Reply to Reviewer #3

General comments

The manuscript describes a field study of a slope, enriched by simulation results from a 1D soil column model. There is relevance in the author's work and presented data, however, unfortunately, I struggle to find the novelty and contribution to current research in this work. All in all, it seems more like an analysis of the behaviour of a simple 1D-model (without considering the model's parameters), and it does not become clear what added value this research offers. I would not count monitoring antecedent conditions as new insight (conclusions, line 655)?

It is clear from the comments of the Reviewer (as well as from those of the other Reviewers), that the way we wrote the Introduction, especially the description of the goal of the study, resulted unclear to the reader. The main aim of the study is to understand how the seasonal slope conditions, related to climate forcing, may affect the capability of the soil of retaining rainwater infiltration for a time long enough to potentially determine critical conditions as a consequence of rainfall events (e.g., the triggering of landslides).

To this aim, to enrich the available field dataset, in which very few times the soil mantle approached critical conditions (as it often occurs in studies dealing with geohazard assessment, as the hazardous conditions are rare per definition), we generated a synthetic dataset with already existing models, calibrated and validated in previous studies based on experimental data collected in the field (Greco et al., 2013; Comegna et al., 2016; Greco et al., 2018), but here the model is just a tool to generate the dataset. Once generated, the dataset represents "the reality", and we analyzed it mimicking what could be done if, rather than synthetically generated data, one was handling real field monitoring data. In fact, we were mostly looking for a way to identify the major cause-effect relationships between (measurable) inputs and outputs before (possibly, but not necessarily) building a model for the interpretation of such relationships, rather than evaluating the sensitivity of an (already available) model output to variations in the input (although the Random Forest analysis also allows quantifying the information content of each considered input variable, but without introducing any mathematical model structure, as it is based on the application of logical rules (IF-THEN-ELSE) to classify the input variables).

This is the reason why we did not test the effects of uncertainty/variability of model parameters: it was simply out of the scope of our study. The obtained results show, for the assessment of the hazard of rainfall-induced landslides and debris flows, the potential value of supplementing the monitoring of rainfall (which is the only monitored variable in nearly all the real world applications) with the monitoring of soil moisture before the rainfall (this is something that is recently being recognized by many researchers, but rarely adopted in operational hazard assessment systems), and the monitoring of the water level in the shallow aquifer developing in the uppermost part of the underlying bedrock. This is quite a novel result, to our knowledge never proposed to predict the response to precipitation of a shallow unsaturated soil mantle. Groundwater level is usually considered an informative variable only for deep-seated landslides, as in that case the (deep) slip surface of the landslide can be below the groundwater table. In the studied geomorphological context, monitoring these antecedent conditions is indeed a novelty.

In the revised manuscript, besides rewriting the Introduction, we will more clearly underline this novel aspect in the discussion of the results, as well as in the Conclusions.

The 1D model comes necessarily with many simplifciations (which of course also has its advantages). But the shortcomints are not adressed (and the advantages - fast runtime etc. - not really exploited). E.g. heterogeneity of hydraulic conductivities on a slope, differences in lower aquifer pressure based on position (top, bottom, local gradient) in the slope, parameter uncertainty in general, differences in layer thicknesses across the slope, etc.

We thank the Reviewer for raising this issue. Obviously, heterogeneities of the soil mantle (either morphological, e.g., slope inclination, soil mantle thickness, or physical, e.g., soil layers with different hydraulic properties) may induce 3D effects in the flow processes. However, 3D effects are expected to be not particularly significant in the soil mantle of the studied slopes, for several reasons. First, owing to the geometry of the slopes (i.e., hundreds of meters long with a soil mantle of few meters), the water potential gradients are such that significant deviations of the flow from the vertical direction (or, more precisely, from the direction orthogonal to ground surface) can occur only when the soil approaches saturation, so that capillarity gradients become small and gravitational gradient prevails (along a steeply inclined slope, in this condition the component of the gradient parallel to the slope becomes significant). In addition, the attainment of soil saturation is very unlikely, owing to the very high porosity (as high as 75%). Furthermore, the high inclination angles, in most slopes larger than 35°, imply that slope failure (landslide) would occur before soil attains saturation. Finally, the very high hydraulic conductivity (as high as 30 mm/h), together with the usually unsaturated soil conditions (soil capillary potential rarely overcomes -0.5 m: Cascini et al., 2014; Comegna et al., 2016; Napolitano et al., 2016), makes overland runoff very small, even during the most intense rainfall events (Greco et al., 2018; Marino et al., 2020). In short, lateral redistribution of infiltration flow can be considered quite small in the soil mantle of the studied slopes. In the revised manuscript, we will add more information about the characteristics of the studied slopes and soil (Section 2.1), and we will give some justification of the use of the simplified 1D model in Section 2.2.2.

About the variability of the groundwater table depth, this is also obviously true (and indeed, observations made in two piezometers, recently installed at two different altitudes along the slope, confirm that the groundwater table depth may be quite different). However, the use that we make of the groundwater level information is to discriminate "low" levels (clusters 1 and 3 of Figures 8, 9 and 10) from "high" levels (cluster 2 of Figures 8, 9 and 10) or "very high" levels (cluster 4 of Fig. 10). Depending on the availability of monitoring instruments, this could be made with a single piezometer, as well as with several piezometers (but, although with different levels, if the groundwater level in a piezometer is high, it will be likely high also in the others, unless they are so far from each other that they are monitoring disconnected groundwater systems). This aspect will be better clarified in the discussion of the results of the revised manuscript.

Given such a relatively simple model, it should be possible to run uncertainty analysis or sensitivity analysis – and that possibility should be exploited.

Then, it is unclear why a Random Forest algorithm is used to emulate the physically-based model outputs – why not simply use an ensemble of the physically-based model, and do some uncertainty and sensitivity analysis on that? Also, some basic considerations when using RF have been ignored (hyperparameter search, sound cross validation strategy, discussion of size of dataset etc).

Therefore, I recommend to address the analysis of the physically-based model more extensively before venturing into a ML analysis of its results.

