Author's response to referee comments of Bernard et al. 2020 by David Milledge.

Thomas G. Bernard, Dimitri Lague, Philippe Steer
Univ Rennes, CNRS, Géosciences Rennes - UMR 6118, 35000, Rennes, France

General response

We first thank David Milledge and Alex Densmore for the positive reviews and for the relevant and detailed remarks they raised that greatly helped to improve the manuscript. Following their comments, the previous manuscript has been subject to important revisions. In essence, our results are not changed, but we believe the MS now makes a more compelling and quantitative case for our previous conclusions. We propose here to give a general overview of the modifications brought to the manuscript but please see the rest of the document for a detailed answer by comment.

1. We clarified what we consider through the generic term of "landslide" which corresponds to the spatially coherent changes of several decimeters detected by our method on hillslopes. This definition has been added in the introduction of the manuscript. We have added a dedicated section that detail the different processes we observed in the discussion (section 5.1.2, L595).
2. Although it would be interesting to compare our landslide inventory to field observations, such data are not available for the studied area. However, we added a detailed comparison with a landslide inventory manually mapped from pre- and post-EQ aerial imageries (section 3.4 and 4.2, fig. 7). Given the impossibility of having the landslide extent mapped by Massey et al. (2020), due to the author's lack of response to our numerous requests, Dimitri Lague carried out a manual mapping. We used this new manually mapped landslide inventory to: (1) compare the two methods in terms of number and area of mapped landslide, (2) detail the origin of under-detection of landslides from both methods and (3) discuss the impact of landslide under-detection on landslide volume estimation using traditional approach. Following this analysis, the discussion now extensively discusses the origin of a rollover in the pdf(Area), and we provide a new analysis showing a systematic size dependent under-detection of landslides in the 2D inventory (fig. 12). This section also replaces the section about reactivated landslides present in the previous version of the manuscript for which we agree that the term "reactivation" was misused.
3. To further improve our treatment of potential errors and false detection, we have added two additional tests to our workflow and have added many elements on the discussion regarding current limits of the workflow. First, we explore in details the source of errors coming from misclassification of the original point clouds and imperfect flight line alignments. We therefore reprocessed the pre-EQ point cloud to remove incorrectly classified points and

quantified residual errors for each flight line alignments (section 3.3.1). We now integrate the residual error from flight line misalignment in the estimation of the registration error. Second, we added a new analysis to deal with false detections after segmentation (resulting in a new step in the workflow, section 3.3.4). We define a new confidence metric for each segmented cluster (source or deposit) called mean signal-to-noise ratio (SNR). An optimal value of SNR, is defined to minimize the proportion of potential artefacts (new fig. 6 and 5b). This results in more than 746 sources and 748 deposits of low confidence removed from our inventory. The final inventory now contains 524 sources and 304 deposits of very high confidence (in the first version of the MS, there were 1431 sources and 853 deposits).

4. We have added further analysis of the impact of the main parameters of our workflow (registration, segmentation and SNR filtering) in the discussion, appendix and supplementary material. Our results, and in particular the lack of rollover in pdf(A) are not significantly affected by these parameters.
5. We have tested segmentation approaches based on density based clustering (DBSCAN and derivatives) which are considered state of the art for 3D rockfall segmentation. We demonstrate that these approaches (1) do not perform better than the segmentation approach we use and do not generate significantly different pdf(A); (2) may have the tendency to oversegment intermediate landslides into small ones or remove border points of large landslides; (3) are sometime much longer to run (45 min vs 3 sec) and have parameters which are not intuitive to set. Because the method part of the article was already dense, the result of this comparison exercise are in the supplementary (supplementary material S5), along with an explanation as to why density based clustering is actually not suitable to segment large landslides in our workflow. As in the first version of the article, we clearly state in the article that our segmentation has limitations, but we are not aware of other approaches that would perform better in segmenting complex 3D objects covering 3 to 4 order in size.
6. The balance between the supplementary materials and the manuscript has been revised, and many figures have been updated following the reviewers recommendation.
7. The title and abstract have been updated to feature the notion of rollover in the pdf(A) of 2D inventories. It reflects the additional data and analysis we provide that clearly demonstrate, at least in our study case, that the rollover in the pdf(A) observed in our 2D inventory but not in the 3D one, is due to a size-dependent under detection in 2D. We believe this is of critical importance for the geomorphic community working on landslide science.

We wish both reviewers a pleasant reading and thank them in advance for the time they may spent in evaluating this new version of the manuscript.

Summary

Major Comments from David Milledge

We really thanks the reviewer for its numerous and detailed comments. Many of them represent area of improvement for the workflow that we have addressed. Others are beyond even an extended revision of the paper and represent what should be an ideal, fully automated 3D workflow for landslide inventory creation backed by extensive field work that we cannot perform. Scientific research is incremental, and we fully expect that our workflow will be improved in coming years by others, as it was the case for 2D landslide inventories. To this end we provide all the elements (code, data) for other researchers to apply the workflow and reproduce our results, or apply the workflow to their data. We also clearly highlights the limits of our workflow. We strongly believe that despite its limitations, it currently represents the state of the art in terms of 3D landslide detection and landslide inventory creation, and that its application in comparison with a 2D landslide inventory shows that landslide under-detection in image based inventories is extremely prevalent in our study area.

MC1) The paper needs to more clearly define: 1) landslides (i.e. what the inventory includes) and 2) what your inventory can and cannot be used for.

Early on you introduce the idea that there are different landslide processes (L43, "process specific") but you don't follow this logic through into your results. Instead, your analysis may contain an implicit definition of landslides as all processes responsible for surface change that cannot be attributed to fluvial processes (L217-9). You certainly need to make this definition of landslides explicit in your introduction.

The introduction needs a much clearer explanation for what you expect the inventory to be useful for. If it is for understanding landslide mechanics then it is essential that you make an effort to distinguish individual landslides on a mechanistic basis (see MC3 on amalgamation). If that is not an expected application of the inventory you should say so, otherwise there is a real risk it will be misapplied.

Appropriate uses of the inventory (e.g. volume estimation or landslide mechanics) depend not only on its purpose but also on entry criteria into, and distinctions within, the inventory. Non-fluvial surface change that might result from earthquake shaking includes: tree-throw, ravel, rockfalls and slides, slow earthflows, rapid soil slides and debris flows. The processes responsible for these surface changes differ from one another to varying degrees. If there is no distinction between them this precludes the inventory's use in analysis of landslide mechanics and therefore prediction. It allows comment on correlation e.g. of volumes and areas of change, but makes it extremely difficult to make any inference about causation. It also opens the work to the criticism that the bulk statistics mask important differences in behaviour between processes. For example: the differing size distributions for rockfalls (where others have reported no detectable rollover) and landslides (where there usually is).

We added the following definition of what we consider as "landslide" in the introduction of the manuscript: "We use the generic term of "landslide" to define the spatially coherent changes detected by our method on hillslopes that result in at least several decimeter erosion (i.e.,scars or sources) or deposition". The discussion now features an entire paragraph (L596) addressing the various type of landslides that can be detected by our approach. The aim of this paper is not to better understand landslide mechanics as we cannot confidently identify the different landslide processes we detect. We are mainly interested by the estimation of co-seismic volume and to overcome issues such as under-detection and amalgamation on volume estimation. The introduction now clearly integrates these two problematics. We also believe that the new filtering approach that we introduce, which results in 3 time less landslide sources compared to the initial MS, results in a much more robust inventory.

MC2) Elevation errors need to be better quantified and more thoroughly discussed

The manuscript needs a more thorough treatment of errors in the topographic data dealing with both: 1) the properties of the elevation errors that you have identified (e.g. spatial pattern, wavelength, covariation with landscape properties); and 2) the possible sources of error.

RE error properties: *The amplitude and correlation length of elevation uncertainty from different sources and how they interact to generate a 2D elevation error field with a particular amplitude and wavelength is really important for this particular application, where landslides are identified by thresholding then segmenting differences. The error appears to have a fairly long wavelength in many areas (tens to hundreds of metres). It also appears to have some aspect dependence. The spatial correlation of these errors is important because you assume uniform isotropic registration errors.*

RE error sources: *It isn't clear to me what you mean by imperfect alignment, nor ICP related errors (L204-5). Identifying errors on Fig 3 and hypothesising the sources from which they derive are useful but need to be discussed in the manuscript as well. You recognise the presence of "internal flight line height mismatch" and indicate that it results in "large scale low amplitude topographic change" (L418-22). They should be introduced earlier in the article with a more complete explanation of what they are and how you found them.*

Error consequences: *It would be useful to say something about the implications of the topographic errors. I can see two implications:* **First**, *incorrectly assuming uniform isotropic errors will result in a confidence interval for identifying significant geomorphic change that is too strict on some slopes (e.g. some aspects) and not sufficiently so on others. This in turn will lead to false negative change detection on some slopes and false positives on others.* ***Second****, change detection false positives will result in false identification of landslide objects or false representation of their geometry. False negatives are equally problematic since they could result in not only changes to landslide geometry but also cluster breakup (biasing the size distribution). These problems are illustrated in Fig 3 where a number of error patches would be identified as landslides by the detection algorithm if these areas were not assumed to be stable. If such false positives exist here it is they likely also exist elsewhere. It is essential that you quantify their impact on your findings.*

Suggested additional analysis: *You could apply your landslide detection method to only the pre-defined 'stable areas' and generate landslide geometries. These geometries and their scaling relationships might indicate the impact of error on your findings (particularly: number and area density, pdfs, and scaling relationships of artefact landslides). If the results are similar to your findings for 'non-stable areas' it would be very difficult to argue that the data support your claims with any certainty (the same results could have been generated purely from topographic errors in the absence of any landslides).*

The reviewer suggested many areas to explore that are extremely interesting, but which, for some of them, would constitute an entire paper by themselves, in particular when it comes to the analysis of error properties suggested by the reviewer. We also aim at developing a generic workflow applicable to a variety of cases for which users may not necessarily perform extensive error properties analysis. Hence, to improve the paper we have worked on two aspects: (i) improving and better identifying errors in our dataset, (ii) defining a new confidence metrics (the SNR) for each landslide source or deposit to filter out landslide with low confidence.

