The authors are grateful to the anonymous reviewer. We will endeavor to respond appropriately below.

[*In general, there is serious issue with respect to quantitative assessment of reliability of the modeling results in terms of calculated landslide volumes, depths and their run-out limits. All of these parameters are very useful and can be applied for practical hazard assessment, but the user needs to know how reliable the model outputs are. This is so far characterized mostly by qualitative, general statements*.]

We respectfully disagree, but we certainly acknowledge that model reliability includes a high degree of expert judgement. The professional must decide whether modeled landslides travel along realistic paths, whether the paths are similar to those of historical events as mapped or as observable in the air photographs, whether the range of deposition and erosion approximates similar events in the same region, and finally, analytically, whether or not the magnitude frequency and area-volume characteristics are sufficiently similar to mapped characteristics, or justifiably different.

Because LABS is both predictive and probabilistic, it may not precisely recreate an existing or historic landslide, but instead tries to credibly produce predictions of landslides that may occur on the existing surface. It's also not a susceptibility model so we don't expect to conduct the type of reliability testing that we normally see (and need) for that type. Our best quantitative calibration tools are the M-F comparisons, and visual comparisons of landslide runout to mapped landslides and geomorphology (more below).

We think that the predictive and probabilistic aspect of the program is a strength, and we include the ability to model many landslides to compare the range of responses between runs. As it happens, we *have* used detailed historic landslide studies to calibrate the current predictions, however, these are the subject of a different paper.

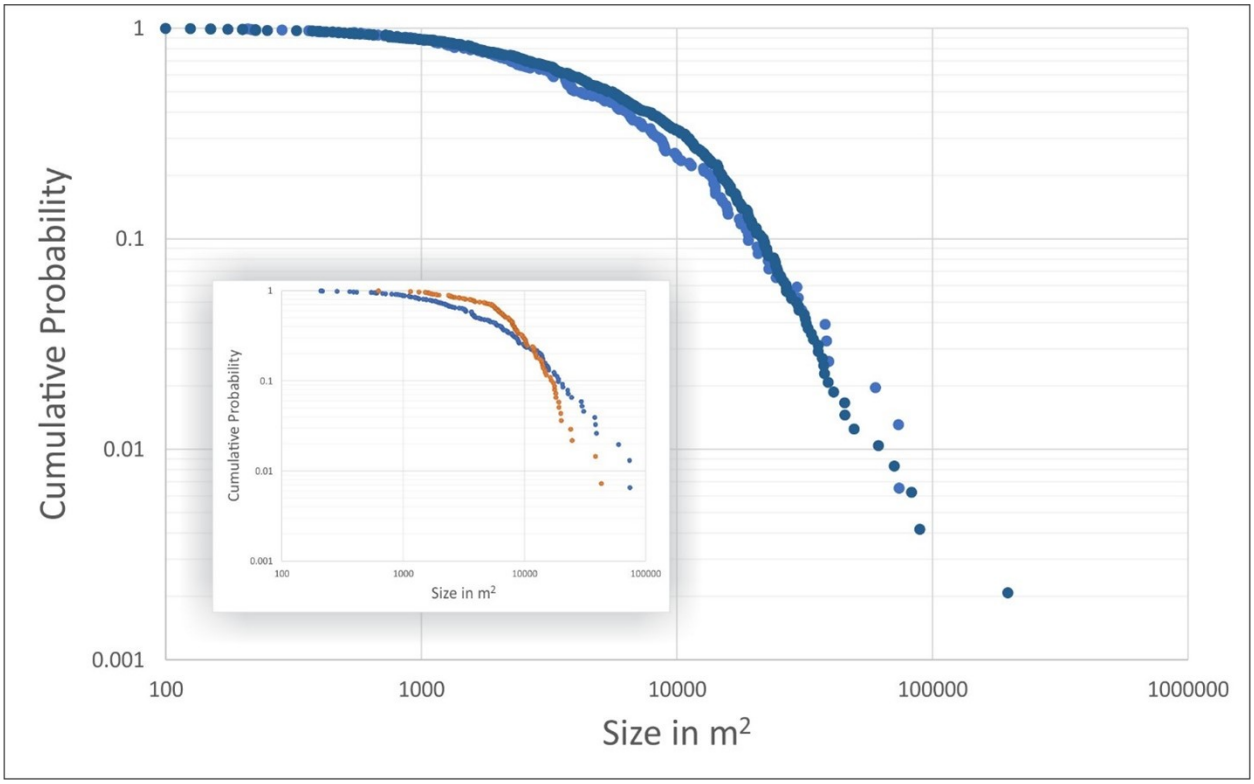
Model calibration is completed iteratively using the controls within the program. The landslide professional runs the model and compares the results to mapped or historical landslides and ground-based evidence for travel distance, scour and deposition. Several methods may be employed including a visual comparison, quantitative comparison of magnitude-frequency of mapped versus modeled results (Figure attached) and volume-area relationships or simple landslide length comparisons.

