Global response highlighting the main modifications to the paper

Note: in the following : **black = referee's comment**, red : quote by referee 2 from our previous answer to referees, BLUE= our new answer. We refer to lines and sections of the revised MS with track changes.

Dear Editors and referees,

We thank both referees for the time and attention spent in evaluating again our work, and the editors for handling and MS. Here is a synthesis of our actions following the previous round of reviews:

- All comments by referee 1 has been addressed, with answers to these along each of referee 1's comments in the pdf he provided.
- In response to referee 2's comment we have significantly improved our workflow which resulted in a new element in the MS dealing, mainly, with the treatment of false detections. We believe these additions, which do not alter our main conclusion, but rather strengthen them will answer the main issues that referee 2 raised. We also addressed minor comments (old ones and new ones) that we thought were fully relevant to the MS, and were helping its readability. Old ones are answered in this document; new ones are answered directly on the pdf file the referee provided. We are extremely thankful to referee 2 for his detailed and extensive review. His persistence has challenged us, but we think that the level of detail or extra work that he requests is sometime unnecessary, and would hamper the readability of the MS. So not all his comments have resulted in change to the text, and we have tried to explain why. The length of his reply has been also a challenge to account for, and actually writing the response to reviewer has been indeed, more complex, than revising the MS itself, while it should, generally be the opposite!

We synthesize below the major modifications made to the MS. They are presented in greater detail attached to the comment of referee 2 with line number referring to the MS with track changes.

We have made 4 significant improvements to our workflow and demonstration:

1. **The M3C2 algorithm has been improved on two aspects**:
   a. the uncertainty model accounting for point cloud roughness, point density and registration error ($LoD_{95\%}$) adheres more strictly to the two-tailed t-statistics when the number of points intercepted by the projection cylinder is smaller than 20, resulting in the LoD95% being up to 46 % larger in the lowest point density area (see L259). This model was originally presented in Lague et al. (2013) but simplified by using a constant value of 1.96 for the t-statistics. This slightly reduces the fraction of surface detected as significant change before segmentation. This has a marginal effect on the end result, but is more statistically robust when working with low point density data such as the pre-EQ lidar survey.
   b. a new iterative procedure detects the <u>rare</u> cases for which the projection cylinder could intercept twice the same point cloud, for instance in very narrow valleys or near ridges. If not corrected, the M3C2 distance was incorrect and would generally indicate a significant change when there was not. The iterative procedure checks for the consistency of the M3C2 distance for various cylinder depth (see L238).

   The combination of these 2 changes results in 1118 potential landslide sources after segmentation compared to 1270 in the previous version of the MS. These changes to the algorithm will be implemented in a new version of the M3C2 plugin in Cloudcompare that will be released when the paper will be published.

2. **A manually labelled 3D source inventory corresponding to 66 % of all potential landslide sources has been created with 2 classes: true landslide sources (384) and false detections (355) (fig 6c)** (see new method section 3.5, L475)**.** The labelling and interpretation was done using orthoimagery, the 3D M3C2 field (to identify nearby deposits) and detailed inspection of the point cloud. The sampling strategy was as follow: all clusters larger than 200 m² were labelled (as they are the most important in terms of volume). The remaining smaller clusters were labelled through stratified sampling ensuring a uniform spatial distribution and no area bias. All area intervals below 200 m² have 60 % of the clusters labelled to avoid impacting the scaling behavior of pdf(A) and the occurrence of a rollover. This labelled source inventory is used as follow:

   a. to quantify the prevalence of false detections in the inventory immediately after segmentation (section 4.2.1, Fig. 6,7) and evaluate their potential impact on scaling properties if they are not properly removed (section 5.3.2 of the discussion, fig. 14a). We document, as expected by referee 2, **a prevalence of false detections decreasing with size**. In forest free areas, the prevalence of false detections is relatively low (24.4 %) and corresponds to correlated elevation errors of the LiDAR data due to intra-survey lidar errors and inter-survey registration error. In forested areas, false detections dominates the inventory (80 %) due to ground classification errors in the pre-EQ data. As explained in the method section 2 (L171) we did not try to further improve the ground classification of the pre-EQ data because (i) we think it is important to clearly highlight the issue which is expected to be frequent in legacy LiDAR data of steep hillslopes with dense evergreen forest and low shot density, and (ii) we do not think the classification can be further improved because laser penetration in the forest was too poor to correctly identify the ground
   b. The labelled source inventory is used to evaluate the performance of filtering metrics (section 4.2.2 & 4.2.3) we apply to create the final inventory on which we compute total volume and geometric characteristics of landslide sources (see point 3 below)
   c. The labelled source inventory is used as a reference case for statistics on landslide sources (pdf(A), pdf(V)) which are unambiguously true labelled landslides, as opposed to predicted one (in which a small fraction of false detection may remain). It is thus used in 3 figures (11a, 11b, 14a)
3. **A new filtering metric is introduced**: **the closest deposit distance (CDD),** calculated along the downslope distance of any potential landslide. Along with other metrics derived from M3C2 and computed for each landslides (fig. 8), we use the labelled 3D source inventory to show how to obtain an optimal balance between preservation of true landslide sources, and removal of false detection. We show that the CDD in combination of the SNR is a very effective way of filtering out false detections while retaining a very large amount of true landslide sources in forest free areas (table 2, balanced accuracy = 0.93). In forested area, the results are slightly worse (balanced accuracy = 0.8, 50 % of true landslide are preserved, 94 % of false detections are removed) owing to the poor quality of the pre-EQ data classification. We also demonstrate that after filtering, false positives do not exhibit a size dependency in forest-free areas and only a weak size-dependent effect in forested area, so that if a rollover where to exist at those areas in the real landslide data, it would be largely preserved and not obscured by the increased occurrence of false positives (fig. 7, section 4.2.1). Finally, deposits, which we recall are not the core of our study, and have thus not be manually labelled, are filtered after the sources, using as condition that a true deposit must be connected to an upstream valid source (L538).
4. **A new predicted inventory has been calculated based on the optimal filtering metrics, and represents the reference landslide source inventory that we use for total volume calculation (source and deposits) and source statistics.**

Consequently the main changes in the organization of the MS are:

1. The issue of false detection is presented in the abstract (L17-21) and in the introduction to emphasize the critical importance of accounting for these issues (L140-145)
2. Method section:
   a. A more extensive presentation of error models in relation to comment MC1 of referee 2 and the new elements of the M3C2 algorithm has been added in the method section (section 3.1, L245-292)
   b. a new section presenting the treatment of false detection has been added (section 3.5, L 475) with the following sub-sections: section 3.5.1 : Construction of a labelled inventory, 3.5.2 Definition of filtering metrics, 3.5.3 Definition of a classification performance index. The nature and type of false detection that can occur with topographic change detection applied to landslide detection is described (section 3.5 fig 5). Figure 3 presenting the workflow has been revised to further highlights the treatment of false detection AFTER the segmentation.
3. Result section:
   a. section 4.1 is now restricted to the presentation of the result of the segmentation before the application of filtering metrics. Figure 6 presenting the map views of 3D distances and segmented data has been recalculated to reflect the results of the new M3C2, the significant change map with the map of forested area (to answer one of the request of reviewer 2) as well as the significant change corresponding to the river bed. Fig. 6c is actually new: it

presents the labelled dataset, the unlabeled data and the deposits. It allows the reader to see the location of labelled false detections. We have added a detailed view at each of these steps for the reader to better understand the segmentation and filtering of small landslides in forested area. The final inventory on which we perform statistical analysis, is now in a dedicated figure 9.

