Reviewer	Section Number	Reviewer comment	Author response	Reviewer response to author action
AE	**1.1**	In general, reviewers suggest that the authors re-proofread the report with a specific eye for conciseness and clarity. The introduction is quite redundant and can be trimmed for ease of reading. However, the main text contains too little information about specific algorithm implementation for readers to understand the models that are used and how they are parameterized. The readibility and interpretability of the report would be greatly improved by minimizing the redundancy in the introduction and providing the necessary information to understand model choice and parameterization, which is currently buried in the appendices.	We made a significant attempt to improve readibility and clarity and trim where possible. However, due to the nature of our program, some longer descriptions and detail need to remain. We recognize that this level of detail is not common in a peer-reviewed document. We attempted to add some model detail to the main text.	
AE	**1.1**	Many readers will be unfamiliar with the terms "bedrock hollow" and "inner gorge." A diagram and description in the text is needed to adequately explain these features. Bedrock hollows and inner gorges were "mapped based on definitions provided in WAC 222-16-050"; however readers need enough information in this report, so that they do not have to go sift through WAC 222-16-050 to understand the definitions of these geomorphic features.	Basic definition, with thresholds and processes, of these two landforms has been added to Section 1.1, and in table 1. The WAC definitions are very brief and not very useful (to your point), but Section 16 of the Board Manual with a link immediately below Table 1 has figures and photos. Added reference to which section in the Board Manual. Added description of Bedrock hollow and inner gorges	
Reviewer 2	**1.1**	How are background rates of landsliding measured?	Here, we are referencing a performance target written during the Forests & Fish Negotiations in the 1999-2001 era; referencing this target is a basic requirement of our overarching process as those of us who are scientists communicate to Policy. That having been said, we haven't actually figured out how to measure this performance target and a few years ago discussions with Policy resulted in a specific project removed from the CMER Work Plan. However, CMER is just now revising the Performance Targets and this legacy target remains in place until this current process replaces it. We have added a sentence to the document to explain HOW we practically interpret this performance target.	
Reviewer 2	**1.1**	Does this require pre- and post-event topography data? If the bedrock hollow fails, is it still identifiable as a bedrock hollow? Does its spatial extent change?	Lidar change detection does require pre- and post-event data. Yes, bedrock hollows are identifiable both pre- and post-event - typically, the debris flow initiates in the lower center of the feature, in a much smaller area than the whole, concave landform. The outer edges of the bedrock hollow usually remain unchanged - the outer concave area contributes groundwater that drives the failure and, by ravel or creep, sediment that fills the actual scarp over a period of decades or centuries. Occasionally, the area we identify as the bedrock hollow increases either uphill or off to one side edge such that the upper failure scarp extends beyond the original concave area. This also happens along the upper edges of inner gorges. Of particular note is that with pre-event data, we can characterize many terrain elements of the bedrock hollow before failure, and we can compare the area of the concave feature between pre- and post-event data.	
Reviewer 1	**1.1**	For the non-expert or somebody in an adjacent field that may have a different intuition for these landforms, what is a bedrock hollow and inner gorge? For instance, my background had me puzzled a few times by what exactly the practitioners here consider to be an inner gorge. I didn't have any issue with the authors' approach to hollows, but my comment still applies. More thorough definitions will aid the reader immensely as they independently evaluate the maps created by the experienced practitioners and automated techniques (I believe currently there is just a reference to a manual with definitions). I think a schematic diagram would also be useful.	Basic definition, with thresholds and processes, of these two landforms has been added to Section 1.1. Section 16 of the Board Manual with a link immediately below Table 1 has figures and photos. Added descriptions.	
Reviewer 2	**1.2**	The difference between the object-based image analysis and the virtual watershed methods are not clear. Why would we choose Object-Based Image Analysis over a watershed-based method? Additionally, why use a virtual watershed over more straightforward topographic metrics (i.e., GIS-based delineation of watersheds)? What is the advantage of the virtual watershed? What are the operational differences between these models? Do they have different tendencies in the production of outliers? Do they have different computation times? What decisions does a user have to make and how ad hoc are these decisions?	All GIS-based methods to delineate landforms use aspects of object-based image analysis. Both OBIA and VW are GIS-based methods. The advantages of these methods for this study is that members of the project team had experience using them, there are established protocols for their use described in the literature, OBIA is well implemented in eCognition, and VW is implemented in an ArcGIS interface (NetMap), in R and python packages, and can be run natively via command line and batch scripts. Both approaches require a high level of expertise to use them appropriately. That is true of all methods for landform delineation. The Virtual Watershed methodology uses "straightforward" topographic metrics, such as delineation of a watershed (see e.g., "drainage wings" in Appendix B, Figure B2), extraction of a channel network from a DTM, and delineation of floodplains using relative elevation, as described in Appendix B. The "users" in this case will initially be the USC project team as we continue working through the sequence of sub-projects. The workflows we develop will be available as R scripts, using our choice of methods (e.g., our choice of eCognition for OBIA may preclude inclusion of OBIA in open-source software). The goal is that future users can replicate what we do in subsequent sub-projects and modify the methods and choices as they determine appropriate. This landform-mapping sub-project is intended to determine if computer-based methods for automatic landform mapping can be used to delineate polygons identifying potential landslide sites based on RILs, not to provide RIL maps. At this point, decisions in model application are "ad hoc" because all we have to evaluate model performance are the variable opinions of human experts. The subsequent sub-projects will provide a quantitative set of methods for evaluating the criteria used to delineate RILs. We are working to make these as replicable as we can by using iterative review and documenting all workflows using software notebooks with executable code.	
Reviewer 2	**1.2**	Include a reference as an example of these more advanced computational methods.	Removed sentence.	
Reviewer 2	**1.2**	Incomplete source list. Additional references to consider: (Clubb et al., 2014; Passalacqua et al., 2010; Pelletier, 2013)	Added citations	
Reviewer 2	**1.2**	Refer the reader to appendix where DTM processing and derivative calculation is discussed	Directed to Appendix A.	
Reviewer 1	**1.2**	On these pages, there are two objectives listed. First, related to my general comment about flow of the Introduction, do these need to be stated twice? I understand they are in different subsections, but there is only a single page separation between them. This feels unnecessarily redundant. Also, the second bulleted objective does not read as an objective the same as the first bullet point. It reads as an expected project outcome. This could easily be resolved through some minor wordsmithing. Finally, in the short paragraph following the second listing of the objectives (on page 8), there is an incomplete sentence (begins with "One using image analysis…).	Removed objectives from this section and revised sentence beginning with "one using" for clarity.	
Reviewer 2	**1.2**	Missing references about using topography to identify landforms, see list at lines 478 – 479 above, but also (Clubb et al., 2017; Gailleton et al., 2019; Grieve et al., 2016) among others.	Added citations	

Reviewer 2	**1.2**	Would suggest instead or in addition: Dietrich et al., 2003; Grieve et al., 2016; Roering et al., 2010, among others.	Added Roering et al., which seems the most appropriate here	
Reviewer 2	**1.4**	Not clear how landslide inventories will work with lidar differencing. I'm concerned that the timescale would be too short to capture the empirical relationships between RILs and actual landsliding that the authors are seeking to make.	That may be true. That is a question examined in the next sub-project.	
Reviewer 1	**1.4**	Overall, I think this is one of the sections of the Introduction that needs to be looked at the most regarding improving clarity to aid the reader. The Introduction, thus far, has been telling the reader the goals and design of the project that will be discussed in the report. Now, there is a whole section that explains how the project is actually different than what was stated? After reading the full report, this section made more sense to me, but I think that it could be more seamlessly merged with previous sections and/or moved to fall before the other sections so that the reader doesn't have to go back and double check what changes correspond with what was discussed earlier in the Introduction. In addition, much of this section is devoted to talking about future work, but at the end of the *Introduction* . This threw me off, since as I approached section 2 on "Data," I expected a final description of *this* project and how data will inform its goals, not a shift in topic to the next sub-project. Consider saving this for later in the report, especially the Discussion.	Thank you for the suggestion. We moved an edited version of this section to the discussion. In the introduction, we now describe that we use manually mapped landforms rather than LHZ Mass Wasting Units to test the ability to parameterize the OBIA methods to observed landform locations and direct the reader to the discussion for a justification of why we did this.	