The "Inspect" tool allows the user to examine the results (including depth) pixel by pixel and the "One By One" advances individual agents through single time steps allowing for a much more detailed analysis of results. These results can be compared to known ground investigations.

Typically, adjustments are made to the control sliders until better results are realized. This might require several runs. Control sliders adjust the shape and spread, and the volume eroded or deposited in each timestep. Note that the volume controls are new since the manuscript was submitted. The attached figure shows the difference between a poorly calibrated result and a well calibrated result using M-F analysis.

[*I also see serious problem with respect to maps you presented in the manuscript. The maps you show lack legends and geographic coordinates, which needs to be added to allow the user to get all information they show.*]

Agreed. We will update the maps.



Example of a M-F graph of modeled and mapped landslides from a well calibrated model run and an earlier poorly calibrated model run (inset).

Marginal comments provided by the reviewer in a markup attachment were largely either editorial, or along the lines of the question above. The questions are generally reasonable, and the authors can certainly update the manuscript to clarify. A couple of comments/questions were unique and we are adding those below.

[*Please add reference which would provide definition of "debris floods" as this would significantly contribute to better understanding of the topic.*]

Agreed. We propose to introduce the recent Church and Jakob (2020) reference.

Church, M, and M Jakob. 2020. "What is a debris flood?" *Water REsources Research* 17.