Errors in our dataset: entirely revisiting our data, we identified two sources of errors: (1) Remaining LiDAR point cloud misclassification in forest areas, inducing local topographic errors, and (2) imperfect flight line alignments from the pre-EQ data, inducing topographic errors of longer wavelength. To address the first issue, we first removed as much misclassified points as possible by interpolating a surface and remove outlier points (see detail in supplementary material and section 2 in the paper). To address the second, we estimated residual errors of each flight lines due to imperfect flight line alignments composing the pre-EQ point cloud and defined the registration error *reg* based on the maximum residual error of the flight line misalignments (section 3.3.1. and S3 in supplements). Compared to the previous version of the MS, the *reg* is now 3 cm larger (20 cm vs 17 cm). We also show that only 1% of points are detected as significant change in the stable area, validating our choice of $LoD_{95\%}$,. While our *reg* is considered uniform over the study area, the $LoD_{95\%}$ (eq.2) also take into account the local point cloud density and roughness which are correlated to the presence of vegetation. The $LoD_{95\%}$ is thus spatially variable. In addition, we are aware that, ideally, a spatially variable model for point cloud error and registration would be preferable for each survey and combined into a more accurate and complete form of *LoD* than what the M3C2 approach currently offers. However, in the absence of the position and attitude information of the sensor (e.g., Smoothed Best Estimate of Trajectory file) and raw LiDAR data - rarely available on LiDAR data repositories -, or of dense ground control which is hard to get in mountainous environment, it is currently impossible. We now discuss this in the discussion (section 5.1.2).

Filtering landslides with low confidence with the SNR: To limit the false detection due to these errors, we also defined a signal-to-noise ratio threshold to efficiently remove suspicious landslides (section 3.3.4 and 4.1). This index is based on the mean ratio between the 3D-M3C2 distance and the $LoD_{95\%}$ for each landslide. We provide a way to evaluate the optimal value of SNR by comparing the number of landslides in the database to a case with no change (i.e., two versions of the same data but with different sampling, a test that we now call Same Data Different Sampling test (section 3.2). This SNR filtering removes a very large number of landslide source and deposits (fig. 5b), and in particular long wavelength low amplitude changes that occurred due flight line misalignment in the pre-EQ data, as well as many small landslides in forested region where point density is very low. We provide a systematic analysis of the impact of SNR (and *reg*) on pdf (A) (fig. 11), pdf (V) (fig. S 13) and V-A relationship (fig. S14).

MC3) You recognise that segmentation results in amalgamation but don't quantify its extent or impact

Severe amalgamation can result from automated segmentation of a thresholded classifier. In the landslide maps that you show here (e.g. Fig 2), amalgamation appears a severe problem for the largest landslides. The argument that it can be solved by tuning Dm (as you suggest on L426-7) is unconvincing since two separate landslides can be within millimetres of one another but have different failure mechanisms. You later say that you "cannot resolve the amalgamation" (L511) and that it "is still a potential issue" (L558). I would argue that it is not potentially but certainly an issue. Your figures show that amalgamation is present (perhaps pervasive) in your inventory but its extent or impact is not quantified. The total volume is insensitive to amalgamation but your landslide pdfs and scaling relationships are not.

Amalgamation makes it difficult to use the inventory to understand landslide mechanics and therefore susceptibility or hazard. *If the inventory is to be useful in understanding landslide mechanics then it is essential that you make an effort to distinguish individual landslides on a mechanistic basis. In the extreme case, an individual landslide where subsets of the failed material moved in opposite directions (e.g. Fig 2) is clearly problematic for the mechanics of that failure.*

Amalgamation introduces bias towards large landslides in size distributions. *Power law exponents and even the appropriateness of any power law can be sensitive to landslide amalgamation. Observations such as that on L569-70 that "largest and deepest landslides deviate significantly from this trend" become questionable since these are the landslides most likely to be the result of amalgamation.*

Suggested additional analysis: *If you are to retain your findings on landslide scaling and size distributions, then you must examine the potential impact of amalgamation on these findings. There are methods available that quantify amalgamation in landslide mapping. You should investigate these. You should preferably also address this amalgamation, either by removing affected landslides or ideally, by decomposing amalgamated landslides into their individual failures.*

We agree that the segmentation approach we use certainly do not allow to solve the amalgamation problem, and this is highlighted in many parts of the MS. However, the problem of amalgamation is inherently subjective and plagues all inventories. Our 3D data reveals a level of complexity, and a density of amalgamated landslides which makes the definition of a single landslide in relation to an ideal rupture mechanism extremely difficult. Even the segmentation of the 2D inventory proved to be extremely complex and is entirely not reproducible. Hence, we favour a reproducible approach, even if currently limited, that can be applied exhaustively to much larger datasets, than a non-reproducible one (2D manual mapping) that we now demonstrate misses a very large number of landslides and incorrectly map their contour.

In this new version of the paper, the landslide amalgamation can be visualized with a map of the landslide source colored by individual landslide as defined by the method (section 4.2 Fig.7). Moreover, the comparison between both inventories shows that while 171 of 2D-sources are shared with 3D-sources, it represents 144 sources 3D-sources. This highlight that 25 landslide sources are amalgamated in the 3D inventory (L-. As in the previous version of the paper, we perform a sensitivity analysis of the impact of D_m showing that landslide statistics are not severely affected by this parameter for $1.5 < D_m < 3$. We also explored density based spatial clustering algorithm used in 3D rockfall segmentation, derived from DBSCAN (Ankerst et al., 1999; Martin Ester, Hans-Peter Kriegel, Jiirg Sander, 1996; Tonini and Abellan, 2014) and HDBSCAN (Carrea et al., 2021; McInnes et al., 2017). None of them managed to provide a significantly better segmentation of the largest landslide and are significantly longer to run than the connected component algorithm we use. We now thoroughly discuss about this in section 5.1.2.

MC4) Topographic errors propagate through segmentation to introduce a bias towards small landslides. Existing experiments to quantify this bias are insufficient.

You say on L559-60 that "our approach has the benefits of more systematically capturing small landslides than traditional approaches". However, this is one of my main problems with the paper given potential propagation of topographic errors through thresholding and segmentation. You show (in SI) that: 1) in the absence of any real topographic change your detection algorithm generates artefact landslides of 1-20 m^2 purely due to spatially uncorrelated topographic noise; 2) this noise generates many more small than large landslides; 3) in this experiment artefacts >20 m^2 were extremely rare. However, this does not demonstrate that predicted landslide size is insensitive to longer wavelength topographic errors (known to be present in the data); nor even to short wavelength noise in the presence of longer wavelength surface differences (e.g. real landslides).

***First,** even without any real topographic change (i.e. no real landslides), the size distribution of erroneous landslide-like clusters will depend on the spatial correlation length of the difference errors, which in turn depends on the correlation lengths of the errors in the surfaces being differenced. Longer error wavelengths will enable the generation of larger error clusters. Your figures show that topographic errors are clearly not uncorrelated and you recognise this yourselves (L418-22); nor do the errors appear to have a single characteristic wavelength. This is a hard problem but one that you must deal with if you are to convince the reader that the landslide inventory you have generated is not hopelessly biased by these landslide-like artefacts. **Second,** the problem is not only that clusters of erroneous negative surface difference due to roughness (or other errors) can create artefacts that appear* *to be landslides, but also that clusters of erroneous positive surface difference are collocated with real topographic change (e.g. due to a landslide); these can interfere negatively with real changes reducing the surface difference below the threshold for detection and breaking a single landslide into multiple patches*.

***Oversampling of small landslides is important because** it undermines your most surprising and high impact claim: that rollover reported in previous inventories is due to under detection (L573-4). I am not currently convinced by this claim because you do not exclude the possibility that the lack of a rollover is solely due to detection errors. You need to quantify these detection biases before you can make these claims.*

***Suggested additional analysis:** A "more advanced segmentation" (L431-3) may be out of scope for this paper. However, an indication of the impact of the simple segmentation on object based classification skill is an essential requirement of this paper if it is to retain the current approach to identifying discrete landslides. Two possible avenues could be followed to provide such an indication. **First,** your analysis of topographic changes on the stable surfaces (Fig 3) would allow you to perform the same analysis that you have performed in the SI but using the pre- and post-EQ surfaces for the stable zones identified in Fig 3. This would enable you to identify the size distribution of artefacts that can be generated from topographic errors with a spatial correlation length closer to that for the unstable parts of the study area. This though still does not account for the possibility that the landslide erosion signal itself is altered by the noise (e.g. by disconnecting clusters). **Second,** you could collect a landslide check dataset using independent observations. This might take the form of an entirely independent inventory but should certainly also involve cross-checking to confirm the existence and characteristics (e.g. area, shape, depth) of your predictions.*

We added further analyses to the method to deal with topographic errors and erroneous landslide. We also added an entirely new 2D landslide inventory as suggested by the reviewer. We are now confident that the actual landslide inventory corresponds to real changes.

Please see our reply to the MC2) and MC8) comments for a detail answer.

MC5) Findings that differ from previous work

There are a number of unusual findings that are worthy of comment because they are some of your most interesting and potentially important findings. It is essential though that each is carefully examined and that the critique that it might have arisen due to methodological errors is dealt with head on.

***First**, it is not unusual to identify more sources than deposits due to amalgamation of landslide deposits. However, it is unusual for deposit areas to be smaller than source areas (L281-2), and for deposit depths to be thicker than source depths (L265-70 and L460). These result should be compared with results from previous studies.*

***Second**, the estimate that Massey et al. (2020) "potentially missed around 169,000" landslides (L491-2), i.e. 92% of the landslides is surprising because of its magnitude. However, to be convincing you need to demonstrate that the landslides that you have detected genuinely are landslides.*

***Third**, you identify areas of deposition where there is no upslope erosion (L256 and Fig 4). I don't think these can be real deposition zones but instead must be a consequence of incorrect landslide detection. Their spatial extent and depth distribution would be a useful indication of the precision of the technique.*

***Fourth**, it is extremely unusual that locations classed as vegetated in post-event optical imagery but identified as a landslide by another technique are considered by the authors to be genuine landslides (L347-8). Instead, the presence of vegetation at the location strongly suggests a false positive.*

***Fifth**, you find no statistically significant difference between total landslide erosion and deposition (L459), does that mean that there was no significant mass loss due to landsliding from the study area? If so, the earthquake's role has been almost exclusively to break up the material and redistribute it within the study area. This seems to differ from the findings of previous studies on the influence of coseismic landslides on earthquake mass balance. It would be very interesting to see the elevation change of the mass (perhaps as elevation pdfs for scars and deposits). This might be a valuable contribution on the instantaneous impact of coseismic landslides on mass balance.*

First: contrary to the reviewer experience, this result does not surprise us, in particular when the runout of landslides is not long. The filtered data clearly support this finding.