   b. An entire new section 4.2 Removal of false detections and the 3D predicted inventory (L631) has been added with 2 new figures: fig 7 presents the proportion of false detections before and after filtering as a function of landslide area. Fig. 8 shows the cumulative distribution functions of filtering metrics for labelled landslide sources and false detections. Section 4.2.2 Optimal filtering metrics explores the best combination of filtering metrics. A new table 2, synthesizes the results. Figure 6 of the previous MS (selection of the optimal SNR) has been removed. Section 4.2.3 The 3D predicted inventory summarizes the characteristics of the predicted inventory we obtain after application of the filtering metrics, with a dedicated new figure 9 showing the map of predicted true landslide sources and deposits and the predicted false detection and a new table 3 summarizing the number, area, and volume of predicted sources and deposits, and predicted false detection (sources and deposits).
   c. The subsequent part of the results (4.3 comparison with the 2D inventory, 4.4 Landslide sources area, depth and volume analysis) use the predicted landslide source inventory containing 433 sources as opposed to 524 sources in the previous version of the MS.
   d. A hillshade has been added to figure 10c to better highlight the presence of retrogressive scars in the forested area coherent with the pattern of deformation we detect (to reflect a comment by referee #1).
   e. Figure 11 has been updated with the new inventory and comments by referee 2 regarding centered bin in X axis (as in all figures now), comparison of depth/area with the relationships predicted by Massey et al. and Larsen. The pdf(A) of the new predicted inventory is extremely similar to our previous MS version. We have added the labelled sources data to the pdf(A) and pdf(V) which follow almost exactly the same trends than the predicted inventory sources. With this new manually labelled data, the lack of rollover above 20 m^2 in the 3D inventory is confirmed. The new inventory has only 4 points below the 3D minimum volume, and the mean-depth/area data no more hints at the existence of two trends in the D-A relationship. We have thus remove these elements from the MS.
4. The discussion has been slightly reorganize with a new section 5.2 . “Benefits of the 3D-PcD approach to create landslide inventories” where we have centralized most of the aspect pertaining to the comparison with 2D inventories. A new section has been added dedicated to discussing the results of the false detections (section 5.1.3 False detections and filtering metrics)

Other minor changes:

- 2D-sources areas are calculated and are now estimated in the same way than the 3D method (sum of core points) so that both source areas can be correctly compared. This slightly reduces the 2D-source areas (from 149,039 m^2 to 146,641 m^2). Moreover, the maximum 2D source area of 40,679 m^2 was a mistake and has been corrected in the new version of the MS."
- The new p(V) shows hint of a rollover above the theoretical lower limit of volume detection. We have updated this part of the discussion (section 5.3.3)

To facilitate the review process, we synthesize here our answers/modification in response to the outstanding main comments of referee 2 that he considered were not properly answered. More detailed answers are provided in the point by point response to reviewer 2.

1. MC1: Clearer definition of 1) landslides (i.e. what the inventory includes) and 2) what your inventory can and cannot be used for. We have changed the text in the introduction and discussion following some of the referee recommendations.
2. MC2: Elevation errors need to be better quantified and more thoroughly discussed
   - New elements in the MS on elevation errors (including classification errors) and how they may generate false detections due to spatially correlated errors. We also explain why elevation errors cannot be precisely quantified with the limited information we have. We also better explain our choice of a uniform registration error and the actual formulation of the LoD95%
   - We use our new labelled 3D inventory to demonstrate that the base inventory contains a significant fraction of false detections due to the above correlated errors with a larger prevalence

in forested due to classification errors. This highlights the importance of developing filtering procedures after the segmentation to remove as much of these false detections as possible.

- The discussion has been expanded to emphasize the importance of developing better elevation error models, the importance of point density and laser penetration for ground detection in forest area and the importance of post-segmentation filtering to remove false detections.

3. MC3 : Segmentation results in amalgamation but don't quantify its extent or impact : we have slightly changed the MS in the discussion to further highlight the complexity of amalgamated landslides that the 3D inventory exhibit. However, we consider that we have adequately answered this point in the previous iteration, even if it is not to the satisfaction of the referee, by showing that no current state of the art 3D segmentation approach manages to correctly segment complex amalgamated landslides, nor does the 2D inventory. As explained in our previous answer, even if we were to provide a manual segmentation, it would be highly subjective such that its relevance would be equally as questionable as the automated segmentation. Our MS was already emphasizing strongly the need to develop better 3D segmentation approaches, but also demonstrating that the current limitations are not impacting the total volume, nor the lack of rollover in the 3D data.
4. MC4 : Topographic errors propagate through segmentation to introduce a bias towards small landslides. Existing experiments to quantity this bias are insufficient. The reviewer was correct: the new addition of a manually labelled 3D inventory covering 66 % of the raw source inventory show that correlated elevation errors do produce a significant amount of false detection with a bias towards small landslides. This is something we were not able to properly quantify in the previous version of the MS, even though the SNR was already a good way to get rid of false detections in forest-free areas. The addition of the labelled dataset allows us to properly demonstrate that our filtering approach removes a large fraction of the false detections and that there is no bias left towards small landslides.
5. MC5: Findings that differ from previous work. We have addressed all outstanding comments raised by the referee when we deemed they were relevant.
6. MC8: Landslide object detection needs to be tested against an independent dataset. The new manually labelled 3D inventory and its exploitation answers the referee comment

Response to Referee 2 – David Milledge

All the text from Referee 2 is included below. Because the new very detailed review of referee 2, adds to another detailed review, I (D. Lague) must admit that never in my life as a scientist or associate editor did I have to write or supervise a response to reviewer that is so long and complex. The way referee 2 responds by quoting us then adding new elements makes it extremely complex to answer concisely, and not turn our response into a discussion forum which is not the point of a review. Also, we feel that referee 2 has difficulties accepting that our work has limitations. We consider that there are limits to the details and additional work a referee should request for a paper that he otherwise qualifies as excellent. We have stated where we considered that this threshold was crossed

This is an excellent paper that makes at least two significant contributions. The first is methodological, detailing a method with which to robustly determine topographic change over large areas that include steep slopes. The second is substantive, demonstrating that the size distributions (volume and area) and geometric scaling relationships for landslides differ from those previously found for landslides. The authors have considerably improved the manuscript since my last review. However, there are a number of major and minor comments from my previous review that have yet to be fully addressed. Below I have retained only these comments and explained (in bold) why I feel that they remain unaddressed. I have also made minor comments on the new draft of the manuscript by commenting on the PDF.