This comments again clearly indicate that, in the Introduction, we failed to describe the aims of the study. In the revised manuscript, we will add paragraphs in the Introduction and in the Materials and Methods sections, to better explain the choice of Machine Learning (and specifically Random Forest) instead of a sensitivity analysis. In fact, we believe that adding a sensitivity analysis, which is out of the scope of our study, would be misleading, for the following reasons:

First, we analyzed the dataset mimicking what could be done if, rather than synthetically generated data, one was handling real field monitoring data. In fact, we were mostly looking for a way to identify the major cause-effect relationships between (measurable) inputs and outputs before (possibly, but not necessarily) building a model for the interpretation of such relationships, rather than evaluating the sensitivity of an (already available) model output to variations in the input (although the Random Forest analysis also allows quantifying the information content of each considered input variable).

Second, the sensitivity analysis is usually carried out to evaluate the effects of input (and parameter) uncertainty on model predictions. In this study, the model chain (already calibrated and validated previously: Greco et al., 2013; Comegna et al., 2016; Greco et al., 2018) is used as a tool to generate a (richer) synthetic dataset (this is a common problem in landslide studies, as field monitoring data records, even when they are relatively long, usually contain very few data representative of potentially critical situations). The model is assumed to represent "the reality", and adding a sensitivity analysis may result misleading, as it would move the focus to the performance of the model (which, in general, could also not exist).

Third, the adopted Random Forest analysis, which allows highlighting the most informative combination of measurable variables to predict the output, is somehow a sensitivity analysis as well, as it gives some indications about the relative importance of the input variables on the possibility of predicting the output, without introducing any mathematical model structure, but simply relying on the application of logical operators (IF-THEN-ELSE) between the variables.

Specific comments

Section 1. Introduction

Line 94 ff: Remember that you are describing location-specific aspects – "fractured limestone", depth of soil above bedrock – i.e. this is not generally applicable.

Thank you for catching this. Indeed, this paragraph refers specifically to the presented case study (i.e., slopes with a shallow pyroclastic soil mantle covering a fractured limestone bedrock). Therefore, it should be moved to the final part of the Introduction, where the characteristics of the case study are briefly anticipated.

Line 105ff: This paragraph with your objectives could be formulated more clearly. E.g. the first sentence – please reformulate and state more clearly what your objectives are.

Apart from some sentences, which should be rewritten to improve the language and style, this paragraph must be totally rewritten, as it misled the Reviewers (and it would mislead all the readers of the paper). In fact, as already pointed out in the replies to previous comments from this Reviewer, the focus of the study is not on the physically based model (which is instead, by our mistake, mentioned firstly in the paragraph of the objectives of the study), but on the interpretation of field monitoring data. Coupled with the NRSP stochastic model of rainfall, the model was just a tool to generate a rich synthetic dataset, which was then analyzed as if it were obtained by field measurements. This approach was chosen because field data series always contain few data representative of potentially critical conditions (in other words, landslides and debris flows, as well as other rainfall-induced geohazards are rare phenomena), so to have enough data to carry out statistically significant analyses. The final part of the Introduction will be completely reformulated to state the goal of the study more clearly.

Section 2. Material and methods

Line 140, line 146: Make clear that you describe/summarize the methods applied in *your* study (e.g. by adding "see section 2.2")

In the revised manuscript we will make more clear that this initial part of section 2 anticipates what is then described in the following subsections.

Figure 1: I suggest a visualizing a DEM as background in one of the images – maybe the smaller inset? However, it remains unclear whether the inset is necessary at all, or one map would do just fine. Indicate the location of the monitoring station. Moreover, the outline of the inset seems incorrect, as well as the red star indicating the main scarp does not match the location indicated in the inset.

We will use a DEM instead of a photo in the smaller inset, so to give the reader information about the morphology of the studied slope. We will also be more precise in selecting the zoomed rectangle, as well as we will move the red star in the correct position (thank you for catching this mistake), which should be the main scarp of the landslide.

Figure 2: This is based on data from Damiano et al. 2012?

Indeed Figure 2 is adapted from Damiano et al. (2012), and this should have been mentioned in the caption. However, as the layered nature of the soil cover is an information that is never exploited in this study, we are considering summarizing its description in the revised manuscript, likely eliminating the figure.

Line 196: By "pyroclastic ashes" you here refer to the entire soil profile? Or only to the layer "Volcanic ashes" in Fig. 2? The terminology used around here should be made more clear.

We agree that we should be consistent, using the word "pyroclastic" throughout the entire manuscript.

Figure 3: Please improve this figure. E.g. north up, show its outline in Figure 1.

In the revised manuscript, we will change the photo, using one with the standard orientation (North upward).

Section 2.1.1 partly lacks details – what has been measured exactly? How long? What temporal resolution? (part of it comes later in section 3.2, and should be noted here)

Thank you for the suggestion. In the revised manuscript we will add more information about the monitoring campaign, and we will also move to Section 2.1.1 the information given between lines 487 and 492 (Section 3.2), leaving there only a small mention.

Section 2.2.1: Reference for the NSRP model is lacking. Also, some kind of comparison (various statistics?) of the synthetic time series with the observed time series would be appreciated.

Thank you for raising these issues. In the submitted manuscript, we decided to describe very briefly the stochastic NRSP model used for synthetic rainfall generation, giving some references to let the interested readers get more information. We understand that we gave too little information, given that the synthetic rainfall series plays an important role in our methodology. In the following, we give detailed information to this Reviewer, so that he can judge about how the generated synthetic rainfall resembles the real experimental record. In the revised manuscript, we will try to find a trade off between the sake of brevity (the synthetic rainfall generation is here only a tool, but it is not the core of the study) and the need for more information. Possibly we will put some of the information in an appendix.

The NRSP stochastic model of rainfall (Neyman and Scott, 1958; Rodriguez-Iturbe et al., 1987a, b; Cowpertwait et al., 1996) describes the process of point rainfall as a superposition of randomly arriving rain clusters, each containing several rain cells with constant intensity. The hyetograph within a cluster is obtained by the superposition of the intensity of the various cells belonging to the cluster. It has been calibrated based on 17 years experimental data (2000-2016) of rainfall depth at 10 min resolution, recorded by the rain gauge managed by Civil Protection in Cervinara. The calibration has been carried out by minimizing, for rainfall aggregated at various durations, the difference between the following quantities, estimated by the model and calculate from the experimental data: mean, variance, lag 1 autocorrelation, probability of dry interval, probability of transition from dry-to-dry interval, probability of transition from wet-to-wet interval. The calibration procedure is based on the one proposed by Coptwertwait et al. (1996), and it is described in detail in Peres and Cancelliere (2014). To account for the seasonality of rainfall, these quantities have been calculated month by month in the experimental record (Fig. R1), suggesting that the calibration of the NRSP model should be carried out separately for seven homogeneous periods (September, October, November, December-March, April, May-June, July-August).