Reviewer 1	**1.4**	Second, in regard to clarity and overall study design: Why was this exercise not done in tandem with existing landslide databases? I understand that this is stated as the goal of the next sub-project, but many datasets already exist. Why not use them?	Existing landslide inventories pose several challenges: 1. Most inventories predate available lidar, so the topography associated with a landslide site, as determined using a lidar DTM, represents post-landslide conditions, not pre-landslide conditions. 2. Imprecise location. If digitized polygon or initiation-point locations are not well registered to the DTM, they may fall outside of the landform they are associated with. This is a primary reason we want to use inventories based on lidar differencing. The landslide location and associated topography are based on the same DTM data source. 3. Imprecise or unknown surface area extent of the landslide scar. 4. Unknown volume of mobilized material. 5. Many inventories do not include runout extent. These are issues that are described in the study design for the next sub-project. Ideally, this would have all been fully laid out in the original scoping document for the USC project (in 2017). We are learning as we go.	
Reviewer 1	**2.0**	This comment applies to all map figures. I think that if there are going to be mention of geologic structures (folds, faults) that may potentially control landslide distribution, it would be useful to include those structures on the maps.	Added geologic maps with structures to study area figures.	
Reviewer 2	**2.1**	Add references to subsequent figures in caption	Added reference to caption.	
Reviewer 2	**2.1**	Which of the study sites have repeat lidar and can be used for the next project?	All of them	
Reviewer 2	**2.1**	Would be helpful to include a geologic map of the sites rather than only a written description.	Added geologic maps with structures to study area figures.	
Reviewer 2	**2.1**	Order of maps is flipped from other maps (hillshade is on the top).	Fixed	
Reviewer 1	**2.1**	Give the name/description of the unit rather than only stating Qapwo1. It is not clear what that is.	Added description, units reference geologic map now.	
Reviewer 2	**2.1**	Suggest replacing current maps with a slopeshade map and a geologic map for each site instead of terrain with watershed boundaries (not that useful) and hillshade (biased representation of terrain).	We have provided watershed boundaries because our broader forest practices audience is very "watershed oriented" and would ask "Where is this?" if not presented this way (lots of field people who may have had boots on the ground in some of these 4 sites). Added geologic maps and replaced hillshade maps with slopeshade maps.	
Reviewer 2	**2.1**	How old? What does stable mean here?	Have rewritten to be more basically descriptive and then summarize the punch line (don't think this area is uplifting).	
Reviewer 2	**2.1**	Drainage density? Or depth of channel incision?	Sorry, channel density (fixed, and explained in text).	
Reviewer 1	**2.1**	More citations in the second paragraph in regards to surface geology and groundwater movement are needed. Also, what does it mean that the Willapa Hills are "old?" The rocks are old? There are relict landforms from different tectono-climatic regimes and/or baselevels? The erosion rate is slow?	Have rewritten to be more basically descriptive and then summarize the punch line (don't think this area is uplifting). We have added geologic references, but this idea that tectonic stability has led to "old" terrain with limited mechanical erosion is really based on the field experience of two of the co-authors and doesn't have a peer-reviewed reference.	
Reviewer 1	**2.1**	What is "significant erosion potential" mean, and how is that determined?	Well, slippery Darrington phyllite at high elevation on steep slopes rather obviously has significant erosion potential (but so too did Site 1 slopes without the phyllite and we didn't feel the need to make point). We have deleted this idea, lacking a specific way to justify the statement.	
Reviewer 2	**2.1**	Choice of sites. Why are there no sites in the Cascades? Is site 3 sufficient to represent recently deglaciated valleys like the North Fork of the Stillaguamish where the Oso landslide occurred?	We were testing model ability to delineate bedrock hollows and inner gorges in a (less than exhaustive) variety of Washington landscapes - active uplift, hard lithology, super steep in Site 1, apparently no uplift and deep weathering profiles/regolith at Site 2, recently and relatively uneroded glacial terrain at Site 3, soft lithology at Site 4. Another CMER study is working on deep-seated landslide terrain in the Northern Cascade; glaciated valleys and DSL's are not our focus in this project. That having been said, we will be leveraging some of the LCD performed for the other project in our next sub-project, so coverage of those valleys will occur. Given that the models work reasonably well in our 4 selected terrains, we will assume equivalent performance in deglaciated valleys until we discover otherwise.	
Reviewer 2	**2.1**	And also due to low erosion rates?	This paragraph has been rewritten to clarify observation from hypothesis. Low channel density and low erosion rates are probably both symptoms of no uplift (and no glaciation), and not really one causing the other but clearly related things. Anyway, while the original paragraph caused several comments, the observation that the Willapa Hills are really different from the rest of Western Washington seems important to acknowledge in our site selection.	
Reviewer 1	**2.3**	How was channel density measured (section 2.3)? From previously field-identified channels? Mapping channel heads on lidar? Using a threshold drainage area for channel initiation? Also, at this point, there has been no definition for the BH or IG acronyms. (Also, as a sidenote, DTM is never defined either).	Channel density was based on channel networks traced from the DTMs. A footnote was added with this information.	

Reviewer 2	**3.0**	For these terrain statistics, would be more useful to see distributions rather than bar charts of mean values. These should be easy-ish to derive from DTMs.	Channel density and zero-order-basin density are single values for each study area, so a bar chart displays all the available information about those values. The cumulative frequency distribution of surface gradient is shown in the subsequent figure (Figure 7). A frequency distribution of zero-order-basin sizes could be generated, but the bar chart of mean sizes illustrates the differences between sites in a simple way adequate for this discussion.	
Reviewer 2	**3.2**	Is it appropriate for a state agency to adopt a method used by the timber company the agency is meant to regulate?	We (broadly the stakeholder participants involved with the scientific effort of the Adaptive Management Program evaluating the effectiveness of forest practices) are investigating 1) if our current RIL are the correct landforms to protect across all the Washington landscape or if some modifications to the list of RIL or nuancing the definitions of the current RIL by region or lithology would improve our ability to limit landslide events; and 2) can we with our new tools (e.g., DEMs from lidar, OBIA capability) create better screens to locate these landforms. Comparing two models is useful, and both will be adjusted to better represent those sites with the detailed characteristics that lead to landslides during the next sub-project. How one or both might be used as a screen is currently unknown, but there is a strong committment to retain the use of field assessment in final harvest/road engineering in Washington State, so neither is likely to become a direct regulatory tool.	
Reviewer 2	**3.2**	A figure similar to figure 9 would be helpful to understand the details of the OBIA via image segmentation and classification	The OBIA techniques used in this study follow the standard image analysis process of segmentation followed by classification. Figure 8 in the main body of the paper describes the segmentation/classification steps used and shows what the resulting objects/classes look like on the landscape. The authors felt key advancements were in determining what mathmatical parameters could be used to extract RILs and the use of data tables in the appendix were a better way to communicate that.	
Reviewer 2	**3.2**	What is a landing?	Edited document	
Reviewer 2	**3.2**	"Peer-reviewed literature" should be accompanied by citations	Added citations in document	
Reviewer 2	**3.2**	More details in the main text would be useful, particularly because Appendix A reveals that the OBIA methods need to be highly tuned based on visual inspection and user decisions.	Due to the nature of our audience (CMER), which is less technically knowledgable on this subject and the intent to have this document be somewhat public facing, we left highly technical details in the Appendix to reduce confusion and to not overwhelm.	
Reviewer 2	**3.3**	Add Dietrich et al., 2003.	Added	
Reviewer 2	**3.3**	What is the difference between landscape position and drainage area?	I added a citation for landscape position - and changed the terminology to "topographic position". Both terms "landscape position" and "topographic position" are used in the literature. Landscape position refers to location of a feature relative to ridge tops and valley floors; drainage area is another term for contributing area.	
Reviewer 2	**3.3**	Really appreciate the flow chart, but I'm having a hard time following the steps. Put boxes around the individual steps and then connect with arrows to show workflow.	This is not indicating sequential steps or a workflow, but the concepts and data structures used to build a virtual watershed data base. Once step 1 is done, everything else can follow in any order depending on the questions addressed.	
Reviewer 2	**3.3**	After reading appendices, understand that "drainage wings" are part of the virtual watershed vocabulary, but this was an unfamiliar term for me. Please define.	Added brief description	
Reviewer 2	**3.3**	Here and elsewhere, other references for identifying channel heads: (Clubb et al., 2014; Lashermes et al., 2007; Passalacqua et al., 2010; Pelletier, 2013)	Added	
Reviewer 2	**3.3**	There are much more robust methods than using drainage area thresholds. See for example (Clubb et al., 2014)	That is true and we could better capitalize on information potentially available in a DEM. However, in this case, our goal was to match expert opinion about the downslope extent of bedrock hollows and the experts do not necessarily key in to all the information available in a DEM. We can work to provide more of that information, e.g., maps of tangential curvature and, say, deviation from mean elevation, both of which could provide other indicators of upslope channel extent (though lidar ground-return density on forested slopes may limit our ability to resolve channel-head location), but for this project, we were simply trying to show that the computer models can be adjusted to accommodate different interpretations of the information available. The VM model can employ multiple data sources to infer channel extent (e.g., slope-area product, curvature, openness, deviation from mean elevation), but for finding a location as consistent as feasible with the manually mapped hollow extents, use of a slope-area threshold alone was a simple way of adjusting downslope hollow extent. We can capitalize on additional topographic indices as we move forward with the subsequent projects using lidar differencing to better resolve topographic correlations with landslide location and runout extent. These are all site specific, rather than global metrics, but are also all limited by the size of features that can be resolved given the spatially variable ground-return density in heavily vegetated areas.	