Second: this statement has been removed, and we now use our own 2D inventory to discuss under-detection and the number of missed landslides.

Third: We now filter landslides by a signal-to-noise ratio (section 3.3.4) and are confident that the actual landslide inventory corresponds to real changes. Some very small deposit areas may not have upslope erosion as we expect deposit are to be easily detectable by our method than source areas (section 5.1.2).

Fourth: we partially disagree with this statement. As now explained in the discussion large landslides that strip out vegetation are obviously mapped in 2D inventory, but small ones that occur on less dense area are extremely difficult to map in 2D imagery as our inventory shows. Moreover, vertical subsidence due to upslope propagation of landslides is entirely missed in forest, while it is detected in our approach. We think the comparison with the new 2D inventory will resolve the reviewer's reserve.

Fifth: indeed, we found no statistically significant difference between the total landslide erosion and deposition suggesting that there was no significant mass loss (given the uncertainty of our method). We also observe that most of the landslide material volume remained on hillslope with only 3 large deposit areas located on the river bed. Although we agree that it would be very interesting to look at the mass balance as suggested, we choose to let this question for another study as the MS is already quite long and dense.

MC6) Areas should be calculated surface parallel

I agree that variation in surface normal orientation could introduce bias in the volume estimate (L238-9) and that volume should be calculated normal to the plane on which landslide area is calculated. However, it isn't clear to me that a horizontal plane is most appropriate for area calculation. This will severely underestimate landslide area on the very steep slopes typical of rockfall initiation. Surface parallel area calculations would be more appropriate and orthogonal measurements for volume calculation would still be possible. I don't find the argument based on retaining consistency with other studies (L279-80) convincing. Your inventory appears to contain a significant number of rockfalls and their area and volume are often calculated in a vertical rather than horizontal reference frame. If you do not alter the reference frame you should certainly report the distribution of surface slopes of your landslides and comment on the effect of vertically projected area given this distribution.

Computing true surface area requires significant new developments of the workflow in relation to segmentation, to better handle complex cases of landslide amalgamation. Surface parallel area calculations is not as trivial as it may seem if it is to be applied automatically over a large number of landslide geometries and configuration. We clearly highlight this issue in the discussion (as we did in the first version of the paper), but have no simple solution to offer (see section 5.1.2). We present the distribution of pre-eq landslide source slopes, and shows that only a very small fraction (1 %) of points haves slopes > 60°. Given that the median of the slope distribution of our landslide source inventory is 34.1, the estimate of the true area gives a total landslide source area of 356,876 m² rather than 286,445 m.

MC7) Reactivation is actually under-detection on bare or sparsely vegetated slopes

I am not convinced that the set of landslides that you are describing are really re-activated landslides (L344-5) but I'm also not convinced that reactivation is really the relevant issue. I think you demonstrate that you are able to detect landslides in vegetation sparse or vegetation free areas that have generally suffered serious underreporting. Some of these areas may be bare due to previous landsliding, others may not but the key point is that inventories have underestimated landslide density due to underdetection in these areas.

However, if you do want to focus specifically on reactivation then you need to clearly define reactivated landslides and detail how you classify them as such. For example: Does a landslide need to reoccur within the footprint of an existing landslide to be a reactivation or is retrogression included within reactivation? Within what time window following the first landslide must the second occur for it to be considered reactivation?

We agree that "reactivation" was not the appropriate definition for what we consider which is more associated with under-detection of landslides. We now describe in details the differences in terms of detection of landslide areas between a 2D manually mapped inventory and the 3D differencing approach in section 4.2.

MC8) Landslide object detection needs to be tested against an independent dataset

The findings in this paper depend critically on the skill with which the proposed method can classify landslide scars and deposits. Thus it is essential that the paper reports testing results that quantify this object based classification skill. At present, "orthophotos were used to visually validate" the classifier (L115-6) but without reporting results of this analysis. I think it is essential that you explicitly explain your sampling and mapping strategy for landslide detection from orthophotos in the methods. You should then include a section in the results where you compare your orthophoto based mapping to the surface differencing approach. However, it is not enough to simply say the orthophoto mapping did not identify landslides that were identified by the surface differencing. You should then go to a carefully chosen (e.g. stratified random) subset of the landslides detected by each method that were not detected by the other (i.e. surface differencing but not ortho-photo mapping and vice versa) to establish as far as possible which of the two methods was in error and why. While finding and mapping thousands of landslides might be time-consuming (L255), confirming their existence and characteristics (e.g. area, shape, depth) would not.

This lack of comparison between the 3D differencing method and a more classical approach has been addressed by adding a manual mapping of landslides based on 2D images. We added a section that specifically explore the differences between the methods (section 4.2). Moreover, the resulted landslide area distribution mapped manually has been added in the figure in section 4.3 and compared with those obtained with the 3D differencing method.

Minor Comments

L32: You need to define landslides early in the paper, it will influence interpretation of your later findings.

We now define landslides as "the spatially coherent changes detected by our method on hillslopes that result in at least several decimeter erosion (i.e.,scars or sources) or deposition" (L 95-96).

L34: "spatial distribution, total volume…"? explaining why these are useful would be helpful because this sets the motivation for everything that follows. Landslide mapping has different requirements depending on its purpose. The purposes you choose here set the metrics against which your own mapping should be evaluated.

This sentence has been modified by: "Both are important for the understanding of landscape evolution and the management of associated direct and secondary hazards ".

L35: "associated direct and secondary hazards" What is the connection between mapping, the first three elements of the list and these later two? It would help to make this explicit. As above, this relates to the purpose of the mapping.

See comment above.

L43: "process specific": here you recognise the importance of different landslide processes. This is an important point that needs to be reflected in your own analysis (see MC1).

Potentially has been added to reflect the current knowledge uncertainties on this question. We also discuss what potential processes we detect with our method in section 5.1.2.

L110&113: It isn't clear what statistic is being referred to here as a metric for vertical accuracy. RMSE? If so it would be better to quote mean error and SDE, giving an indication of the contributions of precision and bias.

The vertical accuracy of both LiDAR data has been estimated from the difference between the elevation of GPS points and the nearest neighbour LiDAR point elevation. While the mean and the standard deviation of the difference are provided by the 2017 LiDAR dataset report (Aerial survey, 2017), and respectively equal to 0.00 m and 0.04 m, only the standard deviations are mentioned in the 2014 LiDAR dataset report (Dolan, 2014) and range from 0.068 m to 0.165 m.

L114: "using the classification provided": More detail is needed on the method used to classify ground points.

The survey report of the LiDAR data only mention that the ground points have been automatically classified using the Terrascan software. The reference to the survey report (Dolan and Rhodes, 2016) have been added to the text.

L115-6: "orthophotos were used to visually validate": The results of this validation are missing from the paper.

This part has been replaced by the comparison with a manual mapping based on 2D images. See section 4.2. The text of this section has been modified to introduce this manual mapping.

L135: "core point": What is the sensitivity of the results to varying core point spacing? The orientation and spacing of the grid of core points should also be described and justified either here or in section 3.2.

We have added a sentence explaining how to select the core point spacing. However, given that it can only varies between 1 and 2 m, we did not deem important to explore the impact it has on the end result.

L136: "first dataset": Which is the first dataset? Are your results sensitive to this choice, both in terms of depth area scaling and disagreement in 1) change surfaces and 2) maps of significant positive and negative change?

The first dataset used in the paper is the pre-event point cloud. To analyse the potential impact of the first dataset to be compared in M3C2, we applied the workflow detailed in the paper with the post-event point cloud as the first dataset and the pre-event as the second one to be compared. We then analysed the results from both versions of the workflow by computing the difference between the 3D-M3C2 distance values obtained on the significant change point clouds. Results show that 98.6% of the significant change point cloud are exactly the same in both cases and that the mean difference between both 3D-M3C2 distance values is 0.001m with a standard deviation of 0.28 m. As these results show that the choice of the first dataset to be compared do not influence the final results and that the text of the method is already long, we did not include this result in the manuscript.

L137-8: This is not clear: do you mean that you calculate the centroid of each point cloud in 3D then take the magnitude of the 3D vector that connects these two points; or that (for each point cloud) you take the arithmetic mean of differences from the reference plane (defined by D) in a direction normal to that plane? In either case you are performing a spatial averaging at length scale d/2 assuming a uniform kernel. First, is it problematic to perform averaging over length scales larger than the core point spacing? Second does it make sense to assume equal weight in the average with plane parallel radial distance from the core point, or should some form of inverse distance weighted average be used? I would have assumed a weighted average was more appropriate but it would be useful for you to explain why an unweighted mean is more appropriate.

To calculate the distance between the two point clouds, the average positions i_1 and i_2 of the point clouds are first defined and then the distance is computed between the two positions along the normal vector. The average positions are defined as the arithmetic mean of the distance distribution of each point of the subset of points (created by the intersection of the cylinder to the point cloud) to the normal vector (or the cylinder axis; see Lague et al., 2013). As this is part of the M3C2 algorithm that we did not modify, we don't discuss the choice of a uniform kernel rather than a non-uniform one.

L139: "if not intercept is found…": This is not clear to me. Do you mean 'if the cylinder does not intersect any points in the second surface? Why would this happen? Does this only occur at the boundaries of the point clouds? How do these intersection failures influence the surface differencing and how do you report them in your later analysis?

In LiDAR datasets, the density of points is non-uniform over the entire point cloud. Consequently, missing data or very low point density (<5 pt/m²) can occur inside the point cloud due to the absence of laser impact on the ground during the data acquisition. This mostly occur in dense vegetated areas or water surface areas (for topographic lidar). When performing M3C2, it is thus possible that the cylinder cannot intercept points or just a few (< 5). In both cases, the M3C2 distance will not be considered significant. In areas with low point density (<5 pts/m²) a solution is to perform M3C2 with a larger projection scale *d* to include more points in the distance and statistic calculations.