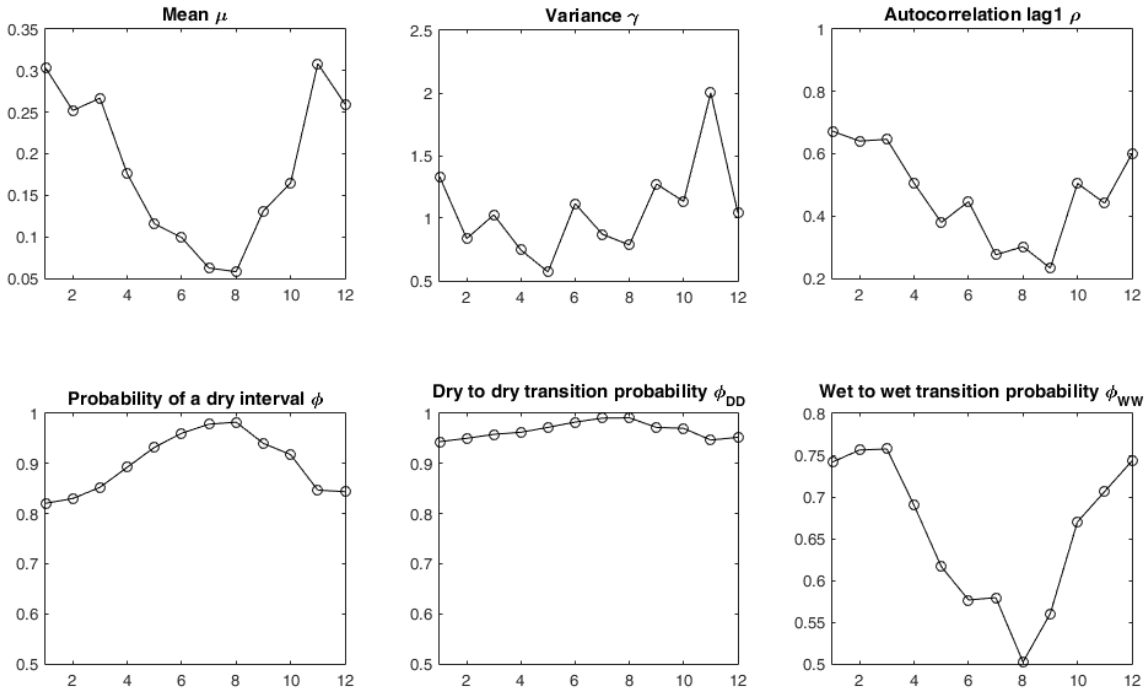


Figure R1. Monthly plot of hourly rainfall characteristics calculated based on the experimental data of the rain gauge of Cervinara.

Table R1 gives the obtained parameters of the NRSP stochastic model, where λ [h^{-1}] represents the parameter of a Poisson process describing the arrival of clusters; ν [-] is the mean number of cells in a cluster, also described by a Poisson process; θ [h^{-1}] is the parameter of an exponential probability distribution describing the arrival times of each cell in a cluster, expressed as the number of time intervals of 10 minutes starting from the beginning of a cluster; η [h^{-1}] is the parameter of an exponential probability distribution describing the duration of rain cells; ξ [$h^b mm^{-b}$] is the parameter of a Weibull probability distribution describing the rain intensity of cells, with cumulative probability function $F(x;\xi,b) = 1 - exp(-\xi x^b)$, in which x is cell rain intensity and the parameter $b = 0.8$ has been set a priori (Cowpertwait et al., 1996).

Table R1. Parameters of the NSRP model.

Parameter	Sept	Oct	Nov	Dec-Mar	Apr	May-Jun	Lug.Aug
$\lambda\ [h^{-1}]$	0.015	0.00524	0.00257	0.0238	0.00809	0.00386	0.00900
$\nu\ [-]$	2.68	36.4	57.1	2.60	38.7	21.6	1.40
$\beta[h^{-1}]$	0.265	0.156	0.0167	0.813	0.123	0.116	24.5
$\eta[h^{-1}]$	1.41	57.3	1.43	0.280	15.5	8.59	1.23
$\xi[h^b\ mm^{-b}]$	0.330	0.047	0.450	0.967	0.186	0.158	0.268

The adherence of the rainfall generated with the stochastic model to the experimental rainfall data has been tested by evaluating rainfall characteristics different from those used for the calibration. For instance, Figure R2 shows the comparison of the rainfall depth cumulated over one year for the experimental data and the synthetic data generated with the calibrated NRSP model.