Reviewer 2	**3.3**	Concerned about adjusting algorithm parameters to match manually drawn polygons when manually drawn polygons have so much disagreement between mappers.	We needed a method for evaluating the computer methods. In the context of this project, the results of any computer model used will be scrutinized by field practitioners, so the study design called for comparison of the computer-generated maps with maps drawn by experienced mappers. The intent was to see to what extent the models could replicate what the mappers interpreted from the DEMs. We anticipated some disagreement among human-drawn maps, but we did not anticipate the magnitude of that disagreement, which is why we devised the voting exercise. Again, the experts vote based on their opinions of what RIL polygons should look like and those opinions vary. We could set model parameters based solely on RIL definitions, but another conclusion of this project is that those definitions may lack sufficient detail to ensure that expert interpretations of those definitions will produce maps that match well. That is a useful finding. As we collect precise data on landslide locations in the subsequent sub-project, we will work to determine how RIL criteria could be expanded and refined to be consistent with the empirical evidence and to provide consistent interpretations across practitioners.	
Reviewer 1	**3.3**	The second paragraph here (begins with "using thresholds…") is largely redundant, as much of this information was just stated.	None of the information about "feathering" was stated. This paragraph describes a strategy for ensuring that each zero-order basin has only one DTM-traced channel extending into it.	
Reviewer 1	**3.3**	Is this figure created by the authors? Or pulled from a different study (I ask because different components appear to have different resolutions). If pulled from a different study, it needs to be cited. Also, this figure feels largely out of place in the current main text, at least without some more of the details that are currently found in the Appendix.	The source of the figure is now cited in the caption.	
Reviewer 3	**4.0**	The report needs more discussion about the implications of what "Reasonable Screening" and "Correctly Mapped" means in terms of utility of these computer-based methods.	These implications are discussed in Section 4.2.3. We do not think we can use these methods and results to evaluate the utility of these methods for identifying landslide-prone terrain until we compare them directly with landslide locations and the spatial density of those landslides. That is a task for the next sub-project.	
Reviewer 2	**4.1**	Box and whisker plots are standard statistical plots; don't think there needs to be a thorough definition in the caption.	We feel the definition is helpful due to the intended audience of this report.	

Reviewer 2	**4.1**	Title of bottom graph needs superscript on the 2	Fixed	
Reviewer 2	**4.1**	Are these the mean values of all pairwise comparisons?	Yes, now stated in text	
Reviewer 2	**4.1**	I found these figures hard to interpret and was not sure what I was supposed to be getting out of them.	Added some text to aid with interpretation.	
Reviewer 2	**4.1**	Precedes the discussion of tangential curvature in the text. Consider reordering.	Fixed	
Reviewer 3	**4.1**	Figure 13: There would be benefit to showing several examples of this (good, bad, ugly) to understand the variability – it appears some of this owes to the level of detail that mappers considered. This is discussed briefly. But maybe one way to think of this is "what are minimum map scales in which accurate delineation of polygons are relevant to practice"?	I agree. In Figure 12 we show the mapper overlap, and chose this area to get a closer look at the difference in delineation. This example is representative of most of the bedrock hollow areas, which is why we showed it. We show the "ugly" in the outlier polygons shown in Figure 32.	
Reviewer 2	**4.1**	Missing scale bar and lat/long, would be nice to have inset location map. Is there overlap between OBIA and VW? If so how is it represented in figure?	Added scale bar, etc. Location map for all sites is in Section 2.1. These sites were chosen specifically without overlap, so there is no overlap in the figure.	
Reviewer 2	**4.1**	Rotate to match other figures	Rotated	
Reviewer 2	**4.1**	Show extent in figure 12	Added square to Figure 12 and mentioned in caption of Figure 13.	
Reviewer 2	**4.1**	Did practitioners know which model was which? I wonder if the VW model is more appealing to practitioners because it is more similar to watershed processes that they are used to considering.	No, the practitioners did not know which model they were voting on. We provided the shapefiles and labeled them "Model 1" or "Model 2"	
Reviewer 1	**4.1**	It was unclear to me why the mappers in the first exercise were using different maps. Why was this done? Does it make results comparable? I recognize that that difference is what motivated the exercise focused on in section 4.1.2.3, but given the shift, is it necessary to report the first exercise with inconsistent mapping tools? (also check spelling in the title of section 4.1.2.3, which currently says "C0mparison")	After we completed the first exercise, there was a lot of discussion among the project team about the area (looking at individual landforms and also the area as a whole) so, in order to prevent a bias going into the next exercise, we chose an area nearby but not in the same spot. Otherwise, it would be hard for a mapper to not take what was discussed into consideration when voting. Added a brief sentence about this in the beginning of section 4.1.2.3.	
Reviewer 1	**4.1**	I would recommend avoiding subjective language related to the exercise being "lengthy" and "mind-numbing." Certainly, from experience, I agree. But I think there are other ways to motivate automated approaches.	Agreed, revised in the document.	
Reviewer 1	**4.1**	Given the mappers occasionally agreed that a model-identified "outlier" polygon was an RIL, I think it is important to restate here if the mappers were able to discuss (and potentially be subject to convincing) whether or not the outliers were indeed RILs or not. This is important for considering how biased the model testing was.	They voted individually without discussing with each other, I've added some clarifying text to the document.	
Reviewer 1	**4.1**	The information in the "Comparison with…" section seems to be coming much too late. This feels like information that should be in the methods section. Maybe I'm missing something?	Agreed, moved into "methods" section (4.1.1)	
Reviewer 2	**4.1**	Interesting to note that mappers are much more likely switch to RIL from nonRIL than vice versa.	It is interesting!	
Reviewer 2	**4.1**	Keep consistent colors for each mapper and for VW and OBIA polygons and colors on plots (bar graphs, histograms, CDFs that come in appendices).	Done	
Reviewer 2	**4.1**	Suggest reversing order of these panels because of order that OBIA and VW are usually discussed. Looking at the errant polygons in VW, appears that ones with 0 or 1 yes votes are in locations where the channel network is likely too dense, suggesting that the model is poorly tuned. Would be interested to see if this is true of the rest of the errant polygons from this algorithm.	It may be true that the channel density is too high. We did extend traced channel flow paths upslope of direct topographic indicators of channel presence because we wanted the traced flow paths to extend into zero-order basins. The voting team did not explicitly include presence of a topographically identifyable channel in determining if a mapped hollow polygon met the RIL criteria; they were focused more directly on the steepness and convergence of the terrain, but presence or absence of a downslope channelized debris-flow corridor could have influenced their voting decisions. The channel-initiation criteria used with a model can be modified to adjust the delineated channel density. We can use well-targeted field visits to verify model estimates of channel extent. However, channel extent is notoriously variable from site to site, and in debris-flow terrain, channel extent can change over time as debris flows scour headwater channels and they then slowly refill over time. (Figures now flipped)	
Reviewer 3	**4.2**	For the OBIA or VW results that mappers did not vote for, were there any specific trends or attributes that resulted in them being a negative vote? Should this be considered?	Yes! These attributes are described in section 5.2 "automated mapping limitations"	
Reviewer 3	**4.2**	Phase 2: Understandably, OBIA and VW yield different outputs in terms of bedrock hollows and inner gorges (both total number and location). The yes votes for reasonable screening and correctly drawn conditions varies, but often appear to be somewhat similar between methods. Would a comparison of OBIA polygons that were voted "yes" for RS and/or CD, but not mapped by VW be worth looking at? What about vice versa? For cases where one model performed better versus one another? This might be a means for discerning the strengths and weaknesses of each model. Similarly, what about polygons that both methods produce and are both voted "yes"? How similar were the polygons?	Yes! In Phase 1 we did side-by-side comparisons like this visually with the project team. We looked at them in a way similar to the "outliers" exercise but as a group instead of individual voting. These visual inspections provided insights for how to interpret the frequency distributions of topographic indices (shown in Appendix F) in the context of model performance. We did not carry this analaysis any further in terms of model adjustments because we felt it more time efficient to wait until we have collected the landslide inventories with the next sub-project to compare with modeled landforms.	