L140: "provides uncertainty": what is the basis / justification for the uncertainty estimate taking this particular form? It looks familiar as it has some similarities to a confidence interval but also some differences. This threshold is important to explain and justify in detail because it is used to threshold discrete landslides in the following analysis. Why threshold at 95% confidence? What is the impact on your findings (total volumes and scaling relationships) of thresholding at a difference CI (e.g. 99 or 90% confidence)?

We refer the reviewer to the original M3C2 paper which has an extensive discussion on the confidence interval, and how to consider surface roughness, point cloud errors, point density and registration in the context of change detection on 3D point clouds. The threshold has been set to 95% to build the segmentation on as many good points as possible. We do not believe that changing this threshold to 90 or 99% significantly change the landslide statistics given the results of the sensitivity analyses of parameters (reg, D_m and SNR threshold) that mainly control the landslide inventory.

L145: "detrended roughness": I think it would be useful to add that these are standard deviations if this is true.

We removed this part of the sentence which was unnecessary

L145: "reg" is quantified using the standard deviation of differences between the surfaces. I think it would be helpful here and elsewhere to use similar notation for the registration error to the other errors being examined here. Why are the local terms converted to standard errors but reg is left as a standard deviation? Finally, the length scale over which reg was calculated would seem to be important here.

We choose to keep the "reg" notation to specifically refer to the registration error when needed in the manuscript. We now provide an explanation in the text as to why reg is not converted as a standard error (L 166). Reg is not measured over a length scale, it is based on the standard deviation of the 3D-M3C2 distance between 2 clouds. This part is now explained in greated detail in section 3.3.1 where we discuss the notion of intra-survey and inter-survey registration quality.

L164: "at least 5 data points": Does removing these points from your analysis alter your results on pdfs and scaling?

The 5 data points considered here is the minimum number of points required to compute the $LoD_{95\%}$ (eq(2)) for each core point at a scale *d*. Even if M3C2 parameters can be changed to evaluate the $LoD_{95\%}$ for fewer points, the test performed by Lague et al. (2013) showed that the statistical model behind the definition of the $LoD_{95\%}$ became incorrect.

L165: "not deemed interesting": I don't think that this is the right phrase, can you rephrase? What was the impact on your results of applying d=10 m throughout?

The sentence has been modified. Applying d=10 m increases the landslide source and deposit by 1% and 0.7% respectively. In terms of volume this represents an increase of 2% and 0.89% respectively.

L173: "may result in…": How do you identify when this has happened? What is the objective that you are optimising?

This can be identified in the bottom of narrow valleys and top of very steep divides where no evidence of mass movement processing can be identified by visual inspection of orthophotos. These cases are now filtered by using the SNR metric as they have very large $LoD_{95\%}$ (see section 3.3.4).

L187: "larger than 1 m": perhaps give a range, >1 could be very large but really it is on the order of 1-2 m.

Updated as suggested (L241).

L187: "rasterizing the dataset": at what resolution?

The grid spacing used here have been added in the text.

L196: "standard deviation…" These stable areas would be an excellent test of your landslide detection method, indicating the scaling relationships, size distributions, and total volumes generated by artefacts alone.

This analyse has not been performed due to the changes we made on how we manage artefacts with SNR filtering (see section 3.3.4). Applied on stable area, the workflow does not detect any landslide.

L199-200: "The registration error…": this definition should come earlier. It is important for interpreting eqn 2.

We disagree with this comment as the registration error “reg” can be defined differently depending on the application of M3C2. In section 3.1, we aim to give a general description of the M3C2 algorithm.

L200-1: “manually selected”: This doesn't seem consistent with L191. Was a threshold of change selected manually then all areas with changes smaller than threshold included? Or, did you take a subset of pixels from the areas with changes smaller than the threshold so that only large patches were considered?

Actually, after applying a first 3D-M3C2, we selected only areas with a 3D-M3C2 distance less than 1m to help the identification of stable areas. We then manually refined the stable areas to select only areas away from landslides. The fine registration is then applied on these stable areas. The text has been updated.

L215: 15-20% it might be useful to show the location of these points on one of your maps.

The percentage of these points on steep hillslope has been revised and actually represents up to 12% of steep slopes.

L217-9: “all geomorphic processes…”: This assumes that all non-fluvial changes are landslides (see MC1).

The definition of what we consider as landslide has been added in the introduction. Please see the reply to the first minor comment.

L221-234: “Landslide source and deposit segmentation”: This process is much more clearly articulated in the SI. When I finished this paragraph I was totally confused but when I read the SI it became clear. The SI text on this analysis is really important and needs to be included within the main manuscript.

This section has been clarified and simplified. The segmentation algorithm segments a point cloud into sub-clouds based on the spatial connectivity between each point. A first parameter defines the minimum number of points (*Np*) a sub-cloud must have and a second parameter defines a minimum distance (D_m) from which two sub-clouds are considered as two individual landslides. In the new version of the manuscript, we explain how we chose the value of the two parameters. *Np* has been set to 20 points (or 20 m² given the spatial resolution of core point of 1m) as we average the position of both LiDAR point clouds in a 20 m² window (projection scale d=5m in diameter). As there is no objective way to a priori choose D_m, we explored various values and chose D_m=2 m as an optimal value between landslide amalgamation and over-segmentation. We also analyse the impact of D_m on the statistical distribution of landslide sources in the discussion (see section 5.2.2, Appendix A and S13, S14 in the supplements).

L225: “compact”: I don't understand what compact means in this context

This word has been deleted.

L225: More detail on how this method works is essential here. This step is a key component of your method and it is very important that the algorithm is given the same clear treatment that others have been in previous sections. I guess that Dm sets the distance between points within which two points are considered connected. This is a tuning parameter and it is not clear to me what objective you are seeking to optimise during tuning.

This question has been addressed in an earlier comment. Please see the answer of the comment for L221-234.

L226: "amalgamation effect and the over-segmentation": These two effects and your method for quantifying them need to be explained. I don't think this can be included only in SI.

Indeed, amalgamation errors are important to consider have plagued 2D inventories for quite some time. Yet, these have been published. We now provide a detailed sensitivity analysis, going far beyond we've ever seen, that should help the reviewer evaluate if he has confidence in our results. Also, contrary to a 2D manual analysis, our 3D detection is fully reproducible and data and workflow are provided for the reviewer to try by himself.

L228: "small artefacts". Do you mean that there are a small number or that they are small in magnitude?

We meant that they are small in magnitude.

We meant they are mall in magnitude. The text has been modified and "small artefacts" has been removed (see section 3.3.4).

L229: Why would it be the smallest landslides that are particularly affected by "these artefacts"? I had assumed that these were due to things like inclusion of points on either side of ridges.

The range of artefact area is now given in section 3.3.4.

L230: "minimum number of points": You haven't explained what this parameter controls. Is it the minimum number of connected components required for a patch to be retained (i.e. minimum landslide size)?

Np is the minimum number of points required to define a sub-cloud from the original point cloud and thus indeed corresponds to the minimum landslide size. This was previously in SI, and now is in the text

L231: I must have misunderstood Np, it does not follow that all detected changes will be artefacts if Np simply removes landslides smaller than some minimum size. On reading the SI I now understand, this is because you aren't comparing before and after point clouds. You explain this well in the SI, but need to bring some of that text in here.

This section has been updated. The choice of the Np value has been addressed in an earlier comment. Please see comments for lines 221-234.

L231-2: Reference to pre- and post- earthquake data here is confusing unless read in conjunction with the SI.

This section has been updated.

L233: "artefacts": What are artefacts? How do you identify them and why do they become smaller/less frequent/less important for larger areas? What does it mean that they are limited (e.g. they make up less than x% of the landslide objects)? The SI text on this analysis is really important point that needs to be included within the main manuscript.

We define artefacts as either false negative or false positive change detection. We identify them with the SSDS analyses (see section 3.2 and 3.3.4).

L238-9: I agree that volume should be calculated normal to the plane on which landslide area is calculated. But disagree that a horizontal reference frame is most appropriate. This will underestimate landslide area (severely on the very steep slopes typical of rockfall initiation). Surface parallel calculations seem more appropriate (see MC6).

Please see reply to the MC6) comment.

L245-7: "specific to each landslide": This potentially presents a nice opportunity to check your method by comparing source and deposit volumes for individual landslides over your study area.

We agree that it is something that would be interesting to look at it. However, this require further development to automatically link landslide source with the appropriate deposits which is beyond of the scope of the manuscript.

L252-3: "Most of the detected changes on hillslopes correspond to…" How do you know this? Is it based on your definition that all non-fluvial change is due to mass movement? Or, are you drawing on additional information to interpret the change map? Ideally you would have some independent check data.

In this section, we now first describe the pattern of large detected changes for which a source area can be associated to a deposit area. These large patterns can easily be linked to landslide processes. For smaller patterns, we now compare the landslide inventory obtained by the 3D point cloud differencing method with a landslide inventory manually mapped from visual interpretation of aerial imageries (see section 4.2). We also added in the discussion (section 5.1.2) a paragraph discussing the different processes we inferred from both the 3D-M3C2 distance field and ortho-imageries and the limit of our approach to confidently identify the dominant landslide mechanism for each sources and deposits.

L254: "previously unstable bare rock": It is useful to know that many are located on bare rock but how do you know that it was previously unstable and what do you mean by previously (this implies a timescale)?

The text has been modified and 'previously unstable" has been removed.

L255: "Their large number illustrates how difficult": Why would their large number make them difficult to manually extract? Perhaps time-consuming rather than difficult. An inventory with 1431 sources is a very modest sample size by the standards of modern landslide inventories.

The text of his section has been subject to significant modifications. This sentence has been removed.

L256: "deposit areas" Some of these areas do not have any upslope erosion. How do you explain this? See MC5

Some deposit areas can be detected without an associated upslope erosion. A possible reason is that the surface change associated to these deposit areas is sufficient to be above the $LoD_{95\%}$ but not for the upslope erosion area.

L259: "stable areas": This argument seems somewhat circular since stable areas were identified from surface differencing rather than independent information.

We agree but our study deals with data for which such information are not available.