The workflow that you have introduced has great potential to improve the quality of landslide inventories. The paper is a significant and rigorous contribution because it: 1) introduces the workflow for a suitable case study, 2) shows that the workflow improves on alternative 3D methods and can detect landslides not detectable in 2D methods; 3) highlights common errors in 2D methods that have been proposed but rarely demonstrated, and 4) demonstrates for the first time that you can calculate a volumetric budget for landslide derived topographic change without the need for volume-area scaling relationships which are

known to (and which you show) introduce considerable uncertainty. It also opens up a discussion about what constitutes a landslide and how this differs from the things that we currently map, whether in 2D or 3D.

I agree with you when you say in your response that *"despite its limitations, it currently represents the state of the art in terms of 3D landslide detection and landslide inventory creation"*. **This is an excellent methodological contribution and I have only very minor comments on presentation of the paper in relation to this aspect of the work. As a methodological contribution I agree when you say that:** *"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.* **In relation to this, it is excellent that you:** *"provide all the elements (code, data) for other researchers to apply the workflow and reproduce our results, or apply the workflow to their data."* **The reason I am so demanding of the checks you apply to the 3D method relative to the 2D method is that the flaws in the 2D method are relatively well rehearsed in the literature but you are presenting the 3D method as a new (and better) technique. To do so you must demonstrate that this is the case.**

We think the new additions demonstrate that this is the case.

However, I strongly disagree that you *"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."* **I disagree because I think that non-trivial errors remain in: 1) the detection of small landslides, where some false positives remain due to spatially correlated errors; and 2) defining the boundaries of landslides, where automated segmentation continues to result in amalgamation. I know that I come to this with a bias: I have interpreted my own field observations as indicating a rollover in landslide size and have developed theoretical explanations for that rollover. So I am probably resistant to the idea that landslide size distributions lack a rollover. I'm trying to avoid this bias but may not manage it.**

We too have come up with a theoretical model for the origin of the rollover (Jeandet et al., 2018) and the results we show in this paper largely contradict these. We do not look specifically for controversy: we try to be as rigorous as possible given the data we have and processing tools we have developed. The referee himself acknowledges that our approach is rigorous. It is even more rigorous now, but the results we obtain are essentially the same. We are however extremely worried that the referee acknowledge himself he has a potential bias in assessing our work. As for the 2 issues raised by the referee:

1: we answer to this point in details in section MC2/MC4 of the present response: our new additions should clear things up for the referee.

2: this point was highlighted in the new discussion of the previous version of the paper and the answer to the reviewer (we again answer to this in detail below). In essence, we fundamentally disagree with the referee, and as explained before, the lack of a rollover is not due to the segmentation approach as amalgamation tend to limit the number of small landslides and favor the occurrence of a rollover (something that can be hinted at in figure 14c as Dm is increased). It obviously impacts the scaling exponent for large landslides, a fact that we illustrated in the MS, and still illustrate in fig. 14c. We made very clear in the previous version of the MS that our segmentation approach was far from perfect, and we are certainly not trying to hide this fact under the carpet !

You could easily address my outstanding major concerns by softening your claims. For example, you say in your response: *"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"*. **I broadly agree but I would say: 1) that the comparison "suggests" rather than** *"shows"* **because you cannot identify which inventory is in error only argue which is more probably the source of the error; and 2) that the under-detection is "present" rather than** *"extremely prevalent"* **because you can argue that it is very likely that some of the size-dependent bias between inventories is extremely likely to be due to 2D under-detection but some is also extremely likely to be due to 3D over-detection and you cannot currently identify their relative share.**

Two of your key conclusions (stated in the abstract) are that the manually mapped 2D inventory ***"severely underestimates total area and volume"*** **[L20] and that there is** ***"a systematic size-dependent under-detection***

***in the 2D inventory"* [L24]. However, both these statements are underpinned by an assumption that the 3D sources are correct (i.e. the ground truth) such that differences between them and the 2D inventory are attributable to error in the 2D inventory. This assertion needs to be justified in the paper but it is not at present. Instead you consistently assert and assume that in cases where the two inventories differ it is the 3D inventory that is correct.**

Following the reviewer requests we have soften some of our claims (see detailed comments). However the addition of a manual labelling of a large fraction of the 3D dataset allow us to better support our claims when it comes to the comparison between the 2D and 3D inventory. In particular, we can now better demonstrate the systematic size dependent under-detection because we obtain a similar tendency (fig. 15) if we use the labelled 3D sources or the predicted 3D inventory. We do not (and did not) claim that the 3D dataset was completely exhaustive, as we showed that it misses some of the landslides detected in 2D, and it cannot detect landslide smaller than 20m^2 while the 2D inventory can.

Another key conclusion of the paper is that the 2D size distribution has a rollover whereas the 3D distribution does not and that this is due to missed landslides in the 2D inventory. However, it is not clear that this is a fair comparison. The rollover is detected in the manual mapping on the basis of a reduced frequency of landslides in the smallest class 13-20 m_2, relative to the class 20-31 m_2 (which is the modal class). This smallest class is below the lower limit of detection for the 3D method. If you enforced a single consistent lower size limit for your analysis and censored all landslides smaller than this limit for both datasets then I don't think you would conclude that the manual mapping displayed a rollover. Note, that the x-axis values for the size distributions in Figure 8 are lower bin limits not central values this is potentially confusing and should be adjusted.

First, we now use central bin values as suggested by the reviewer, as opposed to the lower bin limit for all graphs. For the pdf(A) (fig.11a), it results in an even clearer rollover of the 2D labelled inventory as the first two points shows an increase in frequency with size. Second, it seems important to show that the 2D approach has a higher resolution than the 3D, and thus we do not see why we should censor this graph to the same minimum value than the 3D data. However, it is fair to say that we cannot state that the 3D data lacks the rollover that the 2D data exhibit. The only rigorous and strict conclusion is: if the 3D data were to exhibit a rollover if would be for sizes smaller than 20 m^2, a value much lower than any of the previously published values (see review by Tanyas et al, 2019). We have modified the text (in places highlighted by the detailed comments of the reviewer) to make things clearer in particular in the discussion section 5.3.2.

I think you make one further important finding that you could highlight in the abstract: you demonstrate the variety of types of topographic change that occur in response to an earthquake and show that existing 2D landslide mapping captures only a small part of that range.

Indeed, and we have added a sentence relating to that in the abstract (L38)

Your results prompt questions about what constitutes a landslide within these landscapes and how we should delimit them. This is particularly important for size distributions because the way that you define both your term landslide (to say what is in or out of the class) and the boundaries of your landslide in space on the basis of post failure observations will differ depending on the motivation for examining them. For example, your point on L420 highlights the complexity of mapping post-earthquake topographic change and relating it to processes. Should subsidence / retrogressive slumping upslope of a catastrophic landslide be included within the same source zone? Is this one landslide or two? The processes and perhaps even the timing of movement are quite different. But it is a very important point that these movements will not be captured in conventional inventories though there is widespread recognition of the processes you discuss based on field reconnaissance.

This is a very interesting comment. It addresses the question of the subjectivity of landslide delineation that makes automated landslide segmentation even more complex. We do not know if this comment was a request for further modification of the MS, or just a general comment.

In the following, to avoid adding further comments to our previous comments lacking line numbers, we have added in **blue and bold** line number that were missing. These line number refer to the new version of the manuscript with tracked change.

