nhess-2021-140 -Introducing SlideforMap; a probabilistic finite slope approach for modelling shallow landslide probability in forested situations - van Zadelhoff et al., 2021

Response to the comments of the reviewers

In these responses, we provide the original comments (in italics) and our related responses. The corresponding changes in the manuscript are indicated in the track-change version submitted along with the revised paper at the end of this response.

Response to Reviewer #1

I reviewed with interest this manuscript for possible publication in NEHSS journal. The work describes a comprehensive modeling tool to assess shallow landslides initiated by rainfall, in a probabilistic framework. The manuscript provides an interesting contribution in this field, although some aspects are strongly simplified, in contrast with others. The scientific quality is good, the reading is agile although the manuscript is overall a bit long and often dispersive. The literature review can be improved with additional appropriate references of strictly related works. The description of the climate forcing that initiates (or not) the landslide events requires significant improvement. To my opinion, the work can be published after some important clarifications and revisions.

Thanks for the general positive assessment. We gave the manuscript the best we could to shortening without skipping vital points.

1. Literature review (introduction/discussion).

The discussion on the impacts and costs of the natural hazards, from the point of view of insurance institutes, is interesting. However, in general, I found the introduction a bit dispersive, lacking in some aspects. The work of Dietrich and Montgomery, 1994, (SHALSTAB) represents the pioneering work within this approach, and it has been followed by many other deterministic work that gave different contributions in improving the hydrological modeling at support for the shallow landslide, such as the cited Iverson 2000, and, additionally, Rosso et al., 2006; Claessens et al., 2007, Arnone et al., 2011; Lepore et al., 2013; Simoni et al., 2008, Baum et al., 2002 (TRIGRS), Montrasio et al., (2011) (SLIP) (among the others). With regard to the effect of vegetation, the aspects related to the hydrological effects should be at least discussed, which can sometime be even more significant than the mechanical ones (Feng et al., 2020). An interesting review are by Chae et al., 2017, Gasser et al., 2019 and the just published by Masi et al., 2021.

We extended the introduction and included more references to the pioneering and subsequent research in order to embed our research better. We added a discussion on the hydrological effects of vegetation with reference to Feng et al., in the introduction L96-97.

2. Definition of the stability problem.

I found the definition of the problem of stability estimation (section 2.2, Figure 2) a bit misleading. It is not clear the definition of the volume of soil to which forces are applied. In the method of the limit equilibrium, under the hypothesis that the width of the landslide is sufficiently large so that the deformations are in the plane parallel to the soil thickness Hsoil (i.e. perpendicular to the elliptic *landslide in figure 2), forces are assessed by considering a 'slice' of soil with unit width (in the direction parallel to the elliptic landslide plane). Figure 2 is confusing and the planes of forces are not well drawn. The limit equilibrium method (and infinite slope model) is based on the hypothesis of large and elongated element with respect to the soil thickness, so that a unit in width element can be considered. Also, Pwater is not indicated in the Figure 2. According to the definition in the manuscript, Rlat and Fres apply on different planes. I suggest to modify in a 3D perspective the Figure 2 and specify the hypothesis/assumptions.*

We adjusted figure 2 to a 3D perspective in order to enhance clarity on the dimensions, volume and force application planes of the assumed shallow landslide. In addition, we added the water pressure (Pwater) as a subtraction of perpendicular force and emphasize the points or fields on which the forces apply.

3. Hydrology and precipitation.

Here is my main comment. The proposed modeling framework addresses shallow landslides that are initiated by rainfall, which is the triggering factor. The approach used (based on TOPMODEL) is extremely simplified because based on steady state conditions, which do not take into account the transient of the hydrological processes (Chae et al., 2017). The authors declare the limitation of the approach used in the discussion section, but this should be clearly stated soon in the methodology. As correctly written by the author, the stationarity is supposed to be reached within the hour of timestep. Clearly, this cannot be largely verified. That said I arise two more critical issues that are not mentioned by the authors: Under unsaturated conditions, soil (especially fine and clayey soils) exerts a strong water uptake effect due to suction, which leads to an apparent 'hydrological' cohesion. This represent a further limitation of the Montgomery and Dietrich approach that the authors should mention (see, works mentioned in Chae et al., 2017, e.g. Lepore et al., 2013).