Reviewer 3	**4.2**	Map figures in Phase 2 of polygons with number of votes: Apologies if I missed this, but which votes are shown? Just for reasonable screen? Or for correctly drawn? It would be worth showing a few examples where votes were positive or negative.	They represent "reasonable screen" yes-votes and I have revised the text to make that more clear. Figure 37 shows the distribution of yes and no votes accross Site 1A.	
Reviewer 2	**4.2**	As above, more care should be taken with color choice in the figures.	These colors are representing reasonable screen or correctly drawn votes and do not correspond to a model or mapper, so there would not be a specific color to use here that is used in other figures.	
Reviewer 2	**4.2**	Not clear whether these plots are of CD/total polygons or CD/RS	The last line of the caption explains this. "For CD, the percentages shown represent the proportion of yes-vote RS polygons that received CD yes-votes"	
Reviewer 2	**4.2**	Final column seems unnecessary / redundant – only get a CD vote if there's a yes in the RS prompt. True of other columns also	Completly agree, removed last column in all tables in this section.	
Reviewer 2	**4.2**	Switch panel order. Is this RS or CD?	RS (added to caption), switched panel order.	
Reviewer 2	**4.2**	Maybe just add a final row of this table that is averages and then get rid of the weird bar chart graphs (e.g., Figure 40)	Done	

Reviewer 2	**4.2**	Does incision along main channel not get flagged as inner gorge? Also, I realize these are small features, but they're very hard to see on the figures, so it's hard know what message I should be taking from the figure.	The models as currently implemented do not attempt to identify incised channels. The delineation of an inner gorge is based on the presence of steep channel-adjacent slopes with an upslope break in slope. If channels are incised to an extent resolvable with a 1-m DEM interpolated from sufficiently dense lidar ground returns, we could seek to quantify the degree of incision. For now, if there are ground returns from the base of an incised channel, these will influence the interpolated DEM elevations and be resolved as a steeper, channel-adjacent slope. We tried our best with the symbology to make the inner-gorges stand out and hoped that the vote count next to the colors on the legend would help supplement the information.	
Reviewer 1	**4.2**	I don't believe DSL was ever defined (I'm assuming deep-seated landslide?). Also, for the second equation, how does "our smaller sample size (N)" differ from "our final sample size (n)" given what has been discussed so far? I don't recall there being any sort of trimming or secondary dataset, but maybe this is referring to something else?	Defined DSL acronym. Added clarifying text in Section 4.2.1 for population.	
Reviewer 1	**4.2**	Related to point 3, how would the EPs know if the model had considered deliverability? Would a more accurate statement here be whether the EPs agreed with model output given *their* interpretation of sediment deliverability. Furthermore, there hasn't been discussion of deliverability up to this point. How is it identified and/or interpreted in this context?	The basic question asked of the EPs was poorly worded; conversation among the EPs led to clarity of what was really being asked, and edits to this bullet point have been made to represent this understanding. Results of this deliverability question, following this enhanced clarity of what the prompt was getting at, are summarized in the 2nd paragraph of Section 4.2.3.	
Reviewer 1	**5.0**	Similarly, the next paragraph mentions how "the divergence between manual and automated mapping underscores the importance of refining model parameters and validating them against independent criteria." What is the independent criteria in this case? I don't think it would be the EPs, since they are simply a separate method to which the automated polygons are compared. I think it is referring to landslide maps? Again, I know that comparison to landslide compilations is going to be saved for the next sub-project, but I'm not sure this sentence can be stated yet, because there hasn't *yet* been a comparison to "independent criteria" (i.e., "geomorphic processes," as stated at start of page 90).	Revised and reorganized sections 5 and 6 to address this and other comments.	
Reviewer 3	**5.0**	How generalizable can these mapping tools be for the varied geology observed over the State? While there are criteria to define these RILs and these exercises are intended to provide consistent means of mapping these features, should these outputs be generalized? Should they be calibrated to different environments/geologies? Should there be room for expert judgment?	We do not know how generalizable these mapping tools will be for the varied geology observed over the state. Topography reflects geology, so it seems reasonable to expect that these tools are generalizable, and that they will require regional calibration. We cannot eliminate expert judgement. One goal is to provide tools and data with which expert judgement can be augmented and tested.	
Reviewer 1	**5.0**	This discussion at the bottom of the second-to-last paragraph (beginning with "In phase 2…") is useful, but highlights my general comment about needing more information from the Appendix in the main text. If there is going to be discussion about how the project "…highlighted specific model limitations related to topographic convergence, slope thresholds, and stream adjacency," then there need to be results presented in the main text that highlight that (i.e., figures that show how adjusting these parameters impact modeled polygons).	The limitations were highlighted by the practitioners review of the polygons, not the adjustment of the models. These parameters were identified by the practitioners as reasons for their "no" votes in Phase 2 of the project. Added clarifying text to the document.	
Reviewer 1	**5.0**	Finally, overall, the Discussion section (5.0) is *very short* compared to the volume of information in the other sections, and is even shorter than the future work section. I think there is a lot more information that can be included in this section. Maybe some material needs to be shifted from the results sections to the Discussion? Also, machine learning is mentioned in the very last paragraph of the Conclusions, but as far as I saw, nowhere else. The Discussion would be a good place to dive into this, if the authors think that it is necessary to mention it in the Conclusions.	Thank you for the comment, we moved sections of the report to the discussion, shortened other part of the report that were redundant and added a brief description of the potential application of machine learning techniques for the next project steps in the Discussion.	
Reviewer 2	**6.0**	The project that follows this one, where RILs are tied to actual landsliding, proposes to develop a landslide inventory using repeat lidar. First, it is not clear from this report whether those data already exist or whether new lidar will need to be flown. Second, and more importantly, landsliding is highly stochastic. In a given storm, only a fraction of bedrock hollows will fail, so using a short-term landslide inventory to derive empirical relationships between the topographic characteristics of RILs and landslide occurrence seems highly problematic. Imagine a mapped bedrock hollow, from either an expert practitioner or one of the algorithms. The next sub-project proposed to test whether a landslide issues from this bedrock hollow over the interval of repeat lidar and then use the pre-failure topography of failed hollows to constrain the tuning of the algorithms presented in this document. If the hollow fails, then the topography of that hollow (the average slope, the tangential curvature, etc.) would be taken as predictive of landsliding; if the hollow does not fail, then the topography of this hollow would be taken as predictive of no landsliding. However, there are myriad reasons that a hollow whose shape favors failure might not fail (e.g., Benda & Dunne, 1997), invalidating the proposed linkage between failed hollows and pre-failure topography. To help with this issue, I suggest leveraging other mapping approaches, such as aerial imagery, that could extend the timespan of the landslide inventory, though I am not sure this will entirely overcome the problem.	We did clarify that the landform mapping project is one in a sequence of projects and that scoping of that sequence did undergo UPSAG, CMER, and ISPR review. The study design for the next sub-project has also undergone UPSAG, CMER, and ISPR approval. We included discussion of the next project in this report in response to UPSAG reviews to provide greater context for how the automated landform mapping methods will be used. Subsequent projects will rely on lidar already collected and available on the Washington Lidar Portal. Methods to map susceptibility to shallow landslide initiation and runout recognize the stochastic nature of landsliding. The goal is to identify terrain characteristics associated with past landslide occurrences, with the assumption that future landslide initiations and runout extents will occur in similar locations under similar circumstances. We will measure susceptibility in terms of the proportion of observed landslides associated with specific terrain characteristics. The characteristics we will examine include topographic attributes, substrate, land cover, and estimates of antecedent moisture, storm magnitude, and storm intensity. We can then look at how topographic attributes associated with high proportions of observed shallow landslides relate to the RIL criteria. We do not know yet how much we can infer from available data; that is something to be learned as we work through the next sub-project. We recognize that we are dealing with a "presence only" case in classifying terrain as landslide prone based on observed landslide locations and will apply appropriate sampling and modeling techniques to deal with this limitation. The problems with landslide mapping from aerial imagery is lack of precision in landslide-boundary placement, the generally unknown pre-landslide topography for photography that predates lidar aquisitions, and unknown bias in the proportion and size of landslides that are not visible under canopy. These issues are what motivated our use of lidar differencing for this study.	