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 **(L.127)**: "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.

RE: definitions. Simply defining landslides as decimetre scale change not due to fluvial processes is accurate but will be a very unusual definition to the reader. You can help them to see that your definition of a landslide is consistent with theirs by adding a little more explanation and I think that would be very worthwhile.

Crozier suggests that: *"The three most widely used classifications involving landslides (Sharpe, 1938; Varnes, 1958 and 1978; Hutchinson 1988) separate 'mass movements' (Fairbridge, 1968) into two categories: subsidence (which is the vertical sinking of material-see entry on Land Subsidence) and those movements that occur on slopes. These' slope movements' are then usually divided firstly into 'landslides,' as defined above, and secondly into the slower, more widespread and ill-defined movements such as 'creep,'' sagging,' and 'rebound."* The landslide definition that he refers to is: "*the downward or outward movement of a mass of slope-forming material under the influence of gravity, occurring on discrete boundaries and taking place initially without the aid of water as a transportational agent."*

Crozier M.J. (1999) Landslides. In: Environmental Geology. Encyclopedia of Earth Science. Springer, Dordrecht.

I think you can make the case that most of the change that you detect can be classified as landslides following the definitions of Sharpe (1938), Varnes (1958, 1978), Hutchinson (1988), and Crozier (1999). But you need to make that case. If you explain the timescale over which the change occurs and the spatial limits of detection that you will ultimately impose then you can argue that everything that you detect should fall within the class of landslide. It would help your later discussion if you gave a summary of what that might include (e.g. slides where the failed material is entirely removed from the source zone and those where movements that are small relative to the length of their failure surface). It would also help to explain which non-fluvial mass movement processes are not detected, particularly: tree-throw, ravel and other forms of creep (because the movements are either too small, too localised, or too slow).

First we note that the definition of what constitute a landslide is highly variable (e.g., Tanyas et al., 2019), and that referee 1 did not have any trouble with the definition we propose. Authors publishing 2D inventories of landslides generally use a very loose description of "what's measurable" in terms of optical difference on hillslopes. While we can understand the need for the referee and some readers to place our results in the framework of existing classifications, we think that some of these classifications were based on a descriptive approach which favored highlighting the heterogeneity and peculiarity of sites, while in the end a landslide can be described simply as a large amount of earth and rocks falling down a cliff or the side of a mountain (Collins dictionary). We do not want to err too much on the descriptive side of things in the introduction and prefer to stick with a simple definition on which we elaborate during the discussion as a function of our observations. Because we now use the proximity of deposits in the manual labelling and through the CDD, and because the signal we measure is topographic change which can corresponds to both erosion and subsidence, we have updated this definition in the introduction as follow (L127 to L138):

"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 decimeters of negative topographic change associated with a downstream positive topographic change. Patches of negative (resp. positive) topographic change are called sources (resp. deposits) and correspond to erosion (resp. sedimentation) or subsidence (resp. accumulation). This definition therefore includes all the types of mass wasting processes involving the downward or outward movement of soil, rocks and debris under the influence of gravity, occurring on discrete boundaries and taking place initially without the aid of water as a transportational agent (Crozier, 1999)"

We have embedded the definition of Crozier and reference to Crozier as suggested by the reviewer.

The variety of landslide processes that we capture is discussed in section 5.2.1 and underlined in the abstract (L30) and conclusion (L1297).

It would perhaps also be worth saying that this definition differs from those commonly (implicitly) applied in 2D landslide inventories derived from satellite imagery since these rarely (or incompletely) capture slides where material is displaced by only a fraction of the failure surface. These inventories are censored by their ability to detect change from image properties and thus rarely capture rockfall source zones. The same censoring results in under-sampling of small landslides, landslides in bare or sparse vegetation and landslides obscured by forest canopy because these can't be confidently identified.

This is a very good point, but we think that stating this in the introduction would probably be confusing, because most readers will have no idea of what the 3D topographic differencing actually generate. We have added the following sentence in the discussion section 5.1.3:

"These inventories are also censored in the variety of landslide processes they can capture as they rarely capture rockfall source zones on very steep hillslopes and rotational/translational landslides where material is displaced by only a fraction of the failure surface."

<u>RE: expected uses</u>. Clarifying the focus on co-seismic landslide volume estimation is useful as is the section that you have added on the different processes represented in your inventory.

We did not understand to which part of the manuscript it referred or if it was a comment or a request for change. We did answer to one comment of the referee in the pdf version (L50) for the introduction which seems to relate to this.

<u>MC2) Elevation errors need to be better quantified and more thoroughly discussed</u>

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.

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.

I am not suggesting that all these areas need to be exhaustively explored but that the inferences that you draw from them must be stated with appropriate confidence for the uncertainty in the data that underpins them. Mass balance, which was your primary objective is largely insensitive to the errors that I

highlight here. Landslide size distributions and scaling relationships are potentially very sensitive to these errors.

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.

Your contribution in developing a workflow is extremely valuable and is not compromised by the continued presence of these errors. However, prospective users will also use this paper as a model for what the workflow can be used to calculate. Your discussion of errors and their implications is therefore important because it will influence not only how people interpret your results but also the capability and limitations of the workflow.

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.

Both these aspects have very considerably improved both the workflow and its application in this paper. However, you have not addressed my original concern about spatially correlated error in this comment.

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.

I do not see where this comment is addressed in the response.

This comment was initially addressed in the general answer to the referee comment L 955 to L 960 and in section 5.1.2 of the discussion, specifically in L718 to L724 of the previous MS with tracked change:

Now, we have added new elements in our MS regarding elevation errors:

1. A more extended presentation of the potential source of elevation errors in airborne lidar data in section 3.1 separating uncorrelated errors (survey noise, surface roughness) and correlated errors (time dependent attitude and position error; intra-survey registration error of flight lines; inter-survey rigid registration) (L276-292). These are introduced in the context of the distance uncertainty model we use (eq. (2) and how they have been evaluated in previous work by (i) a spatially explicit direct error model propagation, or (ii) via an empirical analysis on areas that are perfectly stable, horizontal and smooth (e.g., with a classical variogram approach as in Anderson, 2019 that the referee suggested). We explain why neither approach is feasible in our study site given the lack of (i) trajectory data (for a direct model), and (ii) of large enough stable and flat surface (for an empirical approach). Hence, and let me stress that because it seems that the referee think we are lazy or uniformed when it comes to LiDAR data generation and analysis: it is not that we do not want to quantitatively study the properties of elevation errors or generate a non-uniform *reg* model. It cannot simply be done at present, and we expect that it will always be the case when future researchers will apply our workflow for landslide detection.
2. Following this, we have added a new paragraph (L293-308) to better explain our choice of *reg* compared to a more classical approach that would be based on the stated errors of the two LiDAR dataset. In essence, because there are no GCP in the study area, the stated errors of the LiDAR dataset are not really good constraints, and we prefer to have an empirical evaluation based on the standard deviation of intra-survey or inter-survey M3C2 distances on stable area. We expect the lack of GCP will be frequent in future work using our workflow for landslide detection in steep relief.
3. The addition of the manual labelling of true landslide sources and false detections allow us to highlight the occurrence of these errors and study their prevalence, their size distribution, and highlighting the critical role of ground classification errors in forested area (a new result).