We added limitations of the used hydrological approach more explicitly in the methodology L244-249.

In the description of the model application (section 3.4.2) it is not clear how rainfall initiating events are selected. If I understood well, only events of 1hour duration are selected, whose intensity is identified from the Depth-Duration-Frequency (DDF) curve at different return periods (i.e. from 10 to 100 years). Therefore, I guess 10 events of 1 hours are simulated. Is that correct? If so, it should be explained and justified the reason of analyzing events of only 1 hours, which cannot be 'critical' for landslide initiation. Authors should deeply clarify this part in the manuscript, explain the methodology used to define the events, and report the parameters of the DDF curves.

We emphasized the choice for 1-hour events (assumed macro-pore activation time period) in the methodology L256 and the parametrization of the DDF curves in the supplementary material.

4. Data inventory

The proposed methodology used to characterize the hypothetical landslides (extent) is strictly dependent on the data inventory (section 2.3), as also stated somewhere by the authors. However, it is important that the observed landslides used to characterize the model are of the same type, according to the hypothesis of the stability model used and all triggered by rainfall. Is it so? Please specify.

We specified the triggering mechanism and assumed representativeness for Switzerland in the data section L334-335.

5. Calibration/sensitivity analysis

With regard to the best set of parameters, my question is: are the found parameters consistent and realistic?

Consistency in the found parameters is arguable. We suspect equifinality is at play. However, with the data available to us, we believe we made the most realistic assumptions on the parameter ranges. We added a comment on consistency and realism to sensitivity analysis discussion section L540-545, 550.

For example, I argue the choice of including the precipitation intensity as calibration parameter. As discussed in the previous comment, rainfall represents the triggering forcing and it is a dynamic variable. Ideally, we should know the precipitation intensity associated to each observed landslide. Otherwise, if used as parameter, it seems that the model is tuned ad hoc just to reproduce the past events. If so, which could be its utility?

We agree that in an ideal case this should be known, however no detailed information is available. Therefore, we have to rely on the more simplified steady state approach. We added the lack of detailed information in the data section to justify our approach L335-336.

Additionally, it would be interesting to see the AUC curves for the calibrated and the best model combinations. The shape of the curve also tells about the model performance. Then, to my opinion, sensitivity analysis should go before the model calibration. Normally, calibration is done on parameters that are more sensitive. I understand figure 7 and 8, but not sure this is the most efficient way to verify the sensitivity of the parameters. I am curious to see how, for example, the landslide probability varies with the chance of parameters values. This test could be shown with the least and most sensitive parameters.

In our opinion the sensitivity of all these parameters is interesting and can help in future development of SlideforMap or other models employing a similar method. We specified this choice in the description of the sensitivity analysis. As suggested by the reviewer, we added the corresponding AUC curves to the results.

6. Results

The result of high m_f and low m_c is quite obvious; as the author clearly say in the discussion, and as found by other past works, in the end only few parameters really affect the process: the geometry of the slope (i.e. the soil thickness), the mechanical properties (i.e. friction angle) and the characteristic of the trigger (i.e. precipitation) whose effects are controlled by the soil transmissivity. With regard to the vegetation: different vegetation scenarios are analyzed (and this is fine). which is the real configuration? Which is the ultimate target of the simulations?

The ultimate goal is to assess forest (management) scenarios on slope stability. The real configuration is the single tree detection method. This is emphasized in the introduction L124-128 and conclusion L594-596 of this new manuscript.

7. General

I suggest to clearly state which is the ultimate main target of the model. Can we use it as forecast tool in an early warning system? If so, in which way? My impression is that it is too constrained to the calibration parameters, which, in some cases, may lose their physical meaning.

The main application the authors intended to model for is as a tool to quantify the effects of different vegetation scenarios for land managers. We state this more clearly in the introduction L124-128 and conclusion L594-596 of this new manuscript.

TECHNICAL CORRECTIONS

Abstract: I strongly recommend to reduce the abstract to make it more concise.

We reduced the abstract to the best of our abilities.

L3: I do not completely agree with this sentence given that there are of works that take into account the effect of vegetation, although from different perspective such as the hydrological one, together with the mechanical one. Please remove this sentence from the abstract, where you do not have room to discuss.