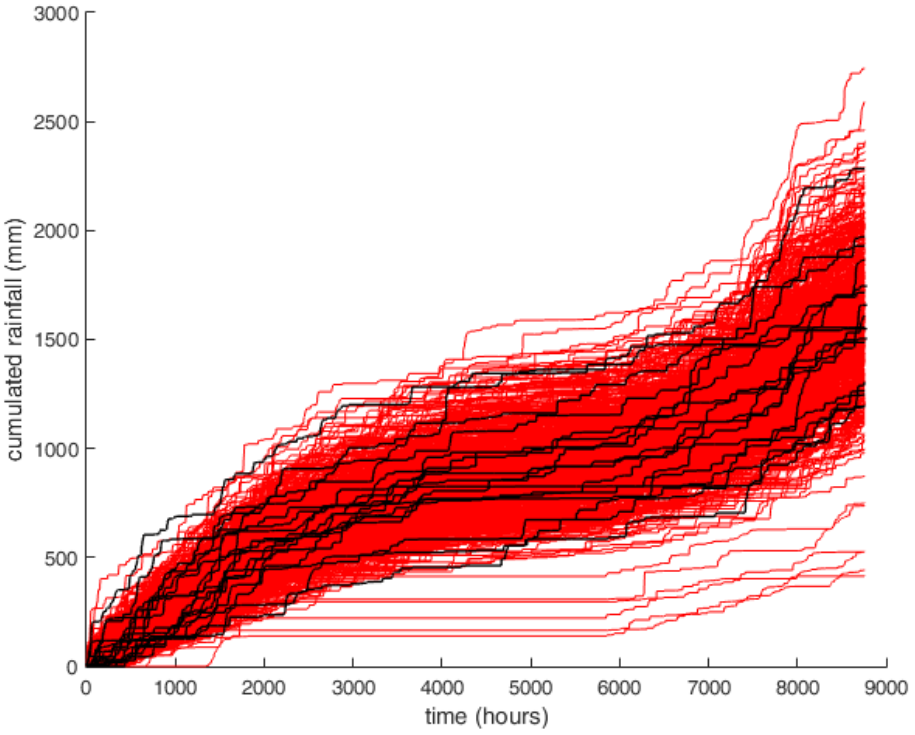


Figure R2. Comparison of observed (black) and simulated (red) cumulated rainfall plots in a year.

In figure R3, the boxplot of the maximum hourly rainfall in one year, observed in the experimental dataset of 17 years, is compared with the same boxplot referred to 20 series of 17 years randomly extracted from the generated 1000 years synthetic rainfall series. Several synthetic 17 years intervals show a distribution of the maximum hourly rainfall close to the observed one.

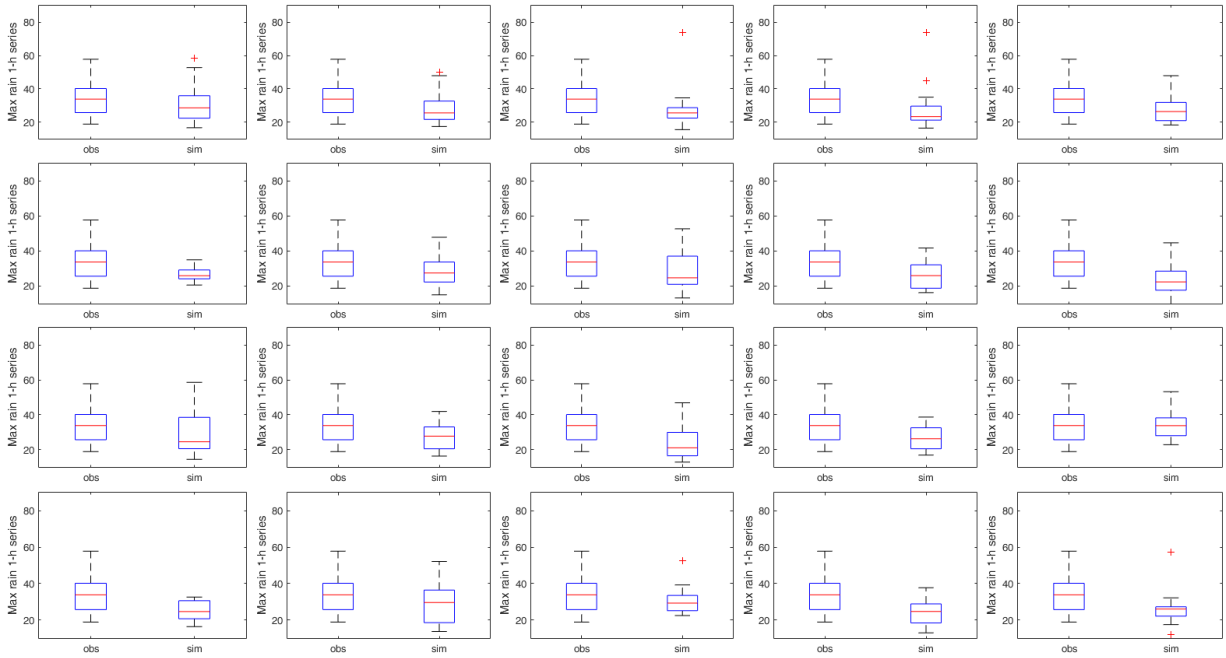


Figure R3. Comparison of observed and simulated distributions (boxplots) of the maximum hourly precipitation in a year, for series of the same length. Each panel shows the distribution for the 17 observed years (boxplot is always the same), and 17 randomly picked simulated years.

Regarding the comparison between synthetic and observed wet and dry intervals, figure R4 shows the scatterplot of duration and total rain depth of the events, sorted with a separation "dry" interval of 24 hours with less than 2 mm rainfall from the experimental dataset (blue dots) and the synthetic dataset (grey dots). The plots show how the synthetic data contain the observed ones, and that the shape of the dot clouds looks quite similar.

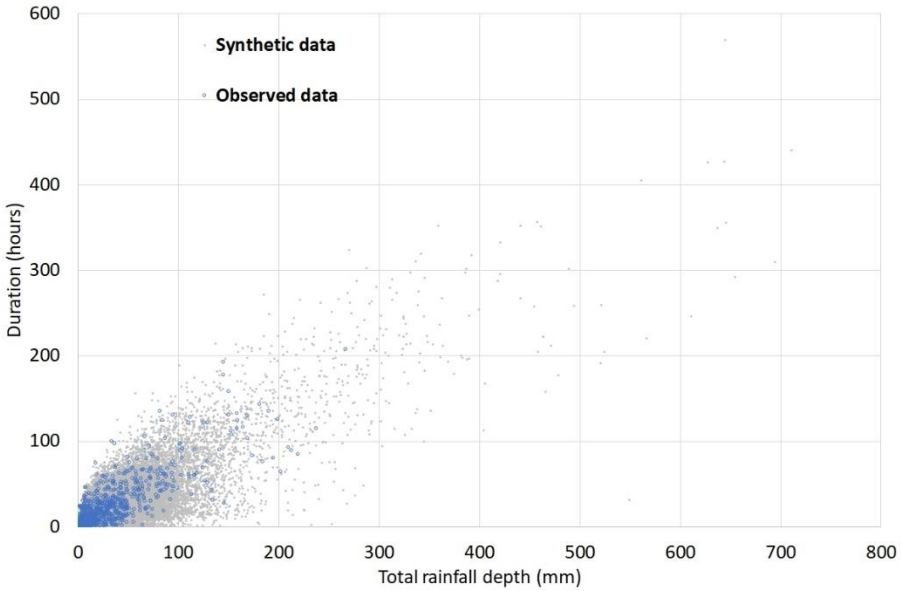


Figure R4. Scatterplot of rainfall event duration vs. total rainfall event depth. The events have been sorted within the rainfall datasets by considering a separation "dry" interval of 24 hours with less than 2 mm rainfall. Blue dots represent events extracted from the 17 years experimental rainfall dataset; grey dots represent events extracted from the 1000 years synthetic rainfall dataset.