As far as I can tell: 1) there are spatially correlated errors in your difference surface; 2) these errors will not be captured by the SSDS analysis; 3) correlation of errors implies that if one core point has errors large enough to exceed LoD then there is a non-trivial probability that one of its neighbours will also have errors that exceed LoD; and 4) the distribution of erroneous patches will be strongly right skewed (i.e. smaller patches more probable than larger patches). I would be keen to know whether you agree. On this basis, I think you must quantify the impact of spatially correlated errors on your size distribution if you are to argue that your measured size distribution is the 'true' distribution or even that it is more correct than the 2D distribution.

Answers to the above comments:

1) we agree there are spatially correlated topographic change associated to elevation errors and this was stated in all versions of the MS, in particular when it comes to flight line misalignments and to ground classification errors of the pre-EQ data. In the previous MS with tracked changes this was indicated:

- in section 3.3.1 (registration error estimate) in L291 to L295 given that bias and error between flight lines for each survey are spatially correlated errors.
- In the result section 4.1 : L461-464 (effect of SNR filtering in removing low amplitude topographic change related to flight line misalignment)
- In the discussion : L709-710

But what matters is the new version in which the above parts have been largely rewritten and extended:

- we have added a detailed description of some typical false detection (new fig. 5) due to spatially correlated errors originating from incorrect ground classification data and flight line misalignment in the pre-EQ. Fig. 6c now shows the labelled false source detections.
- Section 4.2.1 presents the prevalence of these false detections and their size characteristics, and other M3C2 derived metrics. The prevalence of false detections in forest-free area is relatively low (24.4 %). It is however difficult to separate the influence of intra-line, intra-survey and inter-survey errors on these and we did not try to separate these, and we do not see how it would be easily feasible. In forested areas, false detections dominates the inventory (80 %) and are mostly due to ground classification errors in the pre-EQ data, an important result of the new MS highlighting the importance of high quality LiDAR when working on evergreen forests.
- A new discussion section 5.1.3., further address the occurrence of these errors and the importance of filtering.

2) yes, the SDDS is not designed to evaluate these effects. It is now only used to evaluate the optimal parameters of normal scale and projection scale for M3C2.
3) yes, but these errors are largely filtered out.
4) indeed, and using the labelled data this now shown in the new figure 7 highlighting the prevalence of false detections as function of patch size, as well as in fig. 14a of the pdf(A) of false detection compared to true landslides

<u>RE error sources</u>: 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.

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).

This is a useful explanation but the decision to assume that registration error is spatially uniform still needs justifying in the text in a way that addresses the concern that long-wavelength errors might combine with short wavelength errors to generate patches of erroneous change in some places and break up patches of true change in others.

This is a case where we consider that the referee do not know where to stop in his requests. We will not add another page of MS to address this. We already discussed at length in the previous version (see response above) why we cannot generate a non-uniform registration error, and highlighted in the discussion that it would be an important development to have one. We show that our main results are not altered by increasing reg (fig. 14b). And most importantly we now show that our filtering approach get rid of most of the false detections. That being said the new version of the method section in which we present the LoD contains a justification of our choice of reg that should answer the reviewer requests.

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

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 1000 *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.* 1005

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.

I agree, that segmentation is a subjective problem but you have made it reproducible by removing the subjectivity. The problem is that the best reproducible (i.e. automated) segmentations still perform poorly (with respect to the segmentation that a human mapper would choose).

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.

This is a really important point and could be a key contribution of the paper. Your results show that it's complicated. Far more complicated than we capture in conventional 2D inventories.

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.

It isn't clear why reproducibility is favoured over skill. If you think that manual segmentation would outperform connected component segmentation it seems strange to continue with automated approach because it is reproducible. You say that 2D manual mapping "misses a very large number of landslides" but this is on the assumption that the 3D inventory is correct.

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 1025 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 1030 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.

The analysis that you have added demonstrates that the problem is not that your particular segmentation approach is worse than the alternatives but that automated segmentation itself is problematic. You make the argument above the subjective segmentation is equally problematic. I think making this point in a more detailed discussion of the problem of segmentation (both automated and manual) would help to address my concern here.

We have a fundamental disagreement with the referee on the value of adding a manual segmentation to our approach and we consider that the requests for further discussions are unnecessary diversions from the core of the paper and would impact the readability of our work. We already have put a strong emphasis in saying that our segmentation approach is far from perfect and needs to be improved in many places of the MS: abstract (L35), discussion (L916-942), conclusion (L1307). Referee 1 has strictly no problem with the limitations of our approach. We also explain(ed) (L1169-1171) that the lack of rollover cannot be related to the current tendency of the segmentation to generate amalgamation as this tend to reduce the number of small landslides and would indeed favor the occurrence of a rollover. And we show (fig. 14c, fig S12 and S13) that, obviously, the parameter *Dm* of the segmentation changes the values of scaling exponents.

Now, with the new elements on the prevalence of false detection and the importance of filtering, we have added new elements in the discussion highlighting that automatic segmentation is a necessary step to generate a preliminary set of potential sources and deposits on which filtering is applied (L925) : *"As we aim to apply our workflow to very large dataset with potentially several tens of thousands of landslides, and that false detections filtering need to operate on segmented patches, automatic segmentation is a mandatory step.". We have also added (L938) : "Manual segmentation can also be envisioned as a refinement after the predicted true landslide sources and deposits have been produced, but is non reproducible and time consuming when applied to very large datasets."*

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

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 1045 *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.

Most of my concerns in this comment have been left unaddressed. MC2 focusses on topographic errors but your response to MC2 doesn't deal with the problem of spatially correlated errors. Because you are interested in the size of patches generated by thresholding the difference surface it is essential that you examine the spatial structure of the errors. I will deal with each of my comments in turn reflecting on

whether they have been addressed in the new manuscript even if they have not directly been addressed by your response to this comment.

First, you did 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. 9 real landslides)".* **You remove reference to an SSDS test to set the minimum area but retain that test to optimise the SNR. However, when you introduce it you do not recognise that it synthesises uncorrelated noise while the two surfaces that you are differencing both include spatially correlated elevation errors. You do not include any description or explanation in the text for the "stable areas error" shown in Figure 4, these errors appear to be spatially structured on multiple length scales from tens to hundreds of metres.**

In essence, there is no way for us to evaluate how the various elevation errors impact landslide detection, because we cannot evaluate properly the spatial structure of these elevation errors (see response to MC2). We could play with synthetic data, adding another 3 pages to the MS showing that if we add a 2D sinusoidal error to the two dataset with a specific wavelength we would certainly detect significant change if we do not properly set the registration error, and if not filtered out it would generate a landslide size with a certain distribution. However, we will not do it because: (i) we have no more time to do it in the time frame of the PhD project of Thomas Bernard which finishes soon; (ii) it is not necessary to support our conclusions; (iii) it would not help in getting better inventories, because in the end the critical step is the identification of true landslides and false detections which requires well-chosen filtering metrics.