The sentence has been removed.

L72-L80- I suggest to synthesize.

This part is vital for our assumption of macropore flow dominance and the 1-hour rainfall event. We tried to synthesize to the best of our abilities though.

Figure 1: it is useful and appropriate. However, consider to improve it to make it clearer. Not clear from where to start. “extract mean value for each landslide”: do you mean hypothetical landslide? Emphasize the ‘append’ box where everything converges. Avoid text outside from the box. Also, I suggest to use the symbol used in the section (instead of the description). For example: definition of rho_ls; it would improve the correspondence with, for example, section 2.3.

Figure 1 has been improved along the suggestions of the reviewer.

Line 417: Not clear which single rainfall event you refer to. I understand that the database include landslide triggered by different storms across the years.

You are correct. We gave the wrong impression. We mean a single rainfall event per landslide. We corrected the text at line L457

Lines 378-379: it is not really clear the procedure. Please try to write more clearly.

We tried to write this more clearly L405-406

Lines 385-386: I understand the reference, but please give an explanation also here, based on your results.

This has been improved (L415-417)

Section 2.7: how did you define the threshold from daily to hourly??

By dividing by 24. This has been updated in L266.

Response to Reviewer #2

The authors describe a probabilistic model called SlideforMap (SfM) which generates a map of shallow landslide probability across an area of interest. (..) The authors have organized their manuscript well, and they have described a complex workflow in a straightforward way. They also build a convincing case for the utility and need for a model of this type, and the described case studies illustrate the applications well. In my opinion, this manuscript should be published in NHESS after some clarifications and revisions. Most of my criticisms are focused on areas where the authors need to provide additional clarifications, either to adequately explain their approach or to explain how this model could be used by others.

Thanks for the summary and the positive assessment. The revised paper contains more details on how the model could be used by others.

SPECIFIC COMMENTS

1) The authors say that their model demonstrates the importance of root reinforcement on shallow landslides, but the authors need to define what "shallow" means so that it is clear where their conclusions apply.

We stated the definition of shallow landslides to which SlideforMap applies in the introduction and conclusion.

The authors are persuasive about the importance of root reinforcement in modeling landslide hazards, but they do not provide much discussion of how this model compares to other previously published models, including both related models (such as SOSlope or SlideForNET) or other models that compute landslide susceptibility on a regional scale. Some additional discussion of where this model fits within the context of other landslide susceptibility models generally would be helpful for prospective users.

We added a paragraph explicitly comparing SlideforMap to other landslide susceptibility models in the discussion Section 5.6.

In describing the methodology, the authors are not always clear about which values are assumed for their own case study, and which values are fixed in the model. For example, at a number of places in the methodology section, the authors assign values and limits on parameters (e.g., maximum HL surface area, mean tree density, precipitation intensity threshold, etc.) based on data from Switzerland (where the case study is located), but it is not clear whether a given user would have the freedom to change these values.

Future users have the opportunity to change these values and are encouraged to do so if they apply SlideforMap in other areas. We made clear in the revised version which parameters are specifically selected for Switzerland (Table 1)

The structure of the model requires that soil depth, soil cohesion, and the angle of internal friction be modeled as random variables with normal distributions, but the other 16 parameters are assumed to be deterministic. The authors need to explain why these three parameters specifically were chosen to be random variables. For instance, variables can be randomized when the uncertainty in their values is either shown or assumed to have the most significant effects on the results. This is suggested somewhat by the sensitivity analysis for the case of soil cohesion and soil depth, but this choice is not explained explicitly.

To summarize, this choice was made because the soil depth, soil cohesion and friction angle vary in mountainous areas and are suspected to be sensitive parameters. We added this choice explicitly in the methodology (L186-189) on the soil parameters.

The authors make use of two datasets, a tree inventory and a landslide inventory, in their analysis. However, they do not spend much time explaining how a prospective user would apply this model if they were lacking these datasets. It seems that users could still apply this model without these datasets, either by creating synthetic datasets or assuming specific values for the parameters that would be derived from these datasets. Providing some more guidance on applying the model without these datasets this would make the model more accessible to users.

Synthetic parametrization is possible for users lacking certain data/datasets. We made this clearer in the methodology L181-184, 270, 300 of the paper.