L259: What do you mean by "artificial changes" here?

Artificial changes correspond to either false negative and false positive change detection. This definition has been added to the manuscript.

L286: It is not clear what "substantial" means in this context.

This word has been deleted

L291: A rollover is considered characteristic of landslide, but not of rockfall, distributions I think. As I understand it your inventory includes both. It would be useful to comment on this amalgamation of processes within a single inventory and what it implies for interpretation of the landslide distribution. See MC1.

Please see replies to the MC1) comment. We also now thoroughly discuss on the lack of a rollover and the effect of the amalgamation of processes on this result in section 5.2.2.

L292-3: "if we reduce the minimum landslide size to…" I don't think that you can make this statement if you have excluded landslides smaller than 20 m_2 for good reason. Your methods text makes a compelling argument that surface differencing errors will generate landslide-like artefacts that will distort the distribution and thus should be censored.

We agree that this sentence is confusing. It has been deleted.

L305-7: This is connected to my earlier concern about planimetric areas and suggests a need either to address the problem as you suggest here or to include a more detailed explanation for when the problem presents itself, what fraction of the landslides in your inventory might suffer from it. See MC6.

Please see the reply to the MC6) comment.

L308: Why frame your analysis in terms of volume-area relationships rather than depth-area? Since volume is the product of depth and area the x and y variables are correlated by definition. Others e.g. Larsen have examined depth-area scaling and the translation from one to the other is straightforward. But using depth de-trends the y-axis making it easier to see the scatter in the relationship, which is a major advantage.

This version of the manuscript now includes both V-A relationship and depth-area relationship (see section 4.3).

L315: It is not surprising that your depth area scaling relationship is gentler than that of Larsen since you censor core points with difference < 0.33, making it impossible for small landslides to also be shallow.

Indeed, and we now discuss this. In particular, the relationship is barely different if we consider a SNR=1, meaning that compared to the 2D inventory, we miss 32 shallow landslide sources over an inventory of 1270 sources (SNR >=1). Hence, even if the 3d inventory had captured the very shallow landslides that the 2D mapping did capture, it would hardly change the scaling relationship. One could also say that previous 2D inventories have significantly underevaluated the number of small landslides, which in turns affect the representativity of published V-A relationships from 2D inventories.

L325: An analysis of the uncertainty associated with total volume estimates would be very valuable. Your study is unusual in not only being able to resolve scar volumes but also deposit volumes, thus you can calculate a net *volume removed from the hillslopes. This value is of considerable interest and comparing it to the estimates that have been made elsewhere using different approaches seems worthwhile. You should however deal with your own uncertainties as well and within those uncertainties should be included those areas of deposition that have no landslide source upslope.*

The volume of landslide sources and deposits of our inventory is not statistically different. However, we believe that the larger volume of deposit is consistent with decompaction and the likelihood for sources to be more statistically filtered out than deposits (see section 5.2.1).

L344-5: A clear definition of reactivated landslides is needed here. See MC7.

This section has been removed. We do not address reactivated landslide anymore. This analysis has been replaced by the estimation of under-detected landslides based on the comparison with the manually mapped landslides.

L347: In many areas "bare rock areas" surfaces do not necessarily indicate recent landsliding (depending on how recent is defined). They may in this landscape though, so it would be useful to say so and evidence the claim.

This section has been removed. Please see the above comment.

L349: "following classical approaches": This is not a convincing justification for the method. It argues for the method's validity because others have previously adopted it without any further indication of why it might be valid here.

This section has been removed. Please see the above comment.

L384: I don't think you do explain the subsampling exercise in section 3.2 it is currently in the SI as I understand it.

We added a new section explaining the subsample exercise (see section 3.2), that we call SDDS (Same Data Different Sampling) test.

L431-3: "more advanced segmentation": an indication of the classification skill of the simple segmentation of object based classification skill is an essential requirement of this paper.

The classification skill of the segmentation approach can now be assessed with the landslide inventory map show in figure 6b for which each individual landslide has a single color (section 4.2).

L460: "deposits form more concentrated and thicker patches": This result is surprising and should be compared with expectations from other studies on landslide runout.

We think it is not surprising given the small runout of the landslides in this area, and is actually backed by the data in terms of mean 3D thickness of the deposits and sources.

L461-2: I would have expected "very shallow rockfalls" to result in even shallower rockfall deposits because they are initially small in volume and spread out over a wide area during deposition. Reference to rockfalls here increases my concern around use of horizontal (i.e. planform) area calculations.

Please see response to MC6) comment. Concerning the rockfall deposits, it might be the case if the rockfall is triggered in areas free of previous rockfall/landslides. However, if the rockfall is triggered in an active sliding area thus the likelihood of the associated deposits to merge with a pre-existing deposit area increase. In such case, the volume of deposit area will appear larger than source area because the deposit zone collects sediments from different sources. Our study are present such cases. To appease the concerns of the reviewer, we also provide the distribution of pre-eq source slope that shows that in the final inventory, slopes larger than 60° are less than 1 %. Rockfall are thus not expected to be a significant contribution in our inventory.

L465: It is not clear what you mean by "landslide analysis". I suggest cutting this phrase, I don't think it is necessary.

Updated as suggested.

L466: Whether the right tail is a "power law" or not is debated. See Medwedeff et al. (2020) among others.

The reference has been added in the text.

L473-4: Power law scaling can be sensitive to amalgamation errors such as those in your inventory. As a result I struggle to know how much confidence I can place in these values. See MC3.

Please see response to the MC3) comment. Obviously amalgamation will affect power-law exponents, and we now provide a systematic analysis of the impact of various parameters of the workflow that affect the final geometry of landslides (reg, and D_m). We do not argue that our dataset provides a unique, absolute power-law exponent.

L488-90: "rollover … is likely caused by an under detection of small landslides": This is a very important claim with significant implications for our understanding of landslide mechanics. But it depends on both: 1) your findings *being robust to error in segmentation; and 2) the inventories that you are discussing having comparable definitions of landslides. Given that the landslide detection method has not been tested and that there are reasons to expect that the method introduces considerable bias in the size distribution I am not convinced that the claim is true. The lack of a rollover in your dataset might instead be due to detection errors.*

We believe we provide now firm evidence that the rollover is related to under-detection, in particular given the strict filters we apply on our inventory. We also demonstrate that under-detection exists, even for relatively large landslide of the inventories for which there is absolutely no doubt that they are landslides (see section 4.2). We now extensively discuss the potential other causes for the occurrence of a rollover and demonstrate through a sensitivity analysis that the segmentation parameter does not change this (note, the segmentation parameter sensitivity analysis was already presented in the first version of the paper; see section 5.2.4).

L499-500: "most of our inventory is relevant to shallow landsliding": shallow landsliding is usually defined as landslides initiating above the soil-bedrock interface, and thus distinct from rockfall. Can you make this distinction in your inventory?

We agree that we cannot make this distinction. We now compare, in section 5.2.4, our landslide depth-area and volume-area relationship with previous studies (Larsen et al., 2010; Massey et al., 2020). The similarities with the results obtained for soil-type landslide suggest that our landslide inventory may be consistent with shallow landsliding. We discuss about this suggestion at L-776-781.

L511: "cannot resolve the amalgamation…": The problem of amalgamation needs a more detailed treatment. Total volume is insensitive to amalgamation but landslide size distributions and scaling relationships are not. See MC3.

We brought further analysis on how our results are sensitives to amalgamation. See section 5.1.2.

L535: "a much lower detection level than optical methods": This phrase is unclear, can you rephrase?

This section has been subject to significant modifications. This sentence has been removed.

L551: "95% confidence of 0.34 m": This is quoted here as spatially invariant but the method captures spatial variability. The value here doesn't account for the impact of local roughness on the confidence intervals.

That's right. What we describe here is the significant minimum distance. The text has been updated.

L558: "Amalgamation in 3D is still a potential issue": Amalgamation needs a more detailed treatment, given its potential impact on your findings it cannot simply be flagged as a potential issue.

We brought further analysis on how our results are sensitives to amalgamation. See section 5.2.2, appendix A and fig. S13 and S14 in supplements.

L561: "reactivation" needs to be much more clearly defined. However, your argument doesn't really depend on reactivation but on landslide detection in vegetation sparse or vegetation free areas. I suggest restating in these terms.

We do not address the landslide "reactivation" issue anymore. This has been replaced by an evaluation of landslide under-detection based on the comparison with the manually mapped landslides.

L566: "V-A relationship": depth-area should be examined as well as or instead of V-A because it removes spurious correlation with area. All three geometric properties are likely to be very severely affected by detection errors.

This version of the manuscript now includes both V-A relationship and depth-area relationship (see section 4.3).

L569-70: "largest and deepest landslides deviate significantly": these slides are also most likely to suffer amalgamation.

This sentence has been removed.

Figures

Figs: 3,4&6: Though they are very well presented I find the 3D plots (Fig. 3, 4 and 6) very difficult to interpret. I feel strongly that all these Figures should be presented in 2D map view instead of as oblique images.

We modified these figures with a 3D view to a 2D view. We kept 3D views when it helped in the interpretation (e.g., Fig. 10)

Fig 4: The colourmap in Fig4a does not match description in the caption or in the text. It appears to show a single colour for erosion and another for deposition. It also appears to apply a filter to areas with non-significant change. Areas of significant change are shown in Fig4b so the mask is not necessary here. By masking non-significant changes you lose the opportunity to compare the magnitude of change before significance tested. The colourbars in both Fig4a and c are a good length but contain little usable information. As I see it there is no change in colour *for erosion/deposition >1 m in Fig 4a and 5 m in Fig 4c. In addition, the scaling is asymmetric in 4c making it difficult to compare erosion and deposition.*

The colorbar has been modified to match changes up to 4m and has been saturated for higher values.

Fig 5: The 3D minimum volume line needs explaining in the caption. I would have expected this line to be oriented parallel to the depth contours since minimum volume for a given area is set by minimum detectable depth. If so the minimum detectable volume might explain the sharp lower boundary to the volume area point cloud.

Actually, what it is illustrated here is the 3D minimum volume that can be measured here given the minimum area (20 m^2) that we consider and the minimum significant depth (~0.4 m).

Fig 6: The inventory of Massey is useful additional information but scar outlines would be more useful than controids.