We now show examples of false detection occurring on spatially stable areas (fig 5b, fig. 6c, section 3.5.1, section 4.2), but also the much more predominant source of error related to ground classification errors in the pre-eq forested area (fig. 5a and 6c). What matters most, is that our new filtering metrics (CDD+SNR) in non-forested areas results in excellent removal of false detections (97% precision) while preserving a large part of true landslides (79%) and all true landslides above 40 m^2. The new final predicted inventory does not contain any source or deposit in the stable areas (L709).

Second, I argued that *"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"* **and that** *"You need to quantify these detection biases before you can make these claims."* **I'm still not convinced by this claim because I still don't think you have excluded the possibility that the lack of rollover is solely due to detection errors. You need to quantify the size dependent detection bias in the 3D inventory and/or to considerably tone down your claims about rollover in this and other 2D inventories being due to under-detection.**

Using the labelled 3D sources and false detections, we now demonstrate that when considering only manually validated landslide source, the rollover is still lacking (fig. 11a). We also show that if no filtering were applied, then the statistics of small landslides would be indeed significantly biased towards small landslides as expected by the referee (fig. 14a) , in particular in forest area given the large prevalence of ground classification errors. We hope the addition of the labelled data will convince the referee. However, we will not tone down our claim because we are even more confident that the lack of rollover above 20 m^2 is a real feature of our dataset.

Third, I suggested that you: *"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".* **I don't see a response to this suggestion here but when I raised the same point in a minor comment (related to L196). you responded that** *"Applied on stable area, the workflow does not detect any landslide."* **This result would definitely be worth reporting! However, I think we must have misunderstood one another, I can see many patches of significant change (>150) within the stable areas, most of these patches of significant change are removed in Figure 5c. Is that because they are smaller than 20 m^2? However, even after this filtering I can still see several landslides within the stable zones in Figure 5c. I phrased this as a suggestion in my previous review but I really think this is one of the few opportunities that you have to build confidence in your method. It remains a weak test because you chose the stable areas based on areas of limited change in the difference maps and because they**

oversample non-forest vegetation but in the absence of field checks to the inventory this remains one of the best tests I can come up with. My second suggestion to *"collect a landslide check dataset using independent observations"* **is dealt with in a separate major comment MC8 and doesn't need further discussion here.**

We believe our manual labelling of true landslide and false detection offer a richer and more complete approach to the issue of false detections due to elevation errors than just looking at the statistics of false detections only on stable areas. As stated above there are no source on stable area in the final predicted inventory (L680)

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 1115 *and L460). These result should be compared with results from previous studies.*

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.

This has not been addressed. I commented that you needed to compare your results with those from previous studies with respect to scar and deposit depths and scar and deposit areas. I didn't find this new discussion nor a response to explain why it was not necessary.

Because our study focuses on landslide sources, and our analysis of deposits has not been as detailed as the sources, and the discussion is already quite long, we choose to remove this part.

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.

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).

Why do you expect that deposits are more easily detectable than source zones? I would have expected the opposite. In my experience deposits can be very thin, (<50 mm) patchy and extensive whereas source zones are far more coherent.

Deposits now makes it to the final inventory only if they are connected to an upstream source (L539). This resolve the initial issue that the referee had.

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.

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 1135 forest, while it is detected in our approach. We think the comparison with the new 2D inventory will resolve the reviewer's reserve.

How do you know that small ones that occur on less dense area are difficult to map? Is this on the assumption that your LiDAR inventory is correct? If you go to the landslide locations predicted by the 3D method do you find evidence in the orthophotos that there is indeed a landslide at that location (even if it wasn't independently mappable)? This would help to build confidence in your method. The point about vertical subsidence is important and you do a nice job of demonstrating the plausibility of the claim that this is real change. It prompts a series of questions about representation of these landslides within an inventory derived from surface change but you deal with this nicely in the discussion. My only suggestion is that you prepare readers for this finding in the introduction by adding a more complete explanation of the types of landslides that 2D and 3D inventories might include (see MC1).

The referee was correct in doubting the reality of the landslides in forested area! Manual labelling showed a high prevalence of false detection (80%) due to ground classification errors in the pre-eq lidar. We now describe in greater detail these errors (fig 5a, 6c), and emphasize in the results section 4.2.1 and the discussion section 5.1.3 the importance of high point density LiDAR, conducive of less ground classification errors, for change detection in evergreen forested area. We however emphasize, and now better describe with the addition of fig. 10 c (hillshade view) that large subsidence patterns consistent with the occurrence of scars visible in the post-eq hillshade DEM can be detected with our approach

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 timeconsuming (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 1160 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.

The 2D inventory considerably strengthens the paper and addresses the comments in the first paragraph above. However, much of the second paragraph remains unaddressed. You have added analysis of the two inventories and discuss false positives for the 2D inventory where deposit is incorrectly mapped as source zone, this is a secure result and is exactly the type of analysis I was looking for. You use your observations and theory/logic to argue that one method is correct and the other is in error. The remaining areas of disagreement you assign as false negatives for the method that has not identified a landslide at that location. This is not a secure result. You have no objective way of establishing which method is in error (i.e. whether this is a false positive for one or a false negative for the other) and you don't provide any justification for why disagreements should always be treated as false negatives. In fact there is good evidence to suggest that the 3D inventory should contain false positives and that these false positives likely have a strong size bias. First, the SSDS test (which itself is a very weak test because it assumes that errors are uncorrelated in space) results in artefact landslides. Second, the inventory contains landslides in the 'stable' areas of the study area. This problem propagates into the discussion where you describe disagreement between inventories as error in the 2D inventory under the implicit (but untested) assumption that the 3D inventory is correct. I don't think you could use this language in the paper even if the tests above generate only small numbers of false positives from the 3D inventory.

The addition of the manually labelled dataset (which contrary to what the referee states, was extremely time consuming) answers the problems raised by the referee as we have now a very high degree of confidence in the 3D source inventory that we use to compare with the 2D inventory. However, the text comparing the 2D and 3D inventory has not greatly changed because the 3D inventory we used in the previous iteration of the MS was very close to the present one. We have mostly only updated numbers. We consider that the comparison between the 3D and 2D inventory (result section 4.3 and discussion section) are covering all the topics above and clearly highlighting the limits of the two types inventories.

As for tests, the SDDS is no more used to validate the approach (only for parameter estimates for M3C2), and we clearly state in the text that the final inventory of sources does not contain sources in the stable area (L709)

Outstanding Minor Comments from previous review

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.

I don't see Dolan and Rhodes (2016) cited in the text. The sentence above should be added to the text. You have added a manual quality check and reprocessed the data to remove non-ground-points as a result. That is a good additional step that you have introduced since the last version. However, I don't think you can

simply point to the SI for the details of this analysis. This is a key step in your method and should be included in the article itself.

Dolan and Rohdes (2016) was an incorrect reference. The report is Dolan (2014) and is now in the bibliography.