Reviewer 3	**6.0**	The quality of lidar varies quite a bit, especially in steep, forested terrain (e.g. low point densities, etc.). Perhaps a philosophical question for next steps of this work, if these models demonstrate sensitivity to these DTMs and forcing events that trigger landslides may vary, should these tools be used for screening to supplement boots-on-the-ground expert judgment?	Outputs of the next steps in this study include measures of confidence in model results. Lidar DTMs offer far more information about site conditions than we had available when using the 1:24,000 base maps with which much of the original work was done that led to the current RIL criteria. Now we can look at multiscale topographic indices, we can estimate contributing areas for specific storms, we have climatic data, landcover data, and precipitation data that we did not have 20 years ago. We have a huge array of data for statistical analyses that did not exist then. The final results will be used as remote screens, just as our original hand-mapping efforts on 1:24,000 USGS topo sheets have been used as screens over the past 2-3 decades. We expect new products to have fewer false positives (landforms that don't meet RIL definition) and fewer false negatives (unmapped RILs) that the old products, which will help focus field assessment, but there is agreement among practitioners that no remote product should replace field assessment.	
Reviewer 1	**6.2**	"We have demonstrated that the computer models can be adapted by adjusting the input parameters…" Again, as far as I saw from reading the main text, that hasn't been shown. If that is going to be a major takeaway from this sub-project, it needs to be addressed explicitly in the main text. (This comment applies in other places in later pages (e.g., second paragraph of section 6.3), but I won't mention it again in those spots).	Threshold values in the VW model used to determine the extent of bedrock-hollow and inner-gorge polygons were adjusted to better match the frequency distributions of gradient, gradient ratio (profile curvature), tangential curvature, and the proportion of a polygon with gradients less than 70%. We added text to state this (Section 4.1.2).	
Reviewer 1	**6.4**	What is the Post-Mortem study area?	Revised in text	
Reviewer 1	**6.5**	LCD was never defined	Defined in text	
AE	**3.0 and 4.0**	The stated goal of the project is to "develop automated methods to identify bedrock hollows and inner gorges;" however, there is a lack of discussion of automated algorithms in the main text and an over-sized emphasis on the results and methods of the non-automated human-mappers, which have significant uncertainty and disagreement. Additionally, it is concerning that manual delineations, which have significant inconsistencies, are prioritized over process-based algorithim delineations, which would truly be reproducible for a given set of parameters for a given digital terrain model.	1. *Lack of discussion of automated algorithms in the main text.* Our primary audience consists of UPSAG and CMER members so emphasis in the document is directed to their interests and to their review comments and reactions to previous drafts of the report. That is why description of the algorithms were moved to appendices. 2. *Over-sized emphasis on the results and methods of the non-automated human-mappers* . The detailed descriptions of the degree of agreement among mappers also reflect the make-up of our primary audiences: UPSAG and CMER. These are primarily field-oriented people and many are familiar with manual-mapping methods. They predominately expected to evaluate the performance of automated methods in the context of how well they could match expert-drawn maps, which was our plan too as we started this project. It is important, therefore, to document why this approach for model evaluation ultimately proved infeasible. 3. *manual delineations ... are prioritized over process-based algorithm delineations* . This again is a response to our primary audience. We need to show a group, many of whom tend to trust expert opinion and distrust computer models, that the computer models can perform as well as the experts. When models were parameterized solely on current RIL criteria, which are derived using a combination of empirical and process-based values, the resulting polygons did not meet expert approval. This might indicate that our process-based assumptions are incomplete, that the data available for using those process-based approaches are insufficient, and/or that expert opinion needs to change. We intend to better resolve this issue with the empirical and process-based approaches applied by the next sub-projects.	
AE	**3.2 and 3.3**	More information is needed about why the authors chose to use the Virtual Watershed and NetMap tools rather than utilizing GIS-based approaches, which are the standard approach for delineating watersheds and geomorphic features.	NetMap is a GIS-based appraoch. It is an ArcGIS add in, although the methods and underlying software used have also been implemented in R and python spatial analyses packages. Text was added to Section 1.2 to clarify. Note both the scoping document and the study design underwent UPSAG, CMER, and ISPR approval.	
AE	**3.2 and 3.3**	The main text lacks critical information on the technical details of the OBIA and VW methods and how the methods work. The text also needs information on model parameter selection and tuning. Which parameters were tuned and by how much? Are model results more sensitive to certain parameters? Some text regarding these questions is available in the Appendices, but it needs to be in the main text methods section to improve reader understanding of the models.	Please refer to responses in lines 37, 82 and #1 of line 86, 88	
Reviewer 2	**3.2 and 3.3**	The assumptions in the algorithms were buried in the appendices. It also wasn't clear to me why the given algorithms were chosen, given the number of options for topographic analysis (e.g., Schwanghart & Scherler, 2014).	The rationale for choice of models is now described in section 1.2. The choice of algorithms was based on experience of the modelers. For example, an algorithm for calculating gradient was devised to better match the method used by a field practitioner - i.e., to measure gradient over a specified up- or down-slope length. We are not aware of this specific algorithm having been implemented anywhere else. The primary readers of this report are not interested in the algorithms used, only in the performance of automated methods as compared to expert manual mappers. We did not, therefore, spend a lot of text in the main report describing specific algorithms.	
Reviewer 1	**3.2 and 3.3**	This is the key location to provide more detailed information on the OBIA and VW (i.e., moving some materials from the appendix). Specifically, there's statements on this page about how "our model is designed to delineate and classify landforms based on mathematical derivations of key morphometric parameters (e.g., surface elevation, gradient, curvature […])_ that can be derived from DTMs using image analysis tools." That is all great, but while reading, I wanted to know *how* the models do that. While I understand some of these algorithms and associated parameters can be complicated and dense, there needs to be some more detailed information here on how the models actually work. For instance, *how* does the WY OBIA model do on the things listed in points 1-4 on this page and the next? And, in point 4, *how* are RIL definitions and criteria taken into account by the model when classifying landforms? Again, this can be accomplished by transferring information from the Appendix to the main text.	Please refer to responses in lines 37, 82 and #1 of line 86, 88	

Reviewer 1	**3.2 and 3.3**	I think there was some attempt to explain the model by analogy. In other words, describing model inputs and outputs the way a human brain interprets topography to map RILs. While an interesting way to think about it, this doesn't provide the necessary details for a reader to understand or *reproduce* the analysis done here. In other words, if this subsection is about the models, there needs to be less focus on "experts," and more focus on the algorithms. The main reason I focus so much on this particular issue is that it makes the main text of the report non-reproducible. While the appendices are important, I think that the main text should be able to largely stand alone.	Most journal papers now provide algorithmic details in supplimentary material separate from the main text. We have done the same here, primarily in consideration of our primary readers for whom such details would be a distraction from what they want to know. The appendices provide that material for those readers who want to access it. The rule sets and parameter values used for the two models are provided in the appendices. Please refer to responses in lines 37, 82 and #1 of line 86, 88	
Reviewer 1	**3.2 and 3.3**	There is very useful information in the Appendices, and there are methods that are included there that I think should be in the main text. Ultimately, the goal of this sub-project is to test, describe, and highlight the two automated approaches for object-based landform mapping: Object-Based Image Analysis for the sake of the reader who may be reading this document with an expectation of being presented information about automated RIL mapping, they shouldn't be necessarily required to dig mainly through the appendix to get any sort of detailed inform ation on it. (OBIA) and Virtual Watershed (VW) terrain analysis (this is in the title of the sub-project after all!). However, the majority of the methods section of the report goes into detailing how human mappers identified RILs and compared their mapping to each other, with a smaller portion dedicated to how human mappers evaluated automated maps. While that is all useful, there absolutely needs to be more information in the main text that describes *how* the OBIA and VW methods actually work and how parameters are varied in order to produce accurate maps that can evaluated and/or compared to human-produced maps. How much parameter tuning was there? (e.g., page 88-89 and language related to "targeted improvements to input parameterization"). Were the models more sensitive to certain parameters? Currently, most of that information is in the Appendices, with only generic statements about landscape properties (gradient, tangential curvature, drainage area) that go into the models. So, while I am glad that information is available, for the sake of the reader who may be reading this document with an expectation of being presented information about automated RIL mapping, they shouldn't be necessarily required to dig mainly through the appendix to get any sort of detailed inform ation on it.	Please refer to responses in lines 37, 82 and #1 of line 86, 88 Refer to section 5.4 for text about model sensitivity and use moving forward.	
Reviewer 2	**4.0 and 5.0**	The primary uncertainty and limitation that I see in this work is the calibration of the algorithms based on human-mapped data. Most of the space in the paper is dedicated to describing the ways in which the human mappers disagree on the extent and location of RILs. Hence, I was surprised to read in the Appendices that the algorithms are tuned to maximize matching with the human mappers, sometimes to the extent that assumptions rooted in geomorphic process (e.g., that inner gorges are bounded by a break in slope) are relaxed. This uncertainty is underemphasized in the document.	This project is not intended to produce final landform maps, but to evaluate computer-based methods for producing such maps. We needed to show our other reviewers, many of whom actually map and identify RILs both remotely and on the ground, that these models can accommodate their ideas and concerns. The only "data" we are using now to evaluate these models are the opinions of experts, so we wanted to show why it might be challenging to meet expectations across a broad range of practioners.	