This figure has been removed. Unfortunately we were never able to get the scar outlines of the first or the second inventory.

Fig 6b: It would be useful to identify the largest four or five landslides in the right hand panel of Fig 6b either by giving them individual outlines or different colours this would give an indication of the degree of amalgamation.

We now illustrate the amalgamation by showing the polygon of each landslide in figure 7a.

References

Aerial Surveys, Aerial photographs derived from two surveys of the study area carried out in 2014 to 2015 and in 2016 to 2017, by Aerial Surveys Ltd, 2017.

Dolan, J.F.: Data collection and processing report LiDAR survey of five fault segments (Estern Clarence, Western Clarence, Central Eastern Awatere, West Wairau and East Hope-Conway) of the Malborough fault system on the Northwestern portion of New Zealand's south island. Ph.D., University of southern California, 11 pp., 2014

Dolan J.F, Rhodes E.J.. Marlborough Fault System, South Island, New Zealand, airborne lidar. National Center for Airborne Laser Mapping (NCALM), distributed by OpenTopography. http://dx.doi.org/10.5069/G9G44N75, 2016.

Ankerst, M., Breunig, M. M., Kriegel, H.-P. and Sander, J.: OPTICS, ACM SIGMOD Rec., 28(2), 49–60, doi:10.1145/304181.304187, 1999.

Benjamin, J., Rosser, N. J. and Brain, M. J.: Emergent characteristics of rockfall inventories captured at a regional scale, Earth Surf. Process. Landforms, 45(12), 2773–2787, doi:10.1002/esp.4929, 2020.

Carrea, D., Abellan, A., Derron, M., Gauvin, N. and Jaboyedoff, M.: MATLAB Virtual Toolbox for Retrospective Rockfall

Source Detection and Volume Estimation Using 3D Point Clouds : A Case Study of a Subalpine Molasse Cliff, 2021.

Lague, D., Brodu, N. and Leroux, J.: Accurate 3D comparison of complex topography with terrestrial laser scanner: Application to the Rangitikei canyon (N-Z), ISPRS J. Photogramm. Remote Sens., 82, 10–26, doi:10.1016/j.isprsjprs.2013.04.009, 2013.

Larsen, I. J., Montgomery, D. R. and Korup, O.: Landslide erosion controlled by hillslope material, Nat. Geosci., 3(4), 247–251, doi:10.1038/ngeo776, 2010.

Martin Ester, Hans-Peter Kriegel, Jiirg Sander, X. X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Compr. Chemom., 96(34), 226–231 [online] Available from: https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf?source=post_page, 1996.

Massey, C. I., Townsend, D., Jones, K., Lukovic, B., Rhoades, D., Morgenstern, R., Rosser, B., Ries, W., Howarth, J., Hamling, I., Petley, D., Clark, M., Wartman, J., Litchfield, N. and Olsen, M.: Volume characteristics of landslides triggered by the M W 7.8 2016 Kaikōura Earthquake, New Zealand, derived from digital surface difference modelling , J. Geophys. Res. Earth Surf., 0–3, doi:10.1029/2019jf005163, 2020.

McInnes, L., Healy, J. and Astels, S.: hdbscan: Hierarchical density based clustering, J. Open Source Softw., 2(11), 205, doi:10.21105/joss.00205, 2017.

Tonini, M. and Abellan, A.: Rockfall detection from terrestrial lidar point clouds: A clustering approach using R, J. Spat. Inf. Sci., 8(1), 95–110, doi:10.5311/JOSIS.2014.8.123, 2014.

Author’s response to referee comments of Bernard et al. 2020 by Alexander Densmore.

Thomas G. Bernard, Dimitri Lague, Philippe Steer

Univ Rennes, CNRS, Géosciences Rennes - UMR 6118, 35000, Rennes, France

Summary

This is an exciting manuscript that reports on a promising way forward in detection and analysis of landslide inventories. The authors use the M3C2 point cloud analysis approach, previously developed by Lague and colleagues, to detect change in pre- and post-earthquake Lidar point clouds from the Kaikoura earthquake area in New Zealand. This is absolutely the right thing to be doing, and the authors are clear about how their approach gets around some vexing issues with current best practice (which is to map using 2d imagery and to use volume-area scaling to get at landslide volumes). The topic is fully appropriate *to the journal and I anticipate that this work will attract a good deal of attention from the journal readership. do have some questions and suggestions for the authors to consider before the manuscript is published. Many of these are fairly minor and are detailed in the attached PDF; I will not repeat those here. But there are a couple of wider issues that are related to the clarity of what the authors have done, and in part to the division of material between the manuscript and the supplemental information. In brief, I don't think it's possible to follow the authors' approach from the manuscript alone, and there are key parts of their analysis that can only be understood by going to the SI. I don't think that's right for a manuscript that purports to document a new methodological approach. These issues come under three headings (but please see the PDF for more detailed comments): First, because the authors are documenting a new approach for mapping landslides, I would expect to see some quantitative comparison of their results with a landslide inventory or inventories prepared in the more traditional way. The authors show some amalgamated results (Fig 5, Table 2), but I think it would be useful to show a more systematic comparison with the Massey et al. results. Given that there are only 27 landslides in the Massey et al. inventory within the study area, this should be pretty straightforward - but I think it is important to demonstrate the extent to which their approach can match (or not) landslides detected by the alternative approach, as well as the additional landslides that they claim to be able to map. The only place I could see the Massey et al. landslides was as barely-visible centroids on a perspective view of the study area in Fig. 6. Second, and somewhat related, the text is very unclear on how individual landslide sources and deposits are segmented and identified. Because of this, all of the resulting statistics of the area and volume distributions are uncertain in the reader's mind. This is more clearly explained in the SI... but again I don't think it's fair to require the reader to go to the SI to understand a methodological advance that is being proposed. I deliberately read the manuscript without going to the SI to see if I could follow it, and there are places (e.g., section 3.3.3) that are very difficult to understand and don't really address what has been done. I'd really recommend that the authors review the balance between the text and SI and try to flesh out the explanations in the main text. The details of the sensitivity analysis (e.g., to the distance threshold D_m) can be left for the SI, for sure. Third, I found the authors' use of the term 'reactivation' potentially confusing, and I'd suggest that they choose a different way to express this. Assessing reactivation (e.g., of coseismic landslides in post-earthquake storms) is definitely a big issue in multitemporal landslide inventory analysis, and a few different approaches have been put forward, none of them very satisfying. Reactivation implies some renewed landslide activity that partly or wholly overlaps with a landslide from a previous epoch - e.g. further erosion within a pre-existing scar, or headward/lateral progression of the scar edges, or erosion within a scar coupled with deposition on a pre-existing deposit... But that's not what the authors are actually talking about here, because they only have two point clouds spanning one epoch, and there's no independent dataset of pre-existing landslides. A better way of framing this part of the analysis might be around the following question: are there landslides that would not have been recognised using the classical approach, either because they occurred on bare bedrock or because the vegetation contrast was too low, but that they can see with their method? That's an important question - but it's not the same as reactivation. Finally, a fairly minor point: I agree that the approach outlined by the authors is the way forward and that, where suitable Lidar data are available, this should be further pursued and developed. But I think it's equally fair to recognise that (1) suitably accurate and high-resolution Lidar data aren't always available, and (2) there are problems* *and applications for which this approach simply isn't (yet) feasible. For example, for multi-temporal inventory creation over the full landslide affected area, where the goal is understanding patterns of landslide occurrence and hazard but volume estimation is a secondary concern, then a traditional 2d image-based approach might be fine. New Zealand and a few other countries can fly repeat high density Lidar surveys; this capacity doesn't (yet) exist in most landslide-prone regions of the world. I think it would be fair of the authors to acknowledge this - it doesn't detract from their analysis but perhaps places it into a better wider context. To summarise, this is a really exciting piece of work. Once the authors have dealt with these issues, then I look forward to seeing this published.*

Minor Comments

L20: "It's not clear at this point of the ms what this clause means - so you might either expand/explain this point, or leave it out. 90% of what?"

We brought clarification to this sentence. It is now : "(81% of landslides with area < 300 m^2)".

L58:" You might also add that this effect complicates comparison of inventories that were constructed from different image sources, because of differences in image spatial resolution, spectral resolution, and consequent ability to characterise surface change... In other words, we don't even know if we're comparing like with like"

The following sentence has been added to the text: "It can be further complicated by the use of different image sources because of differences in image resolution, spectral resolution, projected shadows and consequent ability to detect surface change".

L97:" This is an awkward phrase - maybe reword as 'identification of individual landslides'?".

The text has been modified as suggested (L110).

L103:" The black box showing the study area is not easy to see at first - perhaps make that white or yellow so it stands out?"

The box has been colored in yellow as suggested.

L114: *"I'm not sure what this means"*

All LiDAR point cloud data contained information provided by the owner. One of these information is the classification which attribute for each point of the point cloud, a number depending on the type of surface the point corresponds to (i.e. ethier it is vegetation, ground, road etc..). Here, we only keep points from the original point cloud that correspond to the class 'ground'. This sentence was replaced by "For both LiDAR point clouds, only ground points defined by the data provider are selected (see details of the classification in Dolan, 2014) ".

L115:" This seems like a really critical part of the analysis, and yet it is given only a single sentence. I'm not sure how a new approach for estimating landslide area and volume can be put forward without some kind of comparison to a more traditionally-created inventory. I agree that those inventories have problems, but otherwise there is no possible way for the reader to evaluate this new approach. So how did the validation work - what did you compare? I think this needs to be more clearly described »

We added an entire section that compare our landslide inventory by 3D point cloud differencing and a manual mapping based on 2D images. See section 4.2. The text in this section 2. has been modified to introduced this manual mapping.

L121:" From who? »

The contact of the platform hosting the data has been added.

L132: "I'm not sure what this means".

It means that as core points represent a regular grid of points and that the results are "stored" on this grid, it can easily be import in any GIS rather than a non-regular LiDAR point cloud. We modified this sentence as follow: "it can be directly reused with 2D GIS as a raster (rather than a non-regular point cloud)" (L152).

L135: "Just to be clear - are these normals to the pre-event or post-event data? The schematics in Fig 2 suggest post-event - is that the case? Might be worth mentioning that, either way"

The normals are calculated from the core points of the post-event data. The text has been modified from "first dataset" to "core points". (L155).

