aquifer recharge, control the conditions that affect the response of the soil mantle to precipitation events. In this sense, the obtained results can be considered representative for large areas that share the major geomorphological features of the slopes of Partenio Massif".

L.488-490: The computational effort of less than 2 minutes per run is concerning the RF procedure, am I right? How many runs did you end up simulate considering all the combinations of the variables? What is the duration of a 1D model run to compare (much less than 2 minutes I assume)? This may also argue in favor of a sensitivity analysis.

The calibration of the hyperparameters of the four tested RF models took in total less than one hour, leading to the choice of the most informative triplet of input variables (θ_{100}, h_a, H), as well as to the scores measuring the contribution of each variable. To have an idea of the computational effort of the ML procedure, this time should be summed to the time needed to carry out the simulation of 1000 years (about 8.76×10^6 hours). Such time is longer than the time required to carry out the 1D model simulations of the SA: the required 65536 rainfall events have variable durations, ranging between 1 hour and few hundreds of hours (see figure 4, above). Considering that the mean rain event duration is about 26.5 hours, the total simulated time results about 1.7×10^6 hours, indeed shorter than the 8.76×10^6 hours, corresponding to 1000 years. However, running 65536 separated simulations, instead of a single long model run, implies additional computational time, thus reducing the difference of computational burden (in fact, the simulations for the sensitivity analysis lasted about 10 hours, while the 1000 years single model run took about 28 hours on the same computer). So, although the SA allows saving some computational time, the gain does not look so significant to be the reason to guide the choice of the method of analysis.

Figure 5: Thanks for clarifying the signification of the scale. You should mention a reservation about simulated values (whether for groundwater or for river level) below the centimeter/millimeter scale, taking into account that you want to mimic a field monitoring (therefore including limit and uncertainty in the measurement, especially for low values).

Thank you for the suggestion. We have modified the caption of Figure 5, which now reads "Field monitored mean volumetric water content in the upper meter of the soil profile (θ_{100}) and water depth in the Castello stream (h_s), compared with synthetic data of θ_{100} and aquifer water level (h_a) (on the vertical axis, plotted in logarithmic scale to help visualizing small water levels and thus not allowing to represent zeros, the values of h_s smaller than the sensitivity of the water level sensor have been plotted as 1 mm; also the smallest simulated values of h_a should be considered equivalent to zero, owing to the limits of any measurement device which could be used for operational field monitoring)".

<u>References</u>

Cascini, L., Cuomo, S., and Guida, D.: Typical source areas of May 1998 flow-like mass movements in the Campania region, Southern Italy, Eng. Geol., 96, 107-125, https://doi.org/10.1016/j.enggeo.2007.10.003, 2008.

Di Crescenzo, G., and Santo, A.: Debris slides–rapid earth flows in the carbonate massifs of the Campania region (Southern Italy): morphological and morphometric data for evaluating triggering susceptibility, Geomorphology, 66(1-4), 255-276, https://doi.org/10.1016/j.geomorph.2004.09.015, 2005.

Greco, R., Comegna, L., Damiano, E., Marino, P., Olivares, L., and Santonastaso, G. F.: Recurrent rainfall-induced landslides on the slopes with pyroclastic cover of Partenio Mountains (Campania, Italy): Comparison of 1999 and 2019 events, Eng. Geol., 288, 106160, https://doi.org/10.1016/j.enggeo.2021.106160, 2021.

Greco, R., Marino, P., Santonastaso, G. F., and Damiano, E.: Interaction between Perched Epikarst Aquifer and Unsaturated Soil Cover in the Initiation of Shallow Landslides in Pyroclastic Soils, Water, 10, 948, https://doi.org/10.3390/w10070948, 2018.

Herman, J., and Usher, W.: SALib: an open-source Python library for Sensitivity Analysis. The Journal of Open Source Software, 2(9), 97, https://doi.org/10.21105/joss.00097, 2017.

Iwanaga, T., Usher, W., and Herman, J.: Toward SALib 2.0: Advancing the accessibility and interpretability of global sensitivity analyses. Socio-Environmental Systems Modelling, 4, 18155–18155, https://doi.org/10.18174/SESMO.18155, 2022.

Marino, P., Santonastaso, G. F., Fan, X., and Greco, R.: Prediction of shallow landslides in pyroclastic-covered slopes by coupled modeling of unsaturated and saturated groundwater flow, Landslides, https://doi.org/10.1007/s10346-020-01484-6, 2021.