The sensitivity analysis is interesting but not entirely convincing. If strong parameter correlation is at play, as the authors suggest, then how would we know which parameters are truly important?

What we intended to say in the original manuscript is that a *potential* correlation between parameters can lead to apparent absence of sensitivity. As an example, we used the paper by Bardossy (2007). We wrote this example more explicitly in the new manuscript L415-417.

In a couple of places within the text (L49-52; L169-170) the authors conflate deterministic models with spatial homogeneity. This is misleading, as it is possible to have deterministic models that account for spatial heterogeneity, and probabilistic models that are spatially homogeneous. I would suggest that the explanation the authors are after is that the spatially heterogeneous values themselves are uncertain, and this is the motivation for using a probabilistic approach.

We stated explicit definitions of both deterministic and probabilistic modelling in 40-50

Is it valid to compare the globally uniform vegetation scenario to the other three scenarios if the globally uniform scenario was used to calibrate the parameters?

We acknowledged this fact and use it to make our case for single tree detection in the discussion L520-521.

It appears that the authors used the same landslide inventory to both calibrate the dataset and to validate the performance of the model against different vegetation scenarios. Did the authors consider using any portion of the landslide inventory as an independent validation dataset?

We wanted to analyze the performance of the model, not do a validation. For a proper validation we think the size of the dataset is too limited. We made this clearer in the methodology L350-351.

L44-45. The authors need to give some additional definition of a deterministic approach and why SHALSTAB is an example of this approach.

We gave a better description of models with similarities in the introduction L40-50 including a better definition of deterministic models.

L128-130. It seems that the unstable ratio is a very limited metric, particularly if the landslide density is already very low. Shouldn't the landslide density be relevant in addition to the unstable ratio? If there is an explicit requirement that the number of HLs be large enough to compute the unstable ratio with a large denominator, does this effectively put a lower bound on the landslide density for this model?

We choose the AUC as main metric since it is a performance measure to the historical landslide inventory. We emphasized this choice in the methodology L393-394. We are not sure on the lower bound of landslide density. We stressed this in the discussion L576.

L152-153. I am surprised that the landslides are generated using a spatially uniform distribution, as this may result in landslides being simulated in areas that are not landslide prone. What is the rationale behind this? Shouldn't they follow a spatially distributed density, or at least be restricted to susceptible areas?

In order for comparability of our results within the study area, between the study areas and with other model, we decided to keep the landslide density constant through a heterogeneous study area.

L278. A 2km buffer seems extremely large, especially if topographic wetness is computed over multiple small catchments. How was this value chosen, and is it adequate for other studies?

The value is arbitrary. Model users are free to have a smaller buffer if they are confident it gives good results. We added this in the methodology L305.

L407-408. What does this mean if the unstable ratio decreases when single tree detection is used? Does this indicate that heterogeneity is important for slope stability, or does it simply mean that the uniform vegetation scenarios are not realistic?

We added some possible explanations in the discussion section 5.4.

Table 7. Why are the AUC and Unstable ratio values different for the globally uniform vegetation scenario compared to the results with the optimal parameters (Table 6)? Is this due to the difference in the landslide density?

We emphasize in the results L444-445 that these are realizations from a probabilistic model.

L471-473. Does this high unstable ratio match with long term observations about landslide occurrence in StA? In other words, is the unstable ratio realistic?

We added a column in Table 2 of the landslide inventory derived shallow landslide density in the study area description and compared this in the discussion L540-542 with the unstable ratio.

L480. This suggests that AUC is a poor choice of performance metric for comparing the three study areas. Are there other metrics which would be better?

In order to compare the results in an easy manner to performance of other models, we decided to stick with the AUC.

TECHNICAL CORRECTIONS

L14. This should be "ratio" instead of "fraction."

Corrected.

L121. Does SfM generate a raster image of probability values?

Yes, we specified this in the methodology L131.

L134. Do the authors mean "greater than 1.0"?

Yes, corrected.

L163. What does "distance of 10" refer to?

Bin size of the histogram. This is specified in the new version L181.

L 271-273. What resolution is the unstable ratio computed at? This is not made explicit here in the paper.

The same resolution as the DEM input. This is specified in the methodology L277 of the new version.