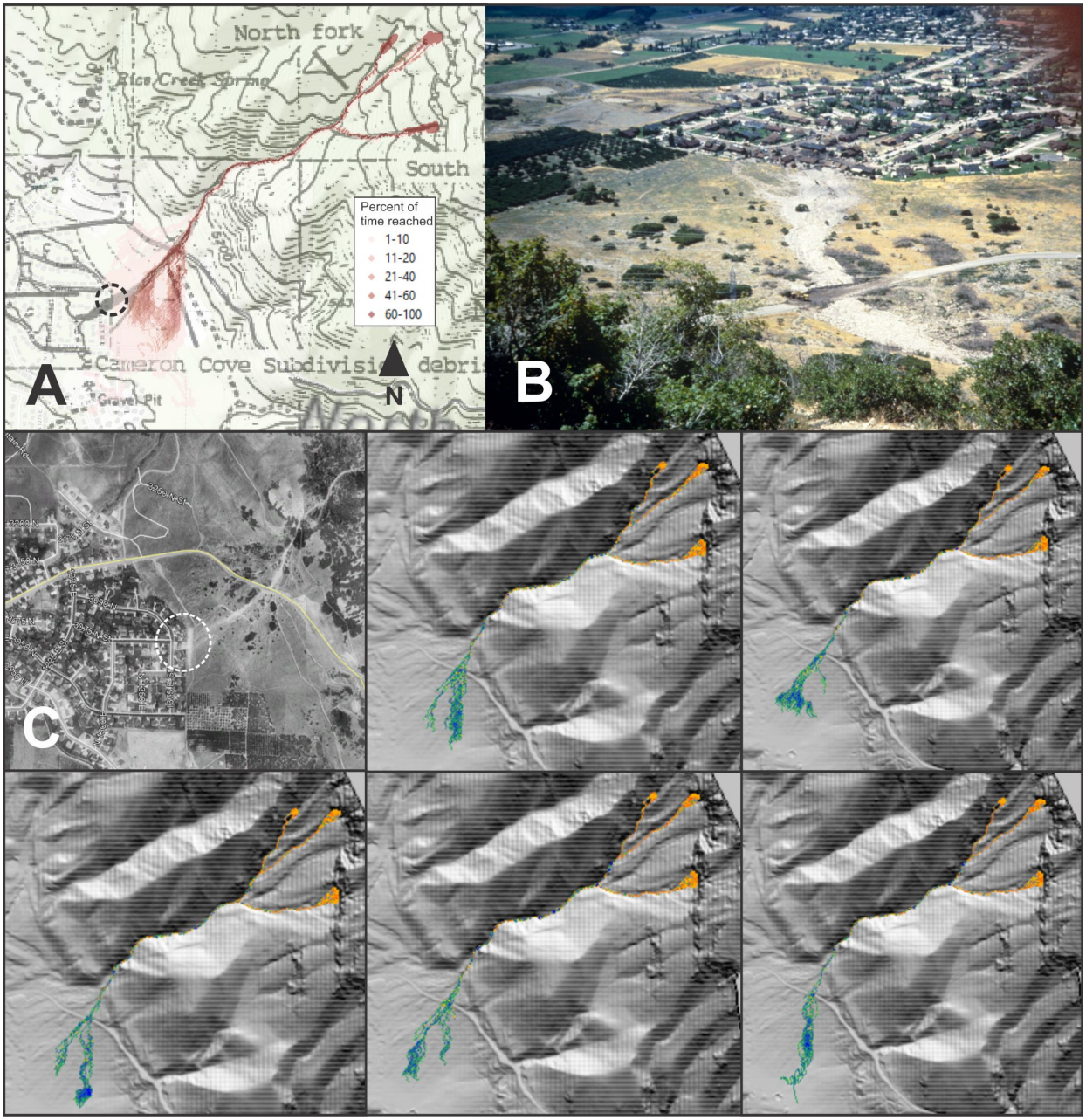
[Several comments related to quantifying observations instead of using qualitative adjectives].

Agreed. This is, by and large, a reasonable request and we will update the manuscript accordingly.