Saltelli, A.: Making best use of model evaluations to compute sensitivity indices. Computer Physics Communications, 145(2), 280–297. https://doi.org/10.1016/S0010-4655(02)00280-1, 2002.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S.: Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Computer Physics Communications, 181(2), 259–270. https://doi.org/10.1016/J.CPC.2009.09.018, 2010.

Sobol, I. M.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation, 55(1–3), 271–280. https://doi.org/10.1016/S0378-4754(00)00270-6, 2001.

Anonymous referee #3

Thanks for the revised manuscript. The authors performed major revisions, addressing many of the raised concerns by the other reviewers and myself. The manuscript now clearly is easier to understand, and its objectives better defined, especially for the typical audience of HESS. The appendices explaining the rainfall generator and details of RF setup also help.

Thank you for acknowledging the improvement of the manuscript.

However, there is one central point that remains open from my side, concerning my questions about the use of a sensitivity analysis (or uncertainty estimation) of the 1D model, instead of the analysis of its results using RF (similar points also were raised by the other two reviewers):

It is a central assumption in your manuscript that this single realization of a 1D model is an adequate representation of reality to base your paper on an ML-based analysis of its outputs. This is so central to your paper, that I think it still deserves more attention.

I do understand the authors' argument that their study is not an evaluation of the sensitivity or uncertainty of the 1D model. And maybe I should have been more precise with my concern in the first review ("[…]You set up a single model (with a single parameter set) – ignoring the heterogeneity of soil thickness, hydraulic conductivites, etc. one would find along the slope? Given such a simple 1D model, I would at least recommend to perform some kind of sensitivity analysis / parameter uncertainty estimation / ensemble model run."). I try to reformulate: You use one single realization of a calibrated 1D model, with a synthetic rainfall dataset, to understand how seasonal conditions may affect triggering of landslides. Your central assumption here is that the model-based synthetic dataset is an adequate representation of reality. Despite using a calibrated model, I would still assume that there is some uncertainty or equifinality in the model? I.e. do you know how well defined are the decisive parameters? Could you not find various parameter combinations that give almost equally well performance of your 1D model? Potentially even quite different parameter combinations? (not even touching on questions of model structure…) Models with different parameters sets could also exhibit quite different physical behaviour/cause-effect relationships, despite similar performance in terms of calibration data. They should be considered an equally adequate basis for your subsequent RF analysis – and might yield in quite different final results and conclusions. This aspect I still am missing – I would really love to see this addressed (how well defined are the parameters of the 1D model? How much equifinality is there?), or at least it has to be discussed as a shortcoming.

What this Reviewer is asking about Sensitivity Analysis is something different from what is asked by the other Reviewer. In fact, here the Reviewer is concerned with the effects of the uncertainty of model parametrization on the cause-effect relationships that we have identified in the synthetic data, and thus he suggests going back to the calibration/validation of the 1D Richards' equation model coupled with the underlying perched aquifer hosted in the epikarst, which had been made long time ago (Greco et al., 2013). Most of the parameters of the model were assigned based either on literature indications or on available measured values. The calibration of the remaining parameters (i.e., those of unsaturated soil hydraulic characteristic functions) was carried out based on field measurements of soil water potential and moisture. The parameters, searched with a Genetic Algorithm, were constrained so to ensure the corresponding hydraulic functions to resemble available measurements of water retention and unsaturated hydraulic conductivity, obtained both in the field and in the laboratory (we have added this information at lines 369-373 in the revised manuscript). The figure below, reproduced from Greco et al. (2013), shows an example of the obtained agreement between simulations and measurements.