It turns out that our attempt at improving the classification was not successful as a large number of ground classification errors remain (e;g., fig 5a and section 4.2.1). The classification improvement is not a key step of our workflow because the user faces two configurations: (i) either the two datasets are high quality lidar survey (as the post-eq data) in which case the classification provided does not need to be improved (ii) either it has poor legacy data as the pre-eq data, where there's basically not much to do. We have added more details in section 2 on the classification (L204-208), and added the following text:

"We did not attempt to further improve the classification as these errors are expected to occur in low point-density LiDAR survey of evergreen forested areas and will generate false landslide sources that our workflow should detect and filter out We note that the classification refinement is not a critical component of our workflow and that other classifications algorithms (e.g., Sithole and Vosselman, 2004) could be used to improve or check the quality of the LiDAR ground points before the application of the workflow."

The discussion also contains reference to the issue of ground classification in relation to the quality of LiDAR datasets (section 5.1.3).

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 1190 *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.

You have not changed the manuscript in response to my comment. Your response is useful, particularly: *"the distance is computed between the two positions along the normal vector"* and *"average positions are defined as the arithmetic mean"*. You could easily amend the sentence to clarify this: *"…as the distance of the arithmetic mean positions of the two point clouds along the normal vector"*.

Indeed, as we explained in our answer, we did not change the text as this is presented in the original M3C2 paper which we consider is a classical paper (> 700 citations) that should be read by anyone wanting to use our workflow. So this should not have been a surprise for the referee that we did not change the text. That being said, we have modified the sentence presenting this aspect of M3C2 following the suggestion of the referee (L230)

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.

You have not updated the text to reflect this discussion. You do mean: 'if the cylinder intersects <5 points in the second point cloud'. If so, this should be added to the paper. Your explanation above is useful but I understand it to mean something different to what you say in the paper.

We have clarified this classifiable feature of the M3C2 algorithm by specifying that no distance is computed in the two point clouds do not overlap or if one of the point cloud has missing data (L233). A distance is always computed if there is at least one point in each cloud, however a distance uncertainty is calculated only if there are more than 5 points in each cloud. This point is mentioned later in the paragraph (L268)

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.

First, I think you can evidence your statement above from your sensitivity analysis and doing so will strengthen the paper. You demonstrate that changes to the LoD do not alter your main findings and the changes you explore cover the range that you would expect from changing the threshold CI from 90-99% (i.e. 0.39-0.52 m). This strengthens your argument that the statistics are robust to model choices.

We do not think it is relevant or necessary to discuss why choosing 95% is better than another confidence level as this is in the end an arbitrary choice. As the referee emphasizes himself, we show in the discussion that our results are robust to the choice of reg (which translates into much larger LoD). The MS is already quite long, so we have not modified the text.

Second, having read the original M3C2 paper (Lague et al., 2013), it discusses the confidence interval and registration error but took a different approach to estimating registration error so it is difficult to translate directly between the two papers. They provide only a brief description of the theoretical basis for the LoD equation citing a statistics textbook.

I found the description of James et al (2017) who you cite and who cite Lague et al. (2013) very useful: *"where reg is the relative overall registration error between the surveys, assumed isotropic and spatially uniform (Lague et al., 2013). Note that Lague et al. (2013) took a conservative approach by adding reg directly (as a potential systematic bias)"*. A similar statement would be useful in your paper to explain that you estimate local uncorrelated random errors using the first two terms and systematic errors under the assumption that they introduce a spatially uniform bias with the final term.

Anderson (2019) describes an approach similar to yours, but highlights spatially correlated random errors as a key component within error analysis for surface differencing and includes this as a term in his total error calculations (eqn 21). This term is missing from your error propagation but seems likely to be very important, particularly because you are interested in the size of thresholded difference patches. Anderson (2019) also argues for direct characterisation of errors within 'stable areas' similar to my suggestion in MC4.

Anderson, S.W., 2019. Uncertainty in quantitative analyses of topographic change: error propagation and the role of 1245 **thresholding. *Earth Surface Processes and Landforms*, *44*(5), pp.1015-1033.**

Here we have added much more justification and explanations on our choice of *reg*, as well as the form of the LoD95% in section 3.1. This changes are also tightly linked to our modifications in relation to the evaluation of elevation errors. We now cite Anderson 2019. His approach is interesting, but also amount at a spatially uniform *reg* as we explain in the text. The best way to improve the level of detection would be, as already stated in the discussion (section 5.1.2 registration error), to build a spatially explicit direct elevation error models for each survey.

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.

You now say: *"The M3C2 definition of the LoD95% makes the conservative choice of adding the registration error to the combined standard error related to point cloud roughness, rather than taking the square root of the sum of squared standard error and squared registration."*

First, I agree that your equation is conservative in the sense that it results in a larger LoD but I don't see a justification for this functional form either here or in Lague et al. (2013).

This is an arbitrary choice to make the uncertainty model more conservative. This is now clearly stated L307: *This arbitrary choice similar to Lague et al. (2013) ensures that the frequency of false detection of statistically significant change is below 5%, at the expense of a reduced capacity to detect real small topographic changes close to the $LoD_{95\%}$.*

The best description of a framework for propagation of both random and systematic errors that I can find is in Anderson (2018). It is similar to yours in its approximation of random and systematic errors (see Anderson's eqns 12 and 20) but differs from yours in how these are combined (see eqns 21-22). Can you explain the difference?

As above this is an arbitrary choice and we were explaining it in the previous version of the MS (L304): *the M3C2 definition of the $LoD_{95\%}$ makes the conservative choice of adding reg to the combined standard error related to point cloud roughness, rather than taking the square root of the sum of squared standard error and squared registration error (e.g., Anderson, 2019; Joerg et al., 2012)*

Second, you have not explained why standard deviation rather than standard error is used for the systematic error (reg). I think this is because standard errors are used to approximate random errors (under the assumption that they are uncorrelated) but standard deviations are used for systematic errors under the assumption that these are perfectly correlated (see Anderson's eqns 1-4, 12 and 20). If that is the case it would be helpful to explain it in the text.

Indeed, this was implicit, because taking the standard error with > 1000 core points over which we calculate M3C2 would result in reg ~0. We now state explicitly that we treat reg as a systematic error (L276) and that it is the standard deviation of M3C2 distances (L301).

Third, I'm happy for you to retain a notation that is specific to registration error but suggest sigma with reg as subscript would make it clearer that this is a standard deviation.

Thanks for the suggestion, but we did not follow it up, because adding a sigma to reg could be confused as the standard deviation of *reg*.

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.

I do not see where you have modified the text to reflect this response.

We did not change the text, and we still haven't because we think these are unnecessary details.

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).

This is still not clear both in the response and in the text (which has changed very little). L211 you say *"This is generally obtained by trial and error"*. This was what I was referring to when I asked *"what is the objective was that you are optimising"* by trial and error. Re-reading the text I am not sure whether it was obtained by trial and error in this paper. If not then how did you find out that the maximum observed change in the study

area was 30 m? If p_{max} is designed to prevent anomalously large changes how do you verify that a change is an anomaly? Something seems ircular in the argument as it is currently presented. I don't understand the connection to SNR filtering and that doesn't currently feature in the text.

The rare issues we had with the projection cylinder intercepting twice the same cloud has been sorted out with a modification of the M3C2 algorithm. We have removed the line referring to this issue, and added information on the algorithm at lines L237-241.

