Dear Authors,

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

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).

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.

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.

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.

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.

Specific comments

Figure 1: Thanks for having taken my advice.

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.

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.

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).