L300-307. What is the spatial format of the landslide inventory? If they are polygons, how are they compared to the unstable ratio map so that the AUC can be computed? Does the landslide inventory need to be converted or rasterized at a specific resolution?

They are rasterized points. This is better specified in the methodology L336-337.

L309. The format for the numbers a,b, and c looks unusual. Please verify that the values and formats are correct.

We verified. They are corrected in our opinion, but we added units to the b and c parameters in accordance with the paper by Malamud et al., 2004.

L312-313. Why are these 11 parameters fixed while the others are varied?

We assume these parameters to be invariable and focus our sensitivity analysis on parameters that are variable and relevant in nature. We specify this in L343-344.

L332. What is n?

Total number of shallow landslides in the inventory with a known USCS soil class. We emphasized this in the revised version L366.

L336-337. Please explain why weighting is being used and how this weighting is calculated.

For representativeness. Weighted according to occurrence. We specified this more clearly L370.

L343. Please explain why the parameter range is using intensity values from different return periods.

We specified this more clearly in the preceding text L378-379.

Table 5. The value for vegetation weight, Wveg, uses a different name and different units than the rhotree in Table 1 (tonne per square meter vs. kg per cubic meter). Is there a reason for this difference?

The tree density (rho_tree in kg/m3) is used in conjunction with the single tree detection to compute a vegetation weight (in tonne per square meter). For the sensitivity analysis parameters in Table 5 we do not use the single tree detection and use the vegetation weight directly with a range of values from literature.

L348. Is 1000 an adequate size to represent the sample space over the 12 parameters used in the sensitivity analysis?

Unfortunately, due to computational constraints it is the best we could do. We specified this more clearly L387.

L360-361. Does this model assume that root reinforcement comes only from trees, and not from shrubs, grasses, or other vegetation types? Is the single-tree detection scenario using the same trees as the tree inventory cited in 3.2?

1st question: Yes, we added a reference to this assumption in the introduction L207-208.
2nd question: Yes, we specified this more clearly L400.

L363. Please verify that the exponent is correct in the expression for landslide density.

We verified and believe it is correct.

Fig. 8. How is “x% best” defined for the unstable ratio?

Best isn’t the correct term, we correct this. Highest in terms of unstable ratio. We corrected this clearly in the paper figure 6.

L403. Do the model runs assume randomization of the three parameters (as in the original model setup)?

No parametrization of the 10 runs is identical. The drawn samples however can vary. We stated this more clearly L444-445.

L508. Are the 12 parameters all included in the 22 original parameters?

Yes, they are, we wrote it more clearly L586.

Response to CC1

This is a really interesting paper that demonstrates the applicability and predictive capability of a new model for shallow landslides to provide a detailed inclusion of the influence of vegetation. The use of LiDAR data to deduce tree properties and thus root characteristics is a really exciting development.

Thank you for the positive overall review.

The model itself is similar to a number of existing models but also makes some important changes. It would be really useful to make these similarities and differences more explicit. The striking similarities to me were: 1) the hydrological model (Eqns 11-12) is exactly that of SHALSTAB (Montgomery and Dietrich, 1994) and SINMAP (Pack et al., 1998); 2) modelling discrete landslides of defined dimensions with lateral resistance due to roots only (Eqns 1-6) follows Montgomery et al. (2000), Schmidt et al. (2001) and Roering et al. (2003); 3) the probabilistic treatment of stability using distributions for parameters follows Pack et al. (1998) who represented c, phi and the R/T ratio as uniform distributions; 4) introducing a slope dependence to failure depth follows Prancevic et al. (2020), though with a different functional form. The similarities are strongest between SfM and Montgomery et al. (1998), they use very similar stability models (both infinite slope with root cohesion only on the margins), the same hydrological model, and both impose discrete landslide dimensions; so differentiating your work from theirs will be important.

SlideforMap is similar in many aspects of the approaches. We added the specific similarities and distinctions to our method section in the revised paper L155, 190-194, 239-245 . A subsequent discussion was added in the discussion section 5.6.

Having read the paper I have one primary outstanding question: What do you gain as a result of the additional data collection and modelling efforts involved in a detailed inclusion of the influence of vegetation?