Reviewer 2	**4.0 and Appendices**	The tuning and parameterization of the algorithms, as described in the appendices, sometimes prioritize matching human mapping results over adhering to process-based definitions of inner gorges and bedrock hollows. For example, it doesn't make sense to extend an inner gorge beyond a slope break just to make it match with an expert practitioner polygon. Particularly given how inconsistent the results between mappers seem to be, changing algorithm parameters to match human mapping seemed misguided.	Process-based definitions are subject to change as we learn more about those processes. Field practitioners might have insights that point to improvements in the criteria used to identify potentially unstable landforms and in our interpration of remotely sensed data (the DTMs). At this point in the sequence of USC projects, we only wanted to demonstrate that computer models can be used to generate landform maps based on a set of landform definitions that may change. In the next sub-projects, we will use precisely measured landslide locations and sizes to evaluate the criteria used to define landform locations and boundaries to see if we can refine those criteria to improve both computer-model and human-practitioner mapping of landslide-prone landforms	
Reviewer 3	**4.1, Table 3**	How were True Negatives determined? By cells that didn't have any mapped RIL?	A "true negative" in this context is a location where neither the manual map or the computer map had a polygon. All areas, both inside and outside of polygons, were measured using a projection onto a horizontal plane (not slope area). In terms of overlap with a DTM cell, a cell containing neither a manual or computer polygon would be included in the true-negative area.	
Reviewer 3	**6.3 and 6.4**	Section 6.3 and 6.4: While mapping landforms is valuable and lidar differencing might serve as a means of supporting calibration of this mapping, it may be possible that storm intensity (which is not consistent between events or often catchments) may affect this density and the locations of landslides. How can this be accounted for in a sensitivity study of this nature? Can the uncertainty in lidar differencing affect such results? What about landslide kinematics (e.g. translation, rotation that results in subsidence and heave at the head and toe, respectively, complete evacuation and entrainment of debris during a flow)?	The study design for susceptibility and runout (next sub-project) specifies examination of how landslide density (measured in terms of landslide number, area, and volume per unit area) varies with storm attributes. These attributes will be estimated using gridded precipitation data, available from a variety of sources. We do not know yet how feasible it is to use these data to correlate landslide density with storm characteristics, but we recognize the confounding influence of spatially varying storm attributes and antecedent moisture on landslide density and seek to quantify it. The USC study is focused on shallow landslides, so we are looking at locations where there was complete evacuation and runout. We are using lidar differencing to identify those locations and to measure the surface area, volume, and runout extent of features identified as potential shallow landslides.	
Reviewer 1	**All**	All axes need to be appropriately labeled (e.g., on Figures 29 and 30, the y-axes aren't labeled. I understand the quantity being measured is in the title for each sub-plot, but I would recommend a y-axis label as well).	The y axis label is indicated in the plot title and caption, instead, due to limitation in plotting software	
Reviewer 1	**All**	Stream networks should be included on the maps, especially if the models use any sort of drainage area or distance to stream metric as a part of the algorithm.	At the scale needed to place maps on an 8.5-by-11-inch page, it is difficult to see the features already plotted. We were trying to keep the maps as uncluttered as possible. Added stream network layer to figures in Section 2.1.	
AE	**All**	Sentences containing phrases like "We know" and "described in peer-reviewed literature" require citations or must be rephrased.	Edited	
Reviewer 1	**All**	Slope maps would be useful in addition to the hillshades.	Hillshade figures have been replaced with slopeshade	

Reviewer 1	**All**	The authors should include a figure that shows full map of automated RILs. Specifically, it would be helpful to move figures C1 and E1 (or similar ones) to the main text. This is because the only maps in the main text of the automated RILs are outlier RILs that are compared to manual maps and/or evaluated by EPs. But, ultimately, I think it is important for the reader to have a clearer idea of what the original automated maps with *all* mapped RILs look like.	Figures C1-C8 show all RIL polygons produced by the models, and we've added text to the document to directly point the reader to that spot if they want a full look at the model results.	
AE	**All**	There are numerous acronyms throughout that readers must keep track of. It would be beneficial to provide a list of acronyms that readers can quickly refer to. Also, the authors need to be sure that all acronyms are defined (e.g., DTM is never defined in the text), and some acronyms should just be spelled out (e.g., "reasonable screens" and "correctly drawn" should not be acronyms)	Included an acronym list in the documnet	
AE	**All**	Please check for consistent use of lidar vs LiDAR throughout the report. Either use is fine as long as its use is consistent.	Edited to ensure consistent use of Lidar	
Reviewer 1	**All**	Labels need to be larger, with names of the sites (beyond simply Site 1, 2, 3, etc.) included.	Added site names to all captions with figures with site number. Revised figures.	
Reviewer 1	**All**	Overall, the report is written with clear language. There are locations, however, where there is significant redundancy and/or where it isn't clear whether methods, results, future work, etc. is being discussed. In general, I'd suggest that the authors re-proofread the report with a specific eye for conciseness and clarity. The report is very long, and certainly there is a lot of relevant material (and indeed, as I point out, I think some material actually needs to be shifted from the Appendices to the main text). But I think there is also a lot of room for trimming	We made a significant attempt to trim language where possible and reorganized for clarity around current and future work. Due to the nature of our program, some detailed descriptions need to remain that we recognize would not normally be part of a peer-reviewed document (e.g., the substantive description of Phase 1). Please refer to responses in lines 37, 82 and #1 of line 86, 88	
Reviewer 2	**Appendix A**	I was struck throughout this appendix by how many of the parameters are user-defined. Reading table A1, several of these definitions are relative or up to interpretation.	Landform definitions are often qualitative and inherently vague, as we see in the Board Manual definitions. This makes formalizing a logical model difficult. Various researchers have used the most noticable or strking feature, as perceived by a user, to help a computer recognize the landform (Gilbert, 2017). With the OBIA method , we use, precisification, a logical process of making a vague concept, landforms, clearer, by defining distinct boundaries where none existed before, for example the outer extents of a bedrock hollow. To be implementable in a model the defining characteristic of a landform and the surrounding region need to be expressed by geometries that can be extracted from a terrain model (Guilbert, 2017). The defining RIL characteristics used in the OBIA model were drawn out of Board Manual descriptions and rephrased where appropriate to reflect model terminology (Table A1). Iteration through segmentation and classification routines allows us to add specificity to vague landform criteria by describing the mathematical process used to determine, for instance, where the uppermost extent of an inner gorge is located, i.e., break in slope. Building thresholds into the model allows us to describe the mathematical process quantitatively. It is anticipated that the thresholds could vary based on context and/or formative geomorphic process that exist in different landscapes.	
Reviewer 2	**Appendix A**	Many practitioners use smoothing algorithms. See for example Roering et al., 2010.	Edited text.	
Reviewer 2	**Appendix A**	Would like to see more thorough discussion of options for calculating curvature. See for example Hurst et al., 2012 and Struble et al., 2014.	Made additions to document	
Reviewer 2	**Appendix A**	What is distance to neighbor?	Added clarification to document	
Reviewer 2	**Appendix A**	How does this threshold vary between landscapes?	The threshold used to classify pixel values that differ from their neighbors is less likely to be influenced by landscape dynamics than the window size used to calculate the metric . When applied to a single pixel, the mean difference to neighbor feature functions like an edge detection filter. We found that the window size (search radius) was more sensitive in unconfined systems with less distinct break in slope along the channels.	
Reviewer 2	**Appendix A**	Is this visual inspection done for each study site?	Text modified slightly but yes, selection of segmentation scales were based on the scale at which RILs were occuring in all study sites.	
Reviewer 2	**Appendix A**	A sensitivity analysis to the user inputs would be appreciated.	I agree. This work would benefit from a structured sensitivity analysis to determine the key parameters and highlight how much the model changes when user defined inputs vary. Will address this in the next sub-project.	
Reviewer 2	**Appendix A**	Greater in what? Length?	Made edits in document	
Reviewer 2	**Appendix A**	What is FAC? What is the "relative border"? In Find enclosed by class row, should read "absorb any object smaller than 1000 pixels into *bedrock hollow* class" not inner gorge class	Made edits in document and changed the Table	
Reviewer 2	**Appendix A**	Units on accumulation area? What does "relatively low border to a stream object" mean?	Made edits in the document.	