Figure R5 shows the frequency distributions of the durations of dry intervals belonging to the 17 years rainfall dataset, and the same distribution for the dry intervals extracted from the 1000 years synthetic dataset: the two distributions look nearly identical.

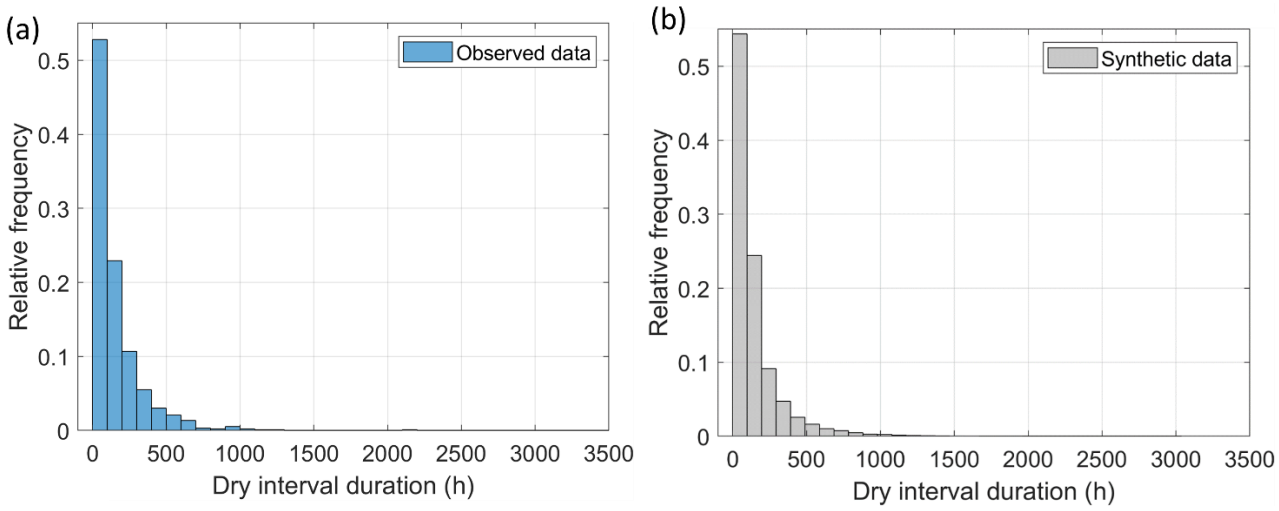


Figure R5. Frequency distributions of dry interval durations for events extracted from the 17 years experimental rainfall dataset (a) and events extracted from the 1000 years synthetic rainfall dataset (b). The events have been sorted within the rainfall datasets by considering a separation "dry" interval of 24 hours with less than 2 mm rainfall.

Line 284ff: For your definition of separate rainfall events – can you argue for the threshold of 24h with less than 2mm rainfall to separate events? This seems very little – does the slope really drain in such a short time? I.e. have effects of preceding events really disappeared after such a short time? I do not follow your argument that the volumetric water content at 10cm depth is sufficient to conclude that. Also, you show model results in Figure 4, right? That is not directly apparent to the reader.

The separation of events within the continuous rainfall record aims at linking the occurrence (or non-occurrence) of critical conditions to a rainfall event, so that they can be considered as a direct consequence of that rainfall event. This is commonly made when empirical predictive tools (e.g., rainfall thresholds: Segoni et al., 2018a, b; Guzzetti et al., 2020; Piciullo et al., 2020) are implemented as part of early warning systems, e.g., against rainfall-induced landslides or debris flows, and the definition of the separation criterion is usually made empirically, looking at the performance of the predictor with different choices of the separation criterion.

From a physical viewpoint, especially when one is interested in the separation between the role of antecedent conditions, i.e., related to previous precipitation (and drainage/evapotranspiration) history, from the direct effects of the last precipitation event, it is quite complex to define a suitable separation criterion, specifically if dealing with slow processes activated by precipitations, such as the infiltration through the unsaturated soil layer. In fact, to completely separate what depends on "previous" precipitation from what is linked to the last rainfall event, one should wait for the infiltration process initiated by previous precipitations to be finished, and, in a soil layer of few meters thickness, it may take several days. Extending so much the dry time interval between two separate events, especially during rainy seasons, would imply the aggregation of several events in a single one, thus leading to long rainy periods, rather than events, thus preventing the desired separation of antecedent conditions from direct effects of events. So, as we have defined the "response" of the soil layer as its attitude to retain infiltrated rainwater after the end of a rain event, looking at the moisture of the topsoil layer seemed a good trade-off: topsoil moisture controls the infiltration at the ground surface, hence when

gravitational drainage from the topsoil is already over (the field capacity has been reached), the infiltration of a new rainfall input through the ground surface would not depend (or it would only little depend) on the remnants of the infiltration process caused by previous precipitation. In this respect, we tested a separation dry interval of 24 hours, commonly used when the available rainfall data are at daily resolution (Berti et al., 2012; Leonarduzzi et al., 2017; Peres et al., 2018), and anyway in line with the empirical choices that are commonly made in the early warning community (Segoni et al., 2018a).

We will clarify that Figure 4 shows synthetic data.

Section 2.2.2: I miss a discussion of the limitations of the 1D model. Assuming that some lateral in the soil layer (and deeper aquifer) exists, this results in different groundwater level across different parts of the slope / for different gradients. Etc… 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.