L180:" There is some repeated text here - can you please clarify?"

The caption of the figure 2 has been modified.

L182:" The caption and panels (a)-(c) don't seem to match - volume estimation seems to be panel (b), while separation of source and deposit seems to be panel (c). Can this be clarified? »

The caption has been clarified.

L186: "Removed? Corrected for? I'm not quite sure what you mean by 'adjusted'"

The text has been modified as suggested.

L191-192:" I'm not quite sure how this works - presumably as the threshold is varied the areas that appear 'stable' vary as well, so how can a single threshold value be assigned? Perhaps I'm misunderstanding what you have done - but as identification of 'stable' areas is important for the registration, it would be good to clarify this".

This section has been re-written and clarified. Actually, after applying a first 3D-M3C2, we selected only areas with a 3D-M3C2 distance less than 1m to help the identification of stable areas. We then manually refined the stable areas to select only areas away from landslides. The fine registration is then applied on these stable areas (see section 3.3.1).

L193:" is this rotation and translation, or just translation?"

We let the algorithm free to adjust the data with rotation and translation. However, the resulting transformation matrix of the ICP only show translation. We let "rigid transformation" in the text to indicate that we did not imposed a translation for the co-registration of the two point clouds.

L200:" This has already been written in the preceding paragraph"

This section has been re-written and clarified.

L218-219: "OK - but how is the location of 'the river' determined? How do you know what is fluvial and what is landslide-driven? I am thinking for example of elevation changes at the base of a steep bank or talus that ends in the river channel - these processes could be difficult to separate, no?"

The river is determined at the bottom of the valley between distinctive banks and where it is free of vegetation. This was done by visual inspection of the orthophotos. Then, fluvial elevation changes were manually separated from 'mass-wasting' elevation changes in the river based on the shape of the patches of negative or positive elevation changes. All elongated patches in the direction of the flow of elevation changes were considered as fluvial erosion or sedimentation. Deposits from mass-wasting processes was recognized as compact 'round' patches with highest elevation changes. Where fluvial sedimentation meets mass-wasting deposits, the latter was delimited by an abrupt change of the slope. Of course, this procedure comes with uncertainties that are difficult to quantify. The authors do not exclude the possibility that few fluvial changes have been merged with mass-wasting processes but we do not expect a significant impact on the total volume. It does not have any impact of landslide source statistics.

L222: "Remaining after what? I'm not sure what this is referring to"

The term "remaining" was used in reference to the resulting point cloud after the removal of points located in the river. The word has been removed.

L223:" Why not describe these in the order in which you carried them out? That also would fit with the order in Fig 2. Not clear why these have been switched around in the text"

The figure 3 (in the current version of the manuscript) showing the workflow have been changed to match the text. However, the vertical-M3C2 needs to be performed before the segmentation of landslide sources and deposits. This aspect is just a matter of workflow efficiency. Doing the vertical-M3C2 after the segmentation would require to perform the calculation on each individual landslides.

L225: "Is Np the minimum number of points or the number of sub-clouds (which is how it reads now)?"

Np is the minimum number of points. It has been corrected in the text.

L226-227:" This is a pretty important part of your analysis, so I think it needs to be better explained in the text. I know what you mean by amalgamation (although I'm not sure how you are assessing it), but I'm not sure what you mean by over-segmentation. I don't think the reader should have to go to the SI to understand this, as it's critical for going from topographic change to landslide process understanding. Can the concept be explained here, even if the details of how it is implemented are left for the SI?"

This section has been clarified (see section 3.3.3) and all the information is now in the core of the text. We explain how D_m operates. As there is no objective way to a priori choose D_m we explored various values and chose D_m=2 m as an optimal value between landslide amalgamation and over-segmentation, that is documented by the number of clusters that are created. We also analyse how the value of this parameter influence the landslide statistics and show that the value of D_m do not significantly influence the results for $1.5 < D_m < 3$ m (see section 5.2.2, appendix A and S13 and S14 in supplements).

L228:" I'm not sure what this means - what kind of artefact? Do you mean 'core points that have a very large distance'? How do you know they are artefacts? »

We use "artefact" to define either negative or positive false detection (L-301). To estimate the proportion of artefacts in the landslide inventory, we first perform the landslide detection workflow on two subsampled versions of the post-EQ point cloud (see section 3.2). Due to the difference of sampling, our workflow detects artefacts and we define for each of them a mean signal-to-noise (SNR) ratio that is the ratio between the mean 3D-M3C2 distance and the associated uncertainty. We also define a mean signal-to-noise ratio for each landslide in the inventory. We then use a SNR threshold from which the presence of artefact is limited in the landslide inventory (see section 3.3.4).

L230-233 : « I found this text hard to follow, and it didn't seem to explain to me the focus of the sub-section - which is how sources and deposits were separated. What does the final sentence mean - how was a minimum surface of 20 m2 'imposed'? Imposed on what? What does the final clause of that sentence mean? Up to here the text has been very clear, but I'm really not sure how the source and deposit have been discriminated. Because that's so critical for the volume distribution information and indeed much of the remaining text, I think that needs to be really clear".

This section has been subject to significant modifications. We clarified how we perform the segmentation of landslide and how the value of the parameters are chosen. By “imposed” we meant that the minimum landslide area is set by the method. We do not use the word anymore (see section 3.3.3).

L239:” Is this still using D=10 m and d=5 m? More generally, this approach seems to introduce some of the issues created by using a DoD approach, except without the cellwise averaging/interpolation problem. But for steep rock slopes/rockfalls, this seems like it would introduce substantial uncertainties. I haven't thought this through, but I do wonder whether sticking with the 3D normal approach, but introducing a set of rules or hierarchies to deal with shadow zones, would be preferable. At the very least it seems worthwhile to explore - whether in this ms or elsewhere.”

Here, the projection scale is 5 m. We agree that the vertical measurement approach to estimate volume may be incorrect on very steep slopes and that it would be preferable to estimate landslide volume in 3D. This is highlighted in the discussion (L568). However, the median landslide slope measured on the pre-EQ point cloud is 34.6° and only 0.74% of the landslide area have very steep pre-EQ slopes (>60°). In addition, we do not expect that measuring volume in 3D from a constant surface normal direction for each sources and deposits would be better than a vertical measurement when we have very complex amalgamated landslides as we observed in the dataset (Fig. 7c). However, new approaches based on 3D mesh reconstruction seem promising (Benjamin et al., 2020).We discuss about this issue in section 5.1.2.

L251:” Was there water in the river? If so, are the lidar data measuring the water surface, or is there some penetration, or a mix of both? Do you have any sense of how changes in water level between surveys could contribute to the distances that you estimate?”

Both LiDAR used to generate the point clouds are topographic LiDAR using NIR wavelength which means that the laser used does not penetrate into the water, and is actually fully absorbed. Hence, where there is water, there is no data. Because the lidar survey were acquired in summer, the water level were extremely low, and do not appear to create significant area without water. We have not added this kind of information in the MS as it is not central to the topic of the paper.

L252:” I'm not sure about the evidence for this statement - is this based on comparison of the orthophotos? Did you overlay the changes in Fig 4 on the post-event orthophoto to see how they correspond? You mentioned using the orthophotos as a validation tool, but didn't really explain what this meant. This is really critical, because without some independent validation, the reader has no basis for understanding (or accepting) what you consider to be a 'landslide' and what you consider to be an 'artefact'.”

This sentence has been changed. We now discuss about large patches for which the association with a process is pretty straithforward. We also added a comparison between the landslide inventory resulting from 3D point cloud differencing and a landslide inventory manually created from interpretation of orthophotos. See section 4.2.

L254:"'previously' - and what is this statement based on? How do you know that it was unstable?

The text has been modified and 'previously unstable" has been removed.

L256:" At this point of the manuscript, you haven't explained how you know (or can state with confidence) that areas of positive distance change are landslide deposits. I think it would be better to first document the pattern of change that you observe, and then demonstrate how you correlate that (or not) with a particular geomorphic process".

In this section, we now first describe the pattern of large detected changes for which a source area can be associated to a deposit area. These large patterns can easily be linked to landslide processes and we do not think there is any ambiguity about that. For smaller patterns of change, the application of the SNR filtering now removes a very large number of positive (and negative) change which could be ambiguous. We also provide details of typical landslide processes that clearly show where deposits occur (e.g., fig. 10). We have also added in the discussion (section 5.1.2) a paragraph discussing the different processes we inferred from both the 3D-M3C2 distance field and ortho-imageries and the limit of our approach to confidently identify the dominant landslide mechanism for each sources and deposits.

L259:" What does this mean? How are these areas identified?"

Artefacts includes both negative and positive false detections (definition L-301). Artefacts are estimated from the Same Surface Different Sampling test (SSDS; section 3.2) and filtered by the signal-to-noise ratio in the landslide inventory.

L270:" Line 264 gives the same maximum but ascribes it to an erosion area, not a deposit - can you clarify?"

This mistake has been corrected.

L271:" Is it intentional that the colorbar is either blue or red? Or is there a gradation at very small +/- values that can't be seen by the reader? If so then I think it would be good to show this, perhaps with a non-linear color bar - because otherwise the pale blue/red colors can't be understood".

The colorbar has been modified to match changes up to 4m and has been saturated for higher values.

L275:" This figure looks to show continuous values of erosion and deposition, which is what I'd expect as the output from the vertical M3C2 step. But the caption seems to suggest that sources and deposits have been segmented, which isn't obvious from the figure. Is it necessary to say that? It would be more clear if you showed the unsegmented raw vertical distance change, and THEN the results of the segmentation (which presumably could be shown as grouped polygons?). Related to this, while the perspective views look interesting, they also hide a lot of the features of the dataset. One perspective view early on might be useful, but I'd argue that these results are better shown in map view. Just an opinion, however - other readers might disagree

"

We agree that the caption can be confusing regarding the map of the vertical-M3C2. The caption has thus been modified by removing the number of landslide sources and deposits. We also added a landslide source map for which each individual landslide is colored by a single color in figure 7 (section 4.2). In addition, 3D view has been changed for map view.