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.
This result is definitely worth reporting. If there are any remaining patches then you should report the size distribution of these patches as they give an indication of the expected bias that elevation errors will introduce into your size distribution.

There are no landslide source remaining on the stable area in the final inventory. False detections on stable areas are studied through the metrics we analyze in comparison to true landslides. We also now report the pdf(A) of false detections (fig. 14a) for all areas (there are not enough on stable areas to construct a robust pdf(A)

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.
I agree that it can be defined differently but you are reporting your method rather than the method in general so you should define it as you have used it.

reg is now define much more extensively and in the first section of the workflow (section 3.1, L244-276)

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.
It would still be useful to show the location of the points on one of your maps.

Useful maybe, but we don't think it is necessary nor actually feasible to show them clearly on a map.

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.
This is useful discussion, I couldn't see it in the revised manuscript. It would be useful in your description of Figure 5.

With our new filtering metrics, a deposit of the final inventory cannot exist if it does not have a upstream valid source

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.
This is useful discussion, I couldn't see it in the revised manuscript but I wasn't sure whether I was looking for parts of the text in your response above or something different.

With our new cleaned inventory, our D-A scaling is very close to the Larsen et al. (2020) relationship for soils (fig 11c) The original comment is irrelevant. We do not want to discuss potential causes for differences, when we do not observe one.

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.
This comment has not been addressed. I cannot find new text comparing your results with expectations from other studies.

We have removed this sentence because our analysis and treatment of deposits is not as detailed as sources.

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.
This comment has not been addressed. You have not altered the text to recognise the debate around the form of the right tail. I don't think you can simply add Medwedeff to the current citation. It reads as though Medwedeff et al. are among those arguing for power law scaling when I understand their paper to argue the opposite.

Indeed, this was ambiguous. We have added the following text: "and although the power-law nature of the tail is debated (Jeandet et al., 2019; Medwedeff et al., 2020)"

Figures

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 m²) that we consider and the minimum significant 1340 depth (~0.4 m).
The minimum area and depth act together to set an absolute minimum detectable volume (when both area and depth are minimised). However, there is also a depth dependent minimum detectable volume that is set by the depth constraint alone. At present the horizontal line that you use to highlight the absolute minimum volume might be misinterpreted by some to be the area dependent minimum volume. You can easily fix this by adding the area dependent minimum volume. It will be a straight line in log-log space and will pass through the points (20,8 and 20000,8000). It would be useful to include this on the Figure and would also bring make the interpretation of the dashed lines in the figure and inset internally consistent.

This has been added to figure 11C (initially figure 5 in the first version of the MS)

Supplement

S1: this text should definitely be included within the paper itself this is a key part of your method. You should also report the parameters that you used for this analysis and the parameter values that you chose, preferably with a justification.

As explained above (response to minor comment L63), the improvement of the classification is not a key part of our method. To keep the method part of the article as compact as possible, and given that we have significantly lengthen the MS with new additions, we have kept this part in the supplement. Our description of the method and how we choose parameters should be enough for anyone to reproduce. We also now provide a classical reference to ground classification in the main text that the interested reader can follow.

L11: "4 standard deviations" I have three questions here: 1) what sample is the standard deviation being calculated from? 2) Do you remove points that differ by >4sigma in positive and negative? I'd expect vegetation to result in only positive residuals. 3) How did you choose the 4 standard deviations threshold?
To answer to the referee's questions:
1) The standard deviation of the cloud-to-mesh distance is calculated from the pre-EQ point cloud. This information has been added to the text to clarify things. (L10).
2) Indeed, we only removed points having a positive cloud-to-mesh distance with a standard deviation >4 as classification errors only result in positive residuals. The text has been modified to include this information (L11).

3) This choice has been found by trials and errors. If chosen too low, point located on steep ridges or steep valley bottom will be also removed as the surface interpolation tends to smooth the surface. This is something we wanted to avoid. Using 4 standard deviations limited this. The following sentence has been added to the text: ” This threshold has been chosen, by trials and errors, to preserve points located on steep ridges and steep valley bottoms that will be removed if chosen too low due to the smoothing of the surface by the interpolation.”

L12: did you repeat three times because there were no outliers after that? If so you should report this, if not you should explain why you chose to stop after three iterations.

The procedure has been repeated only 3 times as further iterations filtered out points located on ridges and valley bottoms. The last sentence has been modified as follow : ”… as further iterations removed significantly points located on ridges and valley bottoms.” (L15).

Figure S2: This is a useful Figure but I can’t distinguish the flight lines based on the legend information, the line styles are not sufficiently distinct. Dashed lines are clearly visible and can be distinguished from the solid line. I think you will need to use different line styles to enable the reader to distinguish this many lines.

The legend of the figure has been modified and now show the dashed lines as suggested.

I don’t understand how you can have only one reference line in Figure S2 and Table S3. Is this a flight line from the post-EQ set? Or is this some combination of points from multiple lines? Either way I think you need to explain this, it will affect how the reader interprets Table S3.

For the pre-EQ point cloud, a 3D-M3C2 calculation is performed between each flight line and the reference line n°215 as shown in fig. S2. For the post-EQ point cloud, the reference line is the n°301. This is now indicated in the caption of Table S3.

S9: “Determination of forested area” Is this a standard technique? It would help if you could give a reference for the technique. I would like to know how returns are classified (i.e. how different do returns need to be to be classified as two distinct returns). However, if this is a very standard exercise it is fine for you simply to point to a reference.

The proxy we use to determine forested areas does not correspond to any standard technique used in previous publications. However, given the quality of the ground class of the post-EQ point cloud (98% of accuracy), we are confident that points with a number of return (NoR) of 1 is located in areas without vegetation. Points with higher number of returns thus are very likely comes from the structure of the vegetation. We think that our proxy is sufficiently indicative whether a landslide is forested areas or in bare-rock / bare-ground area given the strong correlation between the NoR and the vegetated areas.

Figure S10: It would be useful to give more detail in the caption. Something like “corresponding to the number of targets a laser pulse has intercepted” would probably be sufficient.

This figure is now the figure S11. The following sentence has been added to the caption: “The laser returns correspond to the number of targets a laser pulse has intercepted.”

Figure S12: This is an interesting plot. How do you calculate slope here? Is this based on the gradient of the core points? How do you explain the large number of sources with very low slope? Are these all associated with deep seated failures? This slope data provides another really useful way to check your dataset and therefore to build confidence in your results. You should plot slope for each core point against the size of the patch to which each core point belongs. Core points with low slopes associated with large patches may indicate deep seated landslides; those associated with small patches are likely to be errors because landslides require a steeply dipping failure surface (>20 degrees?) to move. Another useful approach would be to examine average slope for each patch plotted against patch size. As above, small patches with gentle gradients are likely to be errors.

This figure is now the figure S11. The slopes are calculated, for each core points in the 3D predicted sources, from the normals of the pre-EQ core points. The large number of core points with a low slope corresponds to areas of vertical subsidence of translational landslides as shown in Fig13b, or retrogressive slumping upslope of a landslide as show in figure 10c. The proposed approaches are interesting to investigate for future analyses but we consider that we have done enough in the present version of the manuscript.