Reviewer 2	**Appendix A**	Avoid using abbreviations in tables if definition comes after table	Edited	
Reviewer 3	**Appendix B**	"We set that radius to what we estimated as the lower cross-slope width limit for a hollow landform, which may vary with terrain (Table B1, Appendix B1)." Considering this, can this be generalized? Can "feathering" be limited by truncating channels shorter than a threshold?	Modified text to better explain the methodology. Basically, channel heads are identified from a DTM using a variety of methods, but we want to extend our traced "channels" upslope of the channel head into the zero-order basin (hollow). But we want only a single traced channel to extend into each hollow. This text described how we prevent the traced channel from branching into more than one channel as it is extended upslope into the unchannelized portion of a hillslope.	
Reviewer 2	**Appendix B**	Suggest labeling road on figure.	Added road label to figure.	
Reviewer 2	**Appendix B**	Why is this not measured along flowpath?	We did not use flow-path length because that is not the way that a human practitioner will generally evaluate proximity of the valley wall to a channel, and because DTM-traced flow paths on valley floors may run parallel to the channel for some distance.	
Reviewer 2	**Appendix B**	Show how this map translates to inner gorge classification visually.	Added lines to the figure to indicate the zones associated with blue-line channels based on Euclidean distance and additional explanation in the text.	
Reviewer 2	**Appendix B**	Here and elsewhere, I understand that you're trying to get a better match with the human mapping, but there's no process-based reason the inner gorge should extend beyond the break in slope, or the bedrock hollow should extend beyond the convergent area. Isn't it possible that the human mappers are wrong and are drawing their polygons too large?	We do not know that there is no process-based reason the inner gorge should extend beyond the (DTM-inferred) break in slope or that the bedrock hollow should not extend beyond the (DTM-inferred) convergent area. At this point in the sequence of projects, our landform maps rely on human interpretation of the information available from a DTM. These practitioners have extensive field experience mapping these landforms and seeking ground-based relationships between topography and other indicators of landslide potential (e.g., tension cracks) that are not resolvable from a DTM. It is possible that the human mappers are wrong. It is also possible that our current understanding of mass wasing processes in these locations is incomplete. We will be able to evaluate the topographic criteria for mapping the extent of a bedrock hollow and inner gorge based on remotely sensed data more completely when we have precisely located landslide locations based on lidar differencing.	

Reviewer 2	**Appendix B**	Given the discrepancies between human mappers, why would you use these frequency distributions to tune the algorithms?	We used these distributions to set threshold values for certain parameters in the models to show that the models can be "tuned" to better match observed patterns. When we collect landslide inventories and can infer precise relationships between landslide density and terrain attributes, the models can then be tuned to better match the frequency distributions of those terrain attributes.	
Reviewer 2	**Appendix B**	Unclear if this is repeating or adding to the written information above. Need a citation for your approach using area times gradient squared. Is there a smoothing scale used that is different than 15 m for slope?	The table was added following a suggestion from UPSAG; it mostly reiterates information presented in the text. The citation for the slope-area function was provided in the text; it is now added in the table.	
Reviewer 1	**Appendix B**	How does the D8-LTD algorithm correct for the bias introduced by the limited 8-connected neighborhood?	I added a very brief explanation, refer to citations for greater detail.	
Reviewer 2	**Appendix C**	Caption: Reverse order of sentences	Done	
Reviewer 2	**Appendix C**	Revise to refer to precision, this is copy-pasted from previous caption.	Revised	
Reviewer 2	**Appendix C**	If this is the case, why include accuracy at all in the report?	We include accuracy because it is a commonly used metric of overlap expected by many reviewers and wanted to show how imbalance in sample proportions hinder interpretation of accuracy.	
Reviewer 2	**Appendix C**	pixelated, need higher res for final draft.	Replaced figure	
Reviewer 2	**Appendix C**	Suggest using empirical kernel density estimates / probability density functions rather than CDFs. It is harder visually to see the difference between CDFs, because the lines by definition have to start at 0 and end at 100.	We have encountered differing opionions on the best way to represent this data, but it seemed like CDFs were satisfactory to many.	
Reviewer 2	**Appendix C**	What kind of complexity?	Deleted this phrase	
Reviewer 2	**Appendix C**	For consistency, write mean values on plot.	We do not understand how to address this comments. We do not see other plots where the mean values are shown.	
Reviewer 2	**Appendix C**	Appreciated the details in the difference you used in the gradient section – enumerate differences if there's anything systematic here.	Patterns that were observed in the frequency distributions for tangential curvature are discussed in the text.	
Reviewer 2	**Appendix C**	Not sure what the standard is, but many of these plots are missing axis titles on the vertical axes.	The y axis label is indicated in the plot title and caption, instead, due to limitation in plotting software	
Reviewer 2	**Appendix C**	In all of these figures, the black outline obscures any color that would be shown within the polygon. Not sure how to better shows this info, but not appearing well in the map.	We could plot the maps at a different scale, e.g., 1:24,000, but they would no longer fit on an 8.5x11 inch page. When reviewing report digitally, colors can be seen easily when zoomed-in.	
Reviewer 2	**Appendix E**	Aren't these in appendix C, which is a prior appendix	Appendix E is included as a stand-alone section to educate readers unfamiliar with the concepts and methods used to characterize topography within mapped landform polygons. There may be some reduncancy with other appendices.	
Reviewer 2	**Appendix E**	Couldn't maps be produced that smooth this variability, thereby reducing differences between mappers?	In general practice, every practitioner applies their own set of protocols. The Forest Practices Board Manual offers guidance, but does not specify how remote and field mapping should be performed. Through experience, each mapper develops their own systematic approach. Smoothing of the DTM or of elevation derivatives could be applied, but the appropriate smoothing length and algorithm may depend on the quality of the DTM, on the nature of the terrain, and on what we want to see. It is possible that smoothing will reduce ability to resolve certain features, such as channel heads or slope breaks. In any case, we tried having our mappers all use consistent elevation derivatives and map scales for mapping of Site 1B, but found that this did nothing to improve agreement among mappers. This was an important finding. It highlights the need to provide tools for consistent use of available data and training in using those tools and data for mapping RILs and other landforms. It also points to the potential of using additional types of analyses for identifying potentially unstable terrain that the next sub-project will provide as an additional resource for practitioners.	
Reviewer 2	**Appendix E**	Colors and style should be consistent with other figures.	Colors represented in this figure do not relate to any other figures in this report.	
Reviewer 2	**Appendix E**	Why not? Mismatch between algorithms and what human mappers are using	Each mapper is relying on contour lines to estimate the degree of curvature; practioners do not generally use maps of calculated curvature, so they may not have a clear sense of how consistent their polygons are in terms of average curvature or even the range of curvature values encompassed. There may be other features that different mappers key into that would also result in differences in the downslope and lateral extent of their polygons, which would result in differences in mean curvature within a polygon.	
Reviewer 2	**Appendix E**	Add note that planar or divergent areas are not colored. I missed the blank spot in the legend at first.	Added to figure caption.	
Reviewer 2	**Appendix E**	Tidy up this graph. Vertical axis label on CDF is incorrect. Should read "cumulative proportion of hollow area less than"	Tidied up	
Reviewer 2	**Appendix E**	Label each panel with mapper name	Added labels for mapper number	
Reviewer 2	**Appendix E**	Suggest putting labels for debris-flow corridor and valley floor off of the actual features so it's possible to see the features in the topography. Right now the labels are blocking the feature	Moved label	
Reviewer 2	**Appendix E**	Refer to figures in other appendices.	Added in text	
Reviewer 2	**Appendix F**	Why don't these go to 100?	Because the horizontal axis does not extend to the maximum gradient-ratio values. The mapper's polygons included areas with much larger gradient-ratio values than the computer-generated polygons. Added text to the figure caption to explain.	
Reviewer 2	**Appendix F**	Notable that OBIA has larger number of polygons for all sites, and is probably the least rooted in geomorphic processes.	This is notable; a feature of having two models to work with. Having two models seeking to delineate the same landform types, but using different approaches, will help us to better identify idiosyncrasies of each and see how those differences relate to where landslides are observed.	
Reviewer 2	**Appendix F**	Clean up these plots (different sizes, not aligned). Plots are small enough that colors are very hard to tell apart. Suggest having a color family for computer-generated polygons that is distinct from the colors of the mappers. Same comment for all CDF plots.	Made the colors consistent across all frequency distribution figures and made them all a bit larger	
Reviewer 2	**Appendix G**	Several of these figures have spelling errors.	Revised	

Reviewer 2	**ES**	Landslide types that are most linked to bedrock hollows (debris flows) may not be represented in landslide inventory data, depending on the type of inventory, because the deposits are not as well preserved (go down channels)	We are familiar with a variety of problems with pre-existing landslide inventories, which is why the Susceptibility & Runout Study Design focuses on the identification of landslides through LCD. This will lead to "correct" placement of the landslide using the same DEM that the models are using, will find all significant landslides even if the deposit is missing, and comparison against a real field inventory will allow us to characterize the minimum observable size. This discussion is covered in Section 5.4; but it seems too detailed to cover it here in the Executive Summary (did add a bit to emphasize new LCD inventory).	