Please, see our answer to a previous similar comment about the limitations of 1D model. Again, it clearly arises that the Reviewer was misled by our unclear description of the goal of the study, which is not about evaluating the performance of a model, but about how to extract information about cause-effect relationship from a dataset of hydrological variables describing the response to precipitation of the pyroclastic soil mantle of the studied slope. We analyzed the data as if they came from field monitoring, although, to get a richer dataset, we generated a synthetic dataset. The model, already developed, calibrated, and validated in previous studies (Greco et al., 2018), was here used just as a tool to generate the synthetic dataset, by coupling it with the NRSP stochastic model of rainfall. Hence, a sensitivity analysis of the model output to parameters is out of the scope of this study.

Section 2.2.3/Line 362ff: Referring back to my comment to line 332ff: You mention that you extract variables before the onset of each rainfall event, as the "would be measurable in the field". E.g. those are aquifer water level – which, I assume, is largely different on the top of the slope from the bottom of the slope.

As we already replied to a previous comment from this Reviewer, the groundwater table depth is indeed variable throughout the slope (observations made in two piezometers, recently installed at two different altitudes along the slope, confirm that the groundwater table depth may be quite different). However, the use that we make of the groundwater level information is to discriminate "low" levels (clusters 1 and 3 of Figures 8, 9 and 10) from "high" levels (cluster 2 of Figures 8, 9 and 10) or "very high" levels (cluster 4 of Fig. 10). Depending on the availability of monitoring instruments, this could be made with a single piezometer, as well as with several piezometers (but, although with different levels, if the groundwater level in a piezometer is high, it will be likely high also in the others, unless they are so far from each other that they are monitoring disconnected groundwater systems). This aspect will be better clarified in the discussion of the results of the revised manuscript.

Line 370: An actual quantification of soil water content based on satellite observations is hard (rather than a relative value), especially on such small scales – this limitation should be mentioned.

We will briefly mention the limitations of satellite products compared to field observations.

Section 2.3.1

Unfortunately, it is hard to understand how you obtain your dataset with triplets of variables. Why did you chose three of the four to predict? You are predicting simulated change in soil water storage, right? What is the time interval? Or is it just aggregated values per rainfall event? How many datapoints? And what are your variable inputs? Only the four mentioned variables? Why did you not also run the 1D physically based model with various parameterizations – I assume the outputs would have looked quite different. Also, why did you chose RF, and how did you decide for number of trees, splits etc (hyperparameters)?

Again, the Reviewer has been misled by the confusion in the description of the goal of this study. So, we have already replied to the point about evaluating the effects of parameters uncertainty/variability on model output, which is out of the scope of this study.

Some of the requested information is already given in the manuscript: we evaluate the change of soil storage between after and before any rainfall event; the number of data is given (around 53000); the choice of the variables is described as the outcome of Random Forest (RF) analysis. We did not test other choices of the variables to be monitored, as we wanted to stick to what can be easily obtained by means of currently available instruments (i.e., satellite products, field soil moisture measurement networks, piezometers, stream water stage sensors).

About the choice of the RF, as we already pointed out in a previous answer, it was made to mimic what could be done if, rather than synthetically generated data, one was handling real field monitoring data. In fact, we were mostly looking for a way to identify the major cause-effect relationships between (measurable) inputs and outputs before (possibly, but not necessarily) building a model for the interpretation of such relationships, rather than evaluating the sensitivity of an (already available) model output to variations in the input, and Random Forest allows quantifying the information content of each considered input variable without introducing any mathematical model structure, but just relying on the application of logical operators (IF-THEN-ELSE) between the variables.

It looks clear that, aiming at brevity, we gave too little information about how the RF was implemented. Here we provide detailed information to this Reviewer about the training and validation of the RF model, as well as about the choice of hyperparameters.

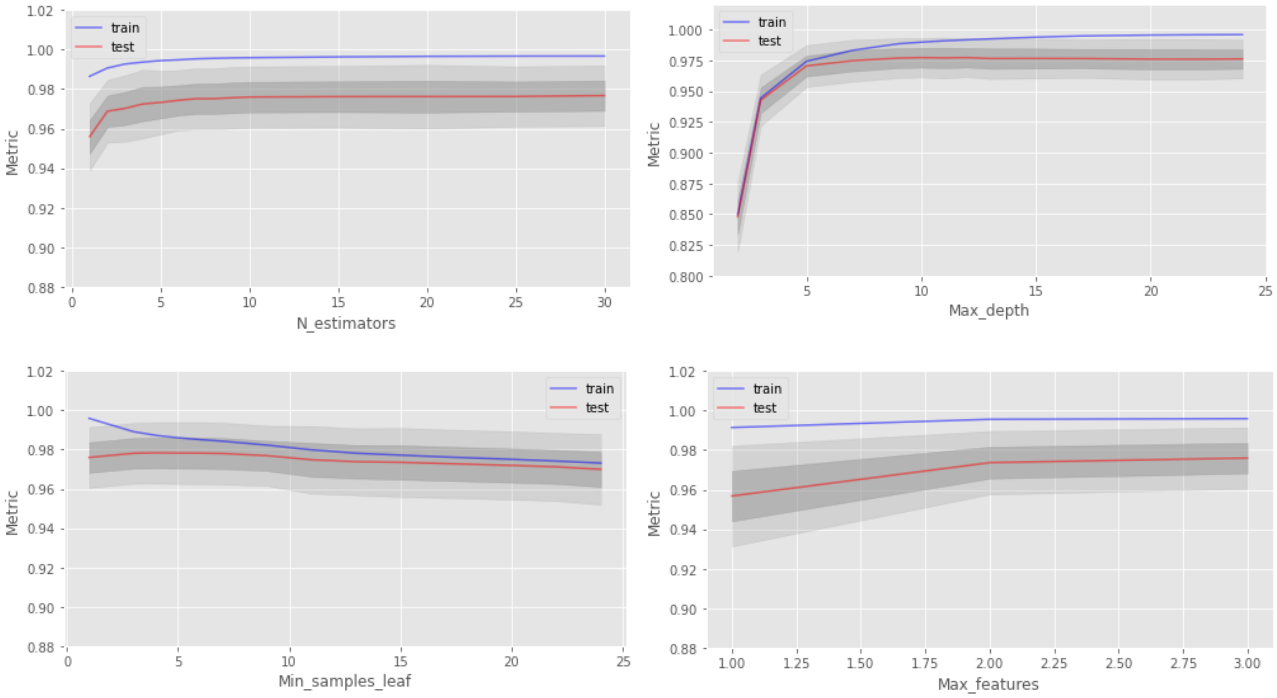
To evaluate the performance of the Random Forest model, the cross-validation technique was used. In cross-validation, the dataset is divided into k equal parts, also known as folds. Then, for each fold, the Random Forest model is trained on the remaining k-1 folds of the data and tested on the remaining fold. We chose k=5, so that the process was repeated 5 times, every time using a different fold (20% of the dataset) as the validation set. A performance metric was calculated for each fold, to estimate how well the RF model perform on new data. We used the explained variance score, computed as follows:

$$metric = 1 - \frac{Var(y - \hat{y})}{var(y)}$$

$Var(y - \hat{y})$ and $Var(y)$ are the variance of prediction errors and actual values respectively. Higher values of explained variance indicate better performance. In addition, the tuning of the hyperparameter of the model was performed based on the cross-validation results to select the optimal set of hyperparameters. In other words, the random forest model was fitted k-times to the data provided by the cross-validation, changing the value of the following hyperparameters once at time:

- *n_estimators: the number of trees to build in the forest;*
- *max_depth: the maximum depth of each decision tree in the forest;*
- *min_samples_leaf: the minimum number of samples required to be at a leaf node;*
- *max_features: the number of features to consider when looking for the best split.*

The following plots show the trend of the hyperparameters vs. the performance metric, the explained variance score. The blue and red lines are the average values of the performance metric computed for the train datasets and the test dataset, respectively, provided by the cross-validation process. The gray and light gray bands represent average values of the metric +/- the standard deviation and two times the standard deviation, respectively.



According to the evidence shown by the previous plots, the search of the optimal parameters was carried out using the following ranges for the hyperparameters:

- *n_estimators was fixed to 10;*
- *max_depth: [5, 10, 15];*
- *min_samples_leaf: [1, 3, 5, 7, 9],*
- *max_features: [2,3].*

The obtained best parameters are max_depth=10; min_samples_leaf=3; max_features=3.

As well known, RF algorithm can work well with high dimensional or multidimensional data, but having a high number of features can lead to overfitting. Therefore, it's important to adjust the hyperparameters (e.g., max_features) to prevent overfitting and create a robust model. Anyway, the synthetic dataset is characterized by only three features (the three variables quantifying antecedent conditions), and a very large number of samples (more than 50000), hence overfitting could be excluded, also owing to the cross-validation method used for model training.

Section 2.3.2

Again, what data exactly do you cluster? The covariates described in the previous section? Also (line 411) – you do not really use “spatial” data here, do you?

The clustering is carried out on the triplets that, based on the results of the RF analysis, seems to be the most suitable to describe the effect of antecedent conditions (prior to the onset of each rainfall event) on the attitude of the soil mantle to retain infiltrating rainwater.

In line 411 we meant that k-means clustering evaluates the distance between the dots in the space of the variables to which the clustering is applied. Nothing to do with distance in the field. We will rephrase the sentence to make it clearer.

Section 3.1

Table 3 vs Table 2: Doesn't this shown effect of the normalization of delta(S) to H simply show that there is a quite good linear correlation between H and delta(S) – why then not simply use a linear regression model?

It is clear that we gave too much emphasis to the analysis of deltaS, so that the Reviewer was misled. The results just show that, obviously, the more it rains, the more soil storage increases, and if one is interested in evaluating the response of the soil mantle to precipitation, one should look at the ratio between deltaS and H. We will remove Table 2 and just briefly explain the choice of deltaS/H.

Figure 5: The water levels should not be shown on logarithmic scale (that just hides deviations?). Also, make clear that you compare simulated aquifer water levels (ha) with observed stream water levels (hs). Can you try to argue better for your statement "a direct relationship links the water level in the aquifer and the water level in the stream" based on this? As you also mention in section 4, line 632: "substantial agreement between synthetic and experimental data" – this has to be quantified.

We beg to disagree on the first remark: if plotted along a Cartesian axis, many of the dots would collapse very close to the zero (the "low" water levels), thus hiding the existence of a cluster containing a large number of antecedent conditions.

The agreement between synthetic and field data is indeed what we meant to show in Figure 5 and Figure 6. From those figures, you can directly compare the few measured values of soil moisture of the upper 100 cm of the soil profile with the synthetic data. About the synthetic groundwater level data, they are compared with stream water level data. The reasons for this choice are several:

- So far, we have measurements of stream water level, while only recently we have installed two piezometers in the epikarst.

- The streams are supplied by groundwater coming from the fractured bedrock with very little contribution of overland runoff (less than 1% of the rainfall) only during the most intense rainstorms (it is revealed by the timing of the observed hydrographs in response to rainfall as well as by measurements of electric conductivity of stream water: Marino et al., 2020), so there might be a close relationship linking stream water level and groundwater level.

- Installing piezometers in the fractured limestone is a complex operation, owing to the mechanical resistance of the rock, which obliges to the use of powerful drilling machines; we have recently installed two piezometers (July 2020), but one of them could penetrate the limestone only for less than a couple of meters, as the machine that could be carried in that steep part of the slope (a light one) was not able to drill more depth; the second piezometer, which is at the foot of the slope, in a much less steep terrain, penetrates 16 meters below the ground, but there we have found a different kind of soil mantle (not only pyroclastic soil, but also some meters of alluvial deposits), in total more than 10 meters thick; as we had no clue of the degree of interconnection of the fractured system in the limestone, we decided to extend the pervious part of the piezometer (the filter) to almost the entire penetration depth in the limestone (1,5 meters for the first piezometer, 5 meters for the second one), as a shorter filter at the base of the piezometer (as it is usually done) would increase the risk of not intercepting any connected fracture; in this way, there is more chance for water to enter the piezometer, but, as it may enter at any height along the filter and then pond at the base

of the piezometer, we cannot convert the water depth that we measure in the piezometer into a groundwater level; during the 2020/2021 hydrologic year we did not measure any water in the piezometers (2020 was a quite dry year), but in December 2021, after a quite rainy autumn (more than 900 mm between September and December), for the first time water appeared in both the piezometers, confirming that the temporary aquifer actually develops in the epikarst during rainy periods; until summer 2022, the piezometric measurements were made irregularly with a freatimeter, but in autumn 2022 we have installed an automatic sensor inside the piezometer on the steep terrain (the first one), and this winter we have observed a slight increase of groundwater level once the cumulated rainfall from September exceeded 800 mm.