We expected an improvement in the performance of the model with the detailed inclusion of vegetation, which we analyzed in the paper. We emphasized better as a goal in the Introduction L124-128 and a discussion on the outcome in the discussion section 5.4.

Your paper focuses on predictive skill (using ROC AUC) and predicted instability (using an unstable area ratio). That focus enables a straightforward assessment of improvement in predictive skill from this more complex model relative to a simpler models such as SHALSTAB or SINMAP. In fact, I think you already have an answer to this in Table 7. The 'no vegetation' case in SfM is very close to the SINMAP model: in this case, there is no lateral resistance (i.e. an infinite slope), probability of failure is calculated from pdfs of friction, cohesion and depth with pore pressure predicted using the SINMAP/SHALSTAB model. The uniform vegetation cases (Global and Forest area) are very close to the SHALSTAB implementation of Montgomery et al. (2000): in these cases landslides have predefined dimensions and lateral cohesion is spatially uniform. The difference is that landslide dimensions (area and depth), and material properties (c and phi) are sampled from distributions to generate a probability of failure rather than using the critical P/T as a metric for propensity to failure (as in SHALSTAB). In all these cases I would expect a direct comparison to SINMAP and the SHALSTAB of Montgomery et al. (2000) to yield almost exactly the same AUCs as those from SfM. The clear structural difference between SfM and previous models comes in the case of 'Single tree detection'.

We added the similarities and distinctions between our model scenarios/metrics and the existing models to our methodology.

Reading Table 7 in the context of these connections to simpler early models leads to three conclusions:

- *Landslide predictions are surprisingly (and encouragingly) skilful even when models as simple as the 'No vegetation' SfM (equivalent to SINMAP) are used. Models like SINMAP are very attractive if they perform so well given their simple structure and parsimonious parameterisation.*
- *Representing landslides as discrete features (as in SfM or Montgomery et al. (2000)) rarely improves predictive skill unless detailed vegetation information is available. Best AUC for SfM with 'Global' or 'Forest area vegetation' are equal to the 'No vegetation' case for 2 of the 3 study sites and only 1% better for Sta.*
- *Detailed vegetation information from single tree detection does subtly improve predictive skill but only in 2 of the 3 sites (slightly worse for Eriz) and only by 3.8 and 3.2% in AUC for Trub and Sta respectively.*

One interpretation of this would be that while SfM is much more satisfying from a process representation point of view it offers only very marginal gains in predictive skill and has considerable cost in that it is more highly parameterised and more complex. An alternative interpretation would be that small skill improvements on an already excellent model are worth the additional complexity (and cost). Reframing the percentage changes in AUC as percentage of the unrealised AUC that has been eroded by the new model (thus changing in denominator from AUC_{pre} to $1\text{-}AUC_{pre}$) the same values are: 6% and 43% for Trub and Sta respectively. I think this interpretation, which recognises the diminishing returns in model improvement is reasonable and if so it suggests the improvement is non-trivial.
It is interesting that the unstable ratio metric is more sensitive to model structure than AUC, and perhaps encouraging that this ratio is reduced by improved process representation. However as you point out (L355), this ratio is a measure of instability rather than accuracy.

The authors agree to a large degree with this interpretation and would like to thank CC for this interesting and concise discussion. We added this to our discussion section 5.4 on the vegetation scenarios and the model in general.

***SfM also makes predictions about the size of landslides** most likely to be triggered in each location (though these are not currently reported in the paper). This is an important difference from previous models. Few models have done this before and those that have are extremely computationally expensive. Therefore the most exciting aspect of SfM to me is its ability to predict landslide size. The authors are clear that the model requires a prior distribution of landslide sizes but this does not prevent SfM from producing useful information on landslide size both in global/lumped terms and spatially distributed terms. In lumped terms, you could compare the size distribution for triggered landslides with the prior distribution. In the current case the prior is the observed size distribution for the study area but you could equally impose a uniform prior and assess the extent to the posterior matches the observed approaches both would be informative. In spatially distributed terms, the pattern of landslide size and its relationship to local conditions would be interesting and you would also be able to assess model performance with respect to landslides size by comparing the areas of predicted landslides that overlap observed landslides and (do they correlate? What is the form of the relationship?). Perhaps this type of analysis is reserved for a later study but it would fit nicely in the current paper.*

We decided, as suggested by CC, to not include this (though interesting) analysis in this manuscript and save it for a later date.