L279-280:" OK... but surely one of the advantages of this approach is that you are not constrained to only show planimetric areas! For example, it would be great to look at both planimetric and true surface area, to see how wrong we typically are by only using 2d imagery to map from"

Computing true surface area requires significant new developments of the workflow in relation to segmentation, to better handle complex cases of landslide amalgamation. We have added in the discussion quantitative elements based on the typical slope of landslide sources to discuss the magnitude of the difference between planimetric area and true surface area (see section 5.1.2). Given that the median of the slope distribution of our landslide source inventory is 34.1, the estimate of the true area gives a total landslide source area of 356,876 m^2 rather than 286,445 m^2.

L291:" These are all fairly old refs... I know that Hakan Tanyas has considered this issue recently, so might be worth making reference to that work?".

The reference to the work of Tanyas et al., 2019 has been added here.

L302-303:" This text doesn't quite make sense - missing word"

The text has been clarified: "the minimum volume that we can confidently measure should be 8 m^3".

L310:" I'm not sure that this is worth including, because the goal of any scaling relationship is capturing the ensemble or inventory behaviour rather than the scaling for an individual landslide"

We suggest to leave the approach here as it is. We illustrate in section 5.1.4 how the method used to fit the data influence the total volume estimation (L655–657).

L319:" There's not enough information for a reader to understand this sentence - what are cleaned or soil-dominated landslides? Are these landslides mapped by Massey et al. 2020 ? »

In their paper, Massey et al. (2020) estimated different V-A scalings in function of different landslide type. To simplify things, we removed the V-A scaling relationship estimated for their soil landslide type as it is close to the V-A scaling obtained with their entire landslide inventory.

L324-325:" This seems like a point that is better made in the discussion"

This sentence has been transferred to the discussion as suggested (see section 5.2.4).

L325-328:" Evidence for these statements? Overpredicts or underpredicts relative to what?"

This sentence has been removed.

L328-330:" Again - this is better left for the discussion. If you cannot distinguish between 'shallow' and 'deep' landslides, then why bring this up? It strikes me that, with direct measurement, this is no longer a meaningful distinction anyway, as presumably your data are consistent with a distribution of landslide depths..."

This sentence has been removed. We now show and discuss the landslide depth-area relationship of our landslide inventory (see section 4.3 and 5.2.2).

L332:" What does this mean?"

N_{LT} is defined at the beginning of the section 4.3 as the number of landslide sources. However, it is true that the notation was not respected which can be confusing. We corrected this in the text by writing "N_{LT}".

L333:" Are the Massey et al. data limited to landslides within the same study area? What does V2 mean? Given that you are making an explicit comparison between their data and yours, I am surprised that you are only mentioning that study here in the results. It would be good to mention their work in the intro or methods, including enough information about what they did so that the reader can both understand what's plotted and appreciate why their results might be different from yours"

The data of Massey et al. (2020) are not limited to our study area but include landslides over the Kaikoura region representing a total area of 6,875 Km^2. In the paper of Massey et al. (2020), they define the "V2" landslide inventory to refer to the landslide inventory started in 2018 and updated in 2020. To simplify, we removed "V2" in the manuscript. We now refer to their study in the introduction (L.101-102).

L335:" This should be 'relationships' because more than one is shown. I don't understand the legend or the wording in Table 2 - you refer to fits on 'averaged data' and 'raw data', but that's different than the way you've described this in the text (lines 310-315). If that's what you're referring to, can you make these consistent? "

The text has been corrected as suggested and we modified the legend in Table 3 to be consistent with the text.

L335:" The green and red colours will be difficult for colour-blind readers to distinguish. Again, without knowing what Massey et al. (2020) did, it's hard to understand the difference between their 'all landslides' and 'soil landslides' scaling relationships - I think this needs to be introduced in the intro sections of the ms"

The color have been changed.

L340:" It took me a few reads to understand this - the column headers could be read as a calculation (log[b-d]/log[alpha]). I think this could be a little more clear".

Clarification have been brought. The column header is now: "log b, log d or lod α".

L346: " Without knowing anything about the study area, this doesn't seem like a convincing definition of 'reactivated landslides'. And I wouldn't necessarily agree that this matches the 'classical approach' - especially because the 'classical approach' of looking for vegetation change, either qualitatively or using NDVI trajectories, is notoriously incapable of spotting reactivation. There is also a question here of what you mean by 'reactivation' - does this include further erosion within a pre-existing scar, or headward/lateral progression of the scar edges, or erosion within a scar coupled with deposition on a pre-existing deposit, or...? I think 'reactivation' is not really what this is about because you don't have an independent dataset of pre-existing landslides, so you can't say what has or has not been reactivated. A better way of framing this part of your analysis might be around the following question: are there landslides that would not have been recognised using the classical approach, either because they occurred on bare bedrock or because the vegetation contrast was too low, but that we can see with our method? That's an important question - but it's not the same as reactivation"

We agree that "reactivation" is not what we actually looking at. This section has been removed and we added a new section in which we compare our landslide inventory based on 3D point cloud differencing to a manual mapping based on 2D images (See section 4.2). We then compare both approaches in terms of similarities and differences and discuss sources of under-detection by both approaches.

L356:" This is exactly the kind of comparison that I would expect in order to evaluate both your change detection and your segmentation routine. I think this needs to come earlier in the results. Why are you comparing against Massey et al. (2018) here and in Fig 6, but Massey et al. (2020) in Fig 5? "

We added a comparison of our result to a manual mapping based on 2D images. See section 4.2. The comparison with the landslide inventory of Massey et al. (2018) here was due to the availability of the data. The landslide inventory of Massey et al. (2020) is not available online and we never received an answer from the authors to our requests to get the extent of their mapped landslide sources and deposits.

L356:" These are not the same thing - and the issues involved in their detection are not the same either. I don't think you should be grouping these together »

This section has been replaced by the section 4.2 in the new version of the paper.

L361: "I'm still not 100% sure how you have separated source and deposit..."

The segmentation procedure has been clarified, please see section 3.3.3.

L361: "The centroids are really difficult to see on the perspective view - and will be impossible for colourblind readers to see. Why not show polygon outlines? That's the most direct comparison between your results and theirs, and you could even summarise in terms of an ROC curve or confusion matrix. As before I think the perspective view decreases the utility of this figure rather than increasing it - I'd really suggest showing this in map view as well, or solely in map view perhaps"

We did not show the landslide inventory of Massey et al. (2018) as polygons because it is not available online and we never received an answer from the authors to our requests to get this information. However, we added a section in which we analyse the results from 3D point cloud differencing compare to a manual mapping based on 2D images. See section 4.2.

L363:" Are these vertical or perspective views? And they need a separate scale bar".

This figure has been removed.

L364: " To help the reader, you could usefully add the dates to these panels - there is plenty of space to do so."

This figure has been removed.

L371:" As written above, I'm not sure that you overcome these limitations because your volume numbers come from a vertical application of M3C2!"

It is true that a vertical distance measurement is not the most appropriate approach to compute volume. However, here we compare the DoD approach to our 3D point cloud differencing approach only in terms of topographic change detection for which with compute distance in 3D.

L376-380:" This partly restates points that were made in the introduction - I think you can leave this out and just focus on the novelty and importance of your approach (which you cover below)"

This paragraph has been removed.

L387:" This doesn't quite make sense because there isn't a single distance measurement here. Do you mean a non-null mean, or median? I think this could be a little clearer"

The text has been replaced by "a non-null mean distance" as suggested.

L397:" I think I know what you are saying, but the wording here is a little odd. In the previous sentence you are talking about the benefits of 3D M3C2, but here this sounds like a drawback. Can you clarify? »

The idea here is that the distance uncertainty compute by M3C2 will be higher in vegetated area due to (1) a lower point density and (2) a higher roughness of the point cloud. Consequently, distance measured by M3C2 in such area needs to be higher than in flat area to be significant to prevent in part for the detection of false positives and false negatives. The text has been clarified.

L407: "This is a very effective figure. You could get rid of some of the trailing zeros on the y-axis, and also format the numbers in the legend to match the text"

Updated as suggested.

L411:" See query above on whether this is translation and rotation, or just translation".

We let the algorithm free to adjust the data with rotation and translation. However, the resulting transformation matrix of the ICP only show translation. We let "rigid transformation" in the text to indicate that we did not imposed a translation for the co-registration of the two point clouds.

L438: "What do you mean by this? Spatial density? Or something else?"

Indeed, we talked about landslide spatial density. However, this paragraph has been modified and integrated to the new section 5.1.2. This sentence has been removed.

*L439-440: "This sentence would make more sense to me if you are talking about landslides for which *little* topographic change occurs in the direction of the surface normal. Is that what you mean? The cartoon in Fig 8b shows such a case - there is negative surface elevation change in the source and positive change at the toe, but little/indeterminate change over much of the body of the slide. Can you clarify if this is what you mean? »*

This is exactly what we meant. However, as explained in the previous comment this paragraph has been modified and this sentence does not appear anymore.

L454 : " repeat the uncertainty given on this number in section 4"

The uncertainty has been added.

L460: "I'm not sure what you mean by this - more concentrated in terms of what?"

We meant that the landslide deposit areas are on average smaller than landslide source area. We have largely rewritten this section to make things clearer (hopefully !)

L470: "what is N_LT?"

N_{LT} is the number of landslide source. It is defined in eq (3).

L475:" You've expressed the exponents as negative numbers so be consistent here as well"

The text has been modified as suggested.

L492: "True... but what fraction of the total area?"

We do not address this question anymore.

L494: "In what sense? I think this sentence would be more effective if you could be specific. For me, why this potentially matters is that small landslides may be negligible in terms of their contribution to regional-scale erosion or sediment transfer, and yet still collectively have tremendous impacts on the exposed population. Even a very small landslide can cause casualties or infrastructure damage. It's worth reminding the reader of that, perhaps... »

We now address this question in section 5.2.2 at lines 749-752.

L501 : « By 'the best estimation' do you mean the estimate that is closest to yours?"

We removed the V-A scaling relationship of soil landslide type of Massey et al. (2020) and removed this sentence.

L503: "You could perhaps remind the reader why this is a 'better result'"

The text has been replaced by: 'gives the closest approximation of the total landslide volume estimation."

L571: "have observed [and you only cite one study - if there is only one, then why write 'inventories'?]"

This sentence has been removed.