- Stream water seems to appear and disappear consistently with the groundwater fluctuations, although, so far, we have too few data to demonstrate it; however, measuring stream water level is much easier than groundwater level in the studied context, and it could be an effective surrogate of groundwater level.

- The use that we do with the synthetic groundwater level data (that could be done with field data, either of groundwater or of stream water level) is just to discriminate between "high" level and "low" level, as a proxy to identify active subsurface drainage conditions.

The colored dots of Figures 5 and 6 also show that the seasonality of the synthetic variables is consistent with that of the observed variables.

Technical corrections

Some general language editing is necessary

We will double-check the English language to remove language, syntax and style mistakes.

Title, and also in the manuscript: "precipitations" cannot be said – maybe replace with "precipitation events" or similar

Based also on similar remarks made by another Reviewer, the title will be changed to "Understanding hydrologic controls of sloping soil response to precipitation through Machine Learning analysis applied to synthetic data", and the wrong plural "precipitations" will be corrected throughout the entire manuscript.

References

Berti, M., Martina, M. L. V., Franceschini, S., Pignone, S., Simoni, A., & Pizziolo, M. (2012). Probabilistic rainfall thresholds for landslide occurrence using a Bayesian approach. Journal of Geophysical Research: Earth Surface, 117(4). https://doi.org/10.1029/2012JF002367

Cascini, L., Sorbino, G., Cuomo, S., & Ferlisi, S. (2014). Seasonal effects of rainfall on the shallow pyroclastic deposits of the Campania region (southern Italy). Landslides, 11(5), 779–792. https://doi.org/10.1007/s10346-013-0395-3

Comegna, L., Damiano, E., Greco, R., Guida, A., Olivares, L., & Picarelli, L. (2016). Field hydrological monitoring of a sloping shallow pyroclastic deposit. Canadian Geotechnical Journal, 53(7), 1125–1137. https://doi.org/10.1139/cgj-2015-0344

Cowpertwait PSP, O'Connell PE, Metcalfe AV, Mawdsley JA (1996) Stochastic point process modelling of rainfall. I Single-site fitting and validation. J Hydrol. https://doi.org/10.1016/S0022-1694(96)80004-7

Damiano, E., Olivares, L., & Picarelli, L. (2012). Steep-slope monitoring in unsaturated pyroclastic soils. Engineering Geology, 137–138, 1–12. https://doi.org/10.1016/j.enggeo.2012.03.002

Greco, R., Comegna, L., Damiano, E., Guida, A., Olivares, L., & Picarelli, L. (2013). Hydrological modelling of a slope covered with shallow pyroclastic deposits from field monitoring data. Hydrology and Earth System Sciences, 17(10), 4001–4013. https://doi.org/10.5194/hess-17-4001-2013

Greco, R., Marino, P., Santonastaso, G. F., & Damiano, E. (2018). Interaction between perched epikarst aquifer and unsaturated soil cover in the initiation of shallow landslides in pyroclastic soils. Water (Switzerland), 10(7). https://doi.org/10.3390/w10070948

Guzzetti, F., Gariano, S. L., Peruccacci, S., Brunetti, M. T., Marchesini, I., Rossi, M., & Melillo, M. (2020, January 1). Geographical landslide early warning systems. Earth-Science Reviews. Elsevier B.V. https://doi.org/10.1016/j.earscirev.2019.102973

Leonarduzzi, E., Molnar, P., & McArdell, B. W. (2017). Predictive performance of rainfall thresholds for shallow landslides in Switzerland from gridded daily data. Water Resources Research, 53(8), 6612–6625. https://doi.org/10.1002/2017WR021044

Marino, P., Comegna, L., Damiano, E., Olivares, L., & Greco, R. (2020). Monitoring the hydrological balance of a landslide-prone slope covered by pyroclastic deposits over limestone fractured bedrock. Water (Switzerland), 12(12). https://doi.org/10.3390/w12123309

Napolitano, E., Fusco, F., Baum, R. L., Godt, J. W., & De Vita, P. (2016). Effect of antecedent-hydrological conditions on rainfall triggering of debris flows in ash-fall pyroclastic mantled slopes of Campania (southern Italy). Landslides, 13(5), 967–983. https://doi.org/10.1007/s10346-015-0647-5

Neyman, J., & Scott, E. L. (1958). Statistical Approach to Problems of Cosmology. Journal of the Royal Statistical Society: Series B (Methodological), 20(1), 1–29. https://doi.org/10.1111/j.2517-6161.1958.tb00272.x

Peres DJ, Cancelliere A (2014) Derivation and evaluation of landslide-triggering thresholds by a Monte Carlo approach. Hydrol Earth Syst Sci 18:4913–4931. https://doi.org/10.5194/hess-18-4913-2014

Peres, D. J., Cancelliere, A., Greco, R., & Bogaard, T. A. (2018). Influence of uncertain identification of triggering rainfall on the assessment of landslide early warning thresholds. Natural Hazards and Earth System Sciences, 18(2), 633–646. https://doi.org/10.5194/nhess-18-633-2018

Piciullo, L., Tiranti, D., Pecoraro, G., Cepeda, J. M., & Calvello, M. (2020). Standards for the performance assessment of territorial landslide early warning systems. Landslides, 17(11), 2533–2546. https://doi.org/10.1007/s10346-020-01486-4

Rodriguez-Iturbe I, Febres De Power B, Valdes JB (1987) Rectangular pulses point process models for rainfall: analysis of empirical data. J Geophys Res. https://doi.org/10.1029/JD092iD08p09645

Segoni, S., Piciullo, L., & Gariano, S. L. (2018a). A review of the recent literature on rainfall thresholds for landslide occurrence. Landslides. Springer Verlag. https://doi.org/10.1007/s10346-018-0966-4

Segoni, S., Rosi, A., Fanti, R., Gallucci, A., Monni, A., & Casagli, N. (2018b). A regional-scale landslide warning system based on 20 years of operational experience. Water (Switzerland), 10(10). https://doi.org/10.3390/w10101297