Beyond these three major points I have several other questions that are more specific but less important. I don't expect any of them to alter the primary messages of either the paper or the points I raise above but I hope they might be useful for the authors during revision. I do not understand the rationale behind some of the assumptions in SfM's boundary resistance representation

- *Neglecting lateral earth pressure. It is true that active and passive earth pressure are maximised at some strain but neglecting them on this basis leaves two problems: a) you still need a treatment for the forces acting at the head and toe of the landslide; b) you need to apply the same criteria to root reinforcement since this is also maximised at some strain.*

In the revised version we specified the phase in the landslide we assume and better explained the resulting force balance in the methodology L152.

- *Neglecting soil cohesion on the sides. It seems inconsistent to apply root reinforcement but not soil cohesion on the lateral boundaries if you apply both on the base*

Like the answer above, we explained this better in the methodology.

- *Lateral root reinforcement acts only over the upslope half of the landslide's perimeter (Eqn 3). I don't see a justification for this and Schwarz et al., 2010 point out that it underestimates lateral reinforcement.*

Like the answer above, we explained this better in the methodology.

- *Lateral root reinforcement in Eqn. 9 is depth independent. This seems inconsistent with observed depth dependent rooting (density and size); and the depth dependence of basal reinforcement in SfM (Eqn 10).*

You are right. We integrated the RBM root probability density distribution and added this as a correction factor to the lateral root reinforcement. This is described in the methodology section 5, equation 8 and the results are recalculated.

- *Calculating root reinforcement using spatially averaged distance to trees within the Gamma function. Previous applications of the Gamma function (Eqn 9) appear to use it to predict root reinforcement at a known distance from the nearest tree (Moos et al., 2016). Given its nonlinearities, is it reasonable to use an average distance in Eqn 9 rather than evaluating Eqn 9 for the distribution of distances then averaging?*

For the current paper, we decided to keep this methodology as it is.

Variability

The form amplitude and spatial pattern of variability in material properties are all likely important in defining landslide location and size (e.g. Bellugi et al., 2021). Representing this variability seems important. I would have liked to see more detail on your rationale for your choice of distribution form and spatial (de)correlation. I recognise that observations to inform this are sparse and these properties are not well known. The normal distribution has some specific problems that you grapple with but that others chose to avoid by using a log-normal (e.g. Griffiths et al., 2007). You deal with *unphysical negative values by truncating, and claim these are rare but this places strict constraints on the variability that you can impose (small coefficients of variation for soil depth and cohesion in Table 5). In the absence of evidence to the contrary, a distribution that is limited to positive values (e.g. lognormal) would seem a more appropriate choice.*

We agreed that the log-normal distribution is a more appropriate choice. We applied this in the paper, section 2.4, in a recomputation of the results. In addition, we added a comparative figure in the appendix.

Soil depth variability is treated slightly differently (spatially de-correlated but slope dependent). I was unsure whether soil depths distribution was parameterised from observed landslide scar depths (L178) or using mean and standard deviation as parameters to optimise (Figure 7). The former seems problematic: landslides likely occur in deeper soils biasing the sample. Perhaps Eqn 7 was designed to account for this? However, I don't understand why the coefficients on mu (1.35) and sigma1 (0.75) in Eqn 7 have these particular values. The second approach, tuning mean depth rather than setting it from observations seems more appealing to me and would also enable a comparison between model results and observed landslide depths, which would be a nice addition.

We adjusted, pointing out that this tuning is optional L203-204.

Hydrology

Your approach is exactly the same as that of SHALSTAB and SINMAP but is considerably different from Topmodel (Beven and Kirkby, 1979). All three use a topographic index to define hydrologically similar units. Topmodel uses these (with simple treatments for evaporation and infiltration) to simulate a time-varying catchment averaged response to a rainfall timeseries that can be mapped back onto the HSUs; the others simply solve for a single steady recharge rate (neglecting these processes). Even the topographic index (i.e. A/sin(B)) differs from that of Topmodel (which uses ln(A/tan(B))). This reflects differences in reference frame (the sin vs tan) and assumed conductivity profile (uniform vs exponential). I don't disagree with the approach but I think it follows Montgomery and Dietrich (1994) and Pack et al. (1998) so it would be simpler to say that. If you wanted to give credit to earlier work then the TOPOG model of O'loughlin (1986) was behind the original derivation of SHALSTAB and the first introduction of a topographic index was by Kirkby (1975).