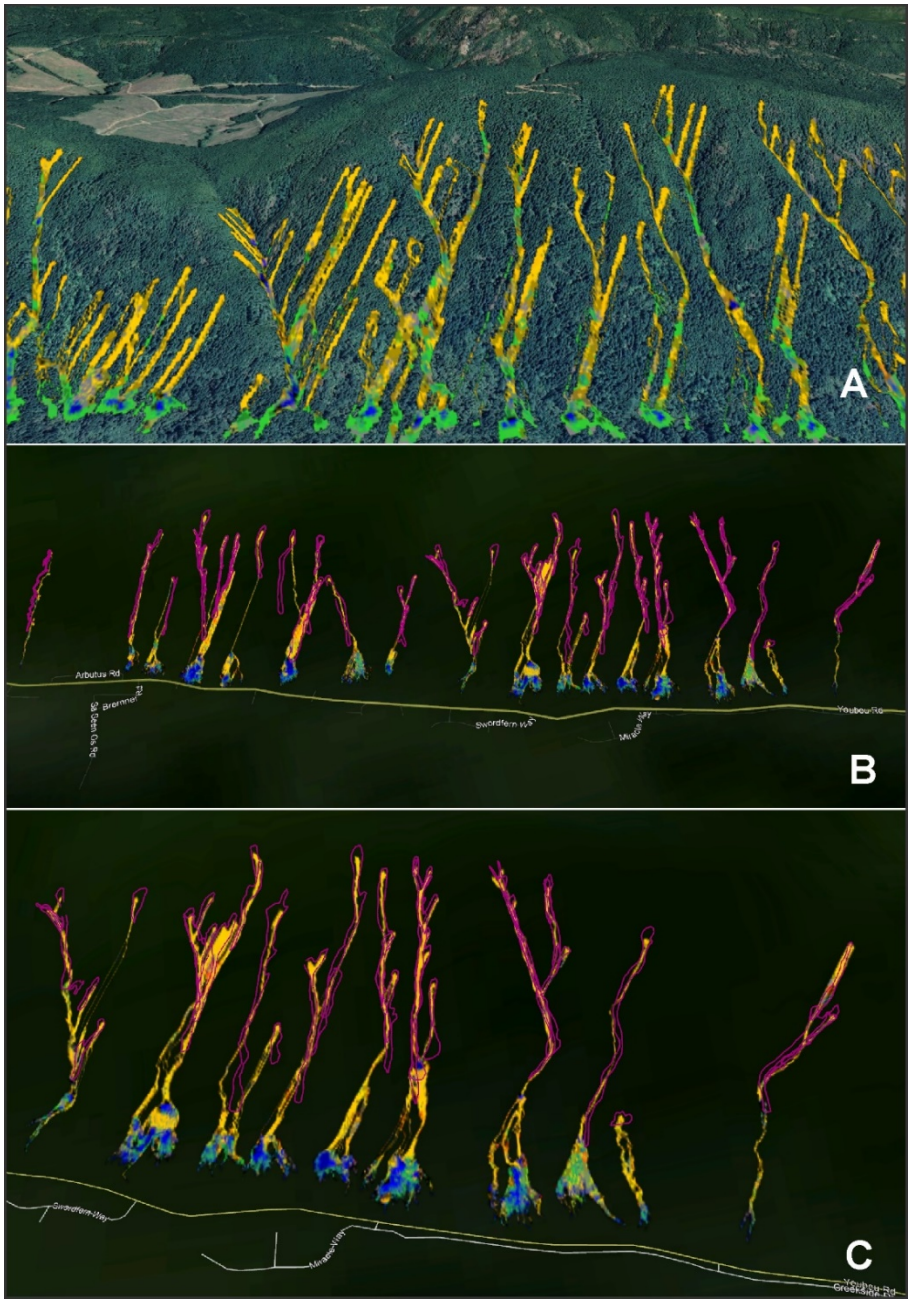
[*This fact* [that distal margins of landslides tend to be inundated less frequently than the main landslide body] *along with characteristics mentioned in the preceding paragraph can be serious limitation of the model if we would search and answer where it is safe to build houses with respect to the expected run-out.*

Could you quantify or describe in more quantitative manner the uncertainty related to the margins of the modeled run-out?]

We consider this to be a strength of the program. The fact that we can run a simulation multiple times and get what we believe is credible variation between runs allows the user to better estimate the potential footprint as in the following example. LABS allows you to show both the overall footprint and the most likely footprint for a specified topography (the current DEM).



[*Please explain … narrow shape of the transportation paths. It seems that some problems with DEM could be involved! Please, check it.*]
The narrow linear shape of the transportation paths and the potential of DEM error is considered in the paper. It is almost certainly a limitation at the DEM scale, however, it is also consistent with the actual mapped landslides. (Figure attached).



Strong linear orientation of modeled landslides on the North Shore when hundreds of landslides are viewed at once (A). The results look more reasonable (though still linear) when compared to just the mapped landslides (B) and (C). Google Earth image in the background of (A).

[*I think that* [the runout] *probability also largely depends on the initial volume of the material. Please consider this in your conclusions. It would be also nice if you may show calculations where the initial volume of landslide mass was larger*.]

Runout does indeed depend on the initial volume, as well as the difference in available entrainment along the landslide path. The professional landslide specialist needs to consider these criteria when calibrating the model. The latest version of the model can increase or decrease the initial volumes, and the scour and deposition to match the geomorphologically interpreted criteria.