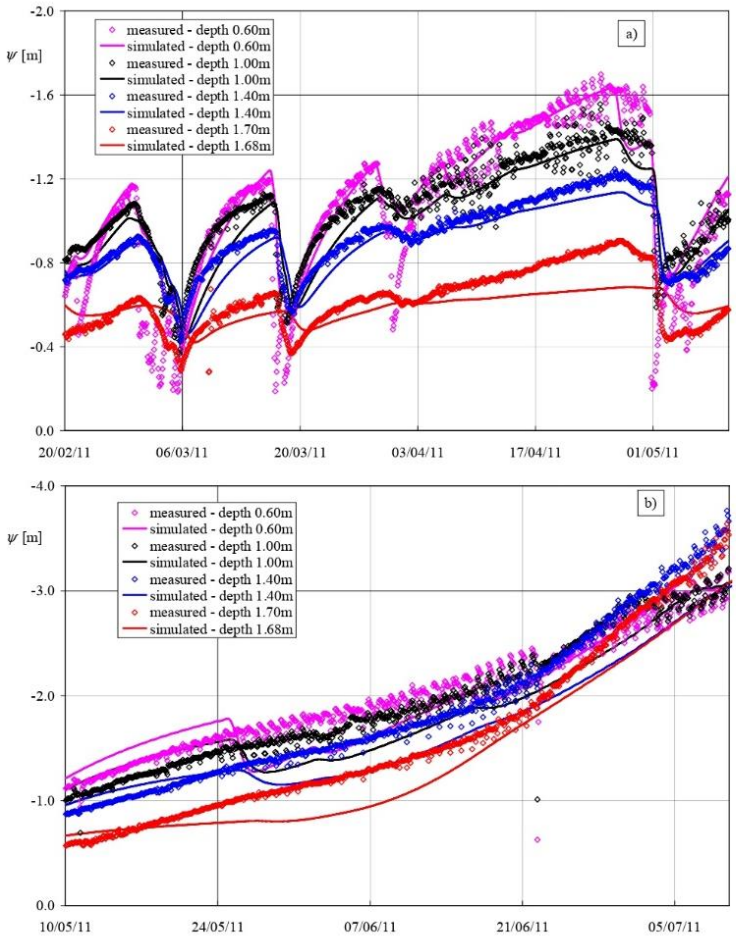

Figure R2-1 (adapted from Greco et al., 2013). Comparison between simulated and observed capillary pressure head at various depths during the period considered for model calibration: slope with leafless trees and low-developed underbrush (above); slope during the vegetation flourishing period (below).

A 2D version of the model, with the same structure about the coupling between unsaturated soil cover and perched aquifer, was then successfully applied to the assessment of the multi-year water balance of the slope of Cervinara (Greco et al., 2018). It also allowed reliably interpreting the response of the soil mantle during intense rainstorms, which might result or not in the triggering of landslides (Greco et al., 2021; Marino et al., 2021).

Re-evaluating the calibration of the model, carrying out a SA of the output to model parameter uncertainty, would imply introducing a completely different chapter in the manuscript, which is far beyond the scope of this work.

In fact, the simplified model (with a single homogeneous soil layer with effective parameters and a simplified slope geometry, with constant inclination and soil mantle thickness, and homogeneous epikarst) aims at representing a geomorphological setting, air-fall pyroclastic deposits overlying calcareous bedrock resting on steeply inclined slopes, which is found in large areas of Campania (southern Italy), and not only in the Partenio Massif, where the studied slopes belong to. The following figure, adapted from the cited paper by Cascini et al. (2008), sketches the areas where slopes share such characteristics.

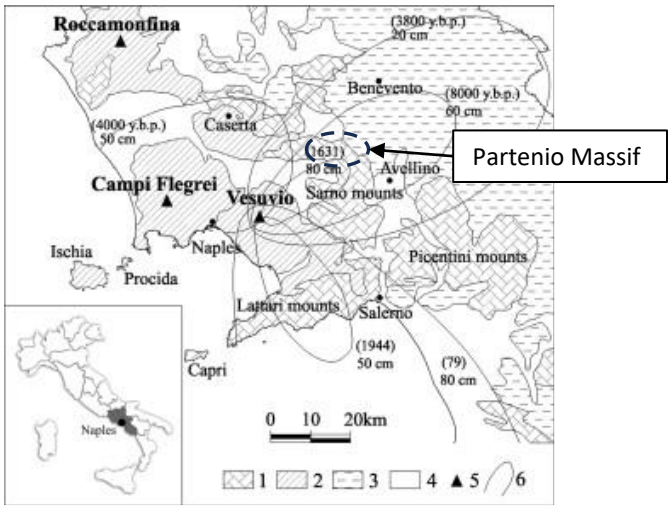

Figure R2-2 (adapted from Cascini et al., 2008). Air-fall pyroclastic deposits in the Campania region: 1) carbonate bedrock; 2) tuff and lava deposits; 3) flysch and terrigenous bedrock; 4) alluvial and continental deposits; 5) volcanic complexes; 6) isopachous lines of the pyroclastic products from the main eruptions (in brackets eruption data).