We let go of the formulation of using TOPmodel or TOPmodel assumptions and gave explicit credit to O'loughlin (1986) and Kirkby (1975) in the methodology L239-240.

Previous papers that apply this hydrological model do not claim that it is particularly well suited to slopes with macropore flow. Montgomery et al. (2002) highlight the importance of macropores and fractures (and a steep soil water characteristic curve) for hillslope hydrologic response but also recognise that "that rapid pore pressure response that controls slope instability [...] is driven by vertical flow, not lateral flow" (Montgomery et al., 2004). ***There is general agreement that lateral flow (modelled here) strongly influences the pore pressure field antecedent to a burst of rain that could initiate a landslide*** *(Iverson, 2000; Montgomery et al., 2002; 2004). This has important implications for the approach though because it implies that Q/T is an index for the 'propensity for landsliding' rather than a parameter to be calibrated within a complete hydrological treatment. This* *explains the apparent problem of predicted pore pressures independent of rainfall duration but observations that landslide triggering depends on both intensity and duration. Broad spatial patterns of pore pressure and instability should be well captured but triggering rainfall properties may not be. In fact, discussion of the influence of macropores on pore pressure tends to focus on the unpredictable localised pressure peaks associated with constrictions or terminations to macropores (e.g. Pierson, 1983; Montgomery et al. 2002). Even given these limitations I don't think this is a bad model relative to the alternatives because it captures broad phreatic surface patterns and I'm convinced that the finer detail of these patterns is set by (unknown and perhaps unknowable) heterogeneity in material properties (e.g. macropores). If so, a more refined and expensive hydrological model may improve predictions of spatial pore pressure patterns very little.*

We updated the introduction L84-86 and methodology section 2.6 stating our assumptions, limitations and similarities to our model more explicitly.

Sensitivity Analysis

As you point out parameter interaction makes it very difficult to infer parameter sensitivity from Figure 7 I think that may make it difficult to support some of your assertions in L388-395 because you cannot guarantee that interactions are not masking other stronger sensitivities. For me the clearest example is the interaction between P and T (Table 6). Both are listed as uncertain parameters within the sensitivity analysis but only feature in pore pressure definition and only in that equation as the P/T ratio. As a result their inclusion as two separate variables in this analysis is likely to lead to severe equifinality (with high or low values will result in the same outcome as long as P/T is constant). Why not include the ratio of the two in your sensitivity analysis?

We added P/T ratio to our sensitivity analysis results and discussed the equifinality L546-547 that appears to be at play.

Queries on equations:

1) I think there is a dimensional problem in either the first term of Eqn 3 or the second term of Eqn 4. Eqn 10 expresses R_{bas} as a function of R_{lat} so I think both should be either a force per unit length or a stress. If R_{lat} (in Eqn 9) is a stress then Eqn3 is dimensionally incorrect because the first term is a force per unit length and the second a force. The first term needs integrating over landslide depth. This could take the form cos(s) H if you assume reinforcement is depth invariant. However, this would then be inconsistent with Eqn10, which assumes that root reinforcement declines with depth. On the other hand, if R_{lat} is a force per unit length (which might be more consistent with Moos et al (2016), Fig 3) then the problem may be more difficult to solve because the lateral depth integrated stress (N/m) is being applied across a basal area (m^2).

Indeed, it uses a dimension correction factor, we added this to Equation 9.

2) Are h and H measured in a vertical reference frame as indicated in Figure 2? If so then I think there is a cos(s) missing from Eqn 12. The first cos(s) converts vertical depth to slope normal thickness, the second converts phreatic surface thickness to pressure head (under assumptions of: uniform steady slope parallel seepage).

You are right, we corrected this and recalculated.

3) Eqn 15 is incorrect because the original equation calculates DBH in cm from tree height in metres (Dorren, 2017) but you use DBH in metres (L292). I think Eqn 15 should be adjusted to $0.01H^{1.25}$.

You are right, we corrected this.