The slopes of Sarno mounts, Picentini mounts, Lattari mounts, as well as those of Partenio Massif, present pyroclastic mantles with similar physical and geometric characteristics, laying upon bedrock with similar geological features, covered with similar vegetation, and they share similar climate. Consequently, all the slopes of this whole large area (i.e., few thousands of square kilometers) are frequently subjected to rainfall-induced shallow landslides, triggered by similar rain events in similar antecedent conditions (e.g., Di Crescenzo and Santo, 2005; Cascini et al., 2008; Greco et al., 2021). We have modified the Introduction to better highlight that the study area is representative of a geomorphological context quite common in the region (lines 74-81:" This research focuses on a case study of a slope located in Campania (southern Italy), representative of a wide area frequently hit by destructive rainfall-triggered shallow landslides (e.g., Fiorillo et al., 2001; Revellino et al., 2013). In fact, such geohazards are recurrent along the carbonate slopes covered with unsaturated air-fall pyroclastic deposits, diffuse over an area of few thousand square kilometres around the two major volcanic complexes of the region, the Somma-Vesuvius and the Phlaegrean Fields (Di Crescenzo and Santo, 2005; Cascini et al., 2008)"), and that the aim of the study refers to slopes of the area with characteristics similar as the (simplified) modelled one (lines 102-105).

Obviously, every slope in the area has its own specific features, but the hydrological processes controlling the response of the soil mantle to precipitation events can be considered similar over wide areas, as they are related to large scale (in time and space) processes such as long-term cumulated rainfall and evapotranspiration, and perched aquifer recharge. In this sense, we believe that our simplified 1D slope model (with a homogeneous soil mantle of constant thickness, constant slope inclination, and homogeneous epikarst) could be useful to approximately assess the antecedent conditions that control the response of the soil mantle to precipitation in a wide area. This is the meaning of the (unclear, we agree) expressions at lines 376-381. To make them clearer, we have modified the paragraph, which now reads (lines 389-399): "The model assumes a homogeneous soil profile and a simplified slope geometry, and indeed it is not aimed at simulating the details of local flow processes through the unsaturated soil mantle. Consequently, the hydraulic properties of the homogeneous soil layer should be considered as effective properties, useful to reproduce the major features of the infiltration and drainage phenomena. The model is rather used to assess how large-scale (in time and space) hydrological processes, such as long-term cumulated rainfall and evapotranspiration, and perched aquifer recharge, control the conditions that affect the response of the soil mantle to precipitation events. In this sense, the obtained results can be considered representative for large areas that share the major geomorphological features of the slopes of Partenio Massif".

This said, we believe that it is not worth carrying out also a SA of the output to the uncertainty of model parameters. As suggested, we have added a "warning" about the limitations of the obtained results at the beginning of section 3 (Results and discussion) (lines 506-513: "The analysis of the physical behavior of the

studied slopes is based on the results of model simulations, as if they satisfactorily resemble what could be measured in the field. Indeed, the uncertainty of model parameters may affect the identified cause-effect relationships. However, during the calibration of model, field measurements of the hydraulic behavior of the involved soil were considered (Greco et al., 2013), thus the major features of the hydrological processes occurring in the slope are considered reliably reproduced in the synthetic dataset").

References

Cascini, L., Cuomo, S., and Guida, D.: Typical source areas of May 1998 flow-like mass movements in the Campania region, Southern Italy, Eng. Geol., 96, 107-125, https://doi.org/10.1016/j.enggeo.2007.10.003, 2008.

Di Crescenzo, G., and Santo, A.: Debris slides–rapid earth flows in the carbonate massifs of the Campania region (Southern Italy): morphological and morphometric data for evaluating triggering susceptibility, Geomorphology, 66(1-4), 255-276, https://doi.org/10.1016/j.geomorph.2004.09.015, 2005.

Greco, R., Comegna, L., Damiano, E., Guida, A., Olivares, L., and Picarelli, L.: Hydrological modelling of a slope covered with shallow pyroclastic deposits from field monitoring data, Hydrol. Earth. Syst. Sci., 17, 4001–4013, https://doi.org/10.5194/hess-17-4001-2013, 2013.

Greco, R., Comegna, L., Damiano, E., Marino, P., Olivares, L., and Santonastaso, G. F.: Recurrent rainfall-induced landslides on the slopes with pyroclastic cover of Partenio Mountains (Campania, Italy): Comparison of 1999 and 2019 events, Eng. Geol., 288, 106160, https://doi.org/10.1016/j.enggeo.2021.106160, 2021.

Greco, R., Marino, P., Santonastaso, G. F., and Damiano, E.: Interaction between Perched Epikarst Aquifer and Unsaturated Soil Cover in the Initiation of Shallow Landslides in Pyroclastic Soils, Water, 10, 948, https://doi.org/10.3390/w10070948, 2018.

Marino, P., Santonastaso, G. F., Fan, X., and Greco, R.: Prediction of shallow landslides in pyroclastic-covered slopes by coupled modeling of unsaturated and saturated groundwater flow, Landslides, https://doi.org/10.1007/s10346-020-01484-6, 2021.