Reviewer 2	**References**	Reference needs to be updated in reference list, doesn't have a publisher or a journal. Also curious how methods developed for ifSAR (a DSM) operate on a DTM – how would this impact the methods.	This is a reference to a report; it has not been published in a peer-reviewed journal. The methods described in that report are currently used for USGS 3DHP elevation-derived hydrography with 1m lidar DEMs.	
Reviewer 2	**4.1**	It strikes me that the outcome of this report may be that there needs to be clearer definitions of RILs before proceeding to subsequent tasks.	That is the motivation for the Unstable Slopes Criteria project, as we described in the introduction. This document reports on the first in the sequence of sub-projects intended to address that issue.	
Reviewer 2	**4.1**	Issue of mapper's patience. Wonder if mapping a smaller area would produce a different result?	It's possible. We attempted to achieve a balance between a larger sample size and reasonable work load.	
Reviewer 2	**4.2**	What are the units of the z-value?	"z" here refers to the "z-value" for a desired level of confidence that we may infer when estimating the proportion of a large population with a given value (a "yes" vote in this case) from a random sample out of a large population. It corresponds to the number of standard deviations to the right of the (normal) distribution mean for which the included area under the normal-distribution density curve equals the desired level of confidence. It is unitless.	
Reviewer 2	**4.1**	I am surprised by the result that the OBIA has better overlap with mappers than the virtual watershed method. I would expect this to be the other way around, given that the VW method seems more closely related to geomorphic processes and topography. To me, this suggests that the VW algorithm is poorly tuned and/or the OBIA algorithm is overtuned – I see in the appendices that rules are relaxed in both cases to try to increase the match between the algorithms and the mappers (even though the mappers weren't consistent).	At this point in the sequence of USC projects, we only wanted to demonstrate that computer models can be used to generate landform maps based on a set of landform definitions that may change. In the next sub-projects, we will use precisely measured landslide locations and sizes to evaluate the criteria used to define landform locations and boundaries to see if we can refine those criteria to improve both computer-model and human-practitioner mapping of landslide-prone landforms	
Reviewer 2	**4.1**	Would be helpful to remind the reader which site this is and why it was chosen over others as the location for this specific exercise.	Added text at the end of the paragraph.	
Reviewer 2	**4.1**	Were they not all instructed to use the same scale?	Not the first time, which is why we specified scale in the following exercises.	
Reviewer 2	**4.1**	Why not use slopeshade? Hillshade biases depending on choice of sun location	We use a multidirectional hillshade in these figures because the polygons tend to show up better than on a slopeshade. I've added slope shade figures to Section 2.1.	
Reviewer 2	**4.2**	Numbers aren't adding up here … average of 0, 39, 68, 21 does not equal 65	Good catch! Revised in-text and added averages to table!	
Reviewer 2	**4.2**	This number doesn't make sense. Some of these bedrock hollows are surely in both datasets?	Each of the polygons drawn by the computer models were different. Even if they were in the same area, they were delineated diferently.	
Reviewer 2	**4.2**	What happened to prompts 3 – 5 in the survey of the expert practitioners?	These prompts are discussed in Section 5 (automated mapping limitations) as they provided a more qualitative analysis than quantitative.	
Reviewer 2	**4.1**	Why do you think the values for inner gorges increased?	The intent of this exercise was to see if having all mappers use the same mapping scale (1:500) and the same set of DTM derivatives resulted in improved alignment across the polygons they mapped. We expected an improvement (increase in balanced accuracy and in recall and precision), so were somewhat dismayed to see no apparent approvement for bedrock hollows. Perhaps the constraints on inner-gorge boundaries are better defined than those for bedrock hollows, e.g., it may be easier to consistently detect a break in slope for an inner gorge than the transition point from convergent to planar or divergent topography for a bedrock hollow.	
Reviewer 1		Similarly, there are too many acronyms, many of which are never returned to. If possible, I would encourage the authors to limit their use and only use an acronym if necessary and if it will be used again. Some acronyms are also never defined.	Added acronym list	
Reviewer 2	**3.3**	This sentence is hard to parse. As written, the method seems circular.	Rewrote the sentence.	
Reviewer 2	**4.1**	Specify this is happening pixelwise and that the confusion matrix values are total areas	Footnote 6 addresses this	
Reviewer 2	**4.1**	Denominator of equation should read "total comparison RIL area"	Fixed	
Reviewer 2	**4.1**	Did not understand this sentence. Please clarify.	Took out this sentence and the following.	
Reviewer 2	**4.1**	Refer to relevant figure that shows this.	Added reference to Figure 20	
Reviewer 2	**4.1**	More details about the nature of the issues that the model was adjusted to address (like an example figure) would be helpful	Figures and details are shown in Appendix G, as indicated in text	
Reviewer 2	**4.1**	I don't see clear evidence that there was a big difference in the distributions of the mappers' polygons and the algorithmic polygons. Please refer to a specific figure if I'm missing something.	Appendix G.Figures and details are shown in Appendix G, as indicated in text	
Reviewer 2	**4.2**	Be specific about where sediment is being delivered to.	Added text to specify delivery to a public resource.	
Reviewer 2	**4.2**	Define acronym	Expanded acronym	
Reviewer 2	**4.2**	If the OBIA doesn't use zero-order basin as a criterion for identifying bedrock hollows, I question the validity of this algorithm.	OBIA could use zero-order basins as a criteria, but those basins would need to be identified some way first. We could use the VW-identified zero-order basins, which would create an interaction between the two models. By not using zero-order basins, the current OBIA rule set provides a potentially useful comparison with the other (VW) model - perhaps it is missing some zero-order basins.	
Reviewer 2	**5.0**	Does this suggest VW is superior to OBIA? The former seems more closely related to actual geomorphic processes than the latter.	A model built using a virtual watershed data structure starts from a geomorphic context, whereas OBIA starts from the context of image segmentation. However, the strategies used with OBIA in determining how image pixels are grouped can also use a geomorphic perspective, which this implementation of OBIA does. Our hope is we can identify and capitalize on the strengths of both strategies; for example, that the image segmentation techniques used with OBIA can be better incorporated into a VW-based model to improve delineation of specific landforms.	

Reviewer 2	**6.1**	hile the board manual might not specify convergence, encourage looking at Clubb et al., 2014 and Pelletier, 2013 for guidance on this question.	Thank you, we will look at those references. We also want to consider the purpose of delineating and potentially restricting timber harvest within a bedrock hollow. Our goal is to prevent harvest-related increases in landslide potential. The current conceptual basis is to prevent loss of root strength within potential failure zones within a hollow and to preclude timber-harvest-related changes in the hydrology of a hollow. These two citations are focused on delineation of the channel head. There are other sources of information found in the literature (e.g., https://doi.org/10.1080/17499518.2019.1616098,https://doi.org/10.1680/jgeen.18.00209, https://doi.org/10.1007/s10346-022-01921-8) that we will capitalize on as we move into the next sub-projects.	
Reviewer 2	**6.1**	Do these data already exist or will new lidar flights be flown?	Added text pointing to existing data available through the Washington Lidar Portal	
Reviewer 2	**6.1, 6.3**	Landsliding is inherently stochastic, so I am uncomfortable with using a landslide inventory as an identifier of these thresholds. Also, the topography of a hollow is altered, likely towards more planar surfaces, post-failure, right?	We will have a sample from the population of possible landslides. Confidence in inferrences made from that sample will depend on the sample size and on the distribution of covariate values across that sample. Results of these analyses will be compared to similar analyses from previous studies. The study design for the next sub-project, Empirical Evaluation of Shallow Landslide Susceptibility, Frequency and Runout by Landform, has gone through UPSAG, CMER, and ISPR review and these issues have been extensively discussed. Generally, shallow landsliding within a hollow increases both the slope normal and tangential curvatures along the hollow axis because the elevation is reduced within the failure zone. We will use pre-landslide topography to calculate topographic indices to use as covariates for statistical analyses of landslide density.	
Reviewer 2	**6.1**	This seems like an obvious enough step that it should have been done as part of this sub-project and doesn't warrant a new sub-project.	We will need to extract the "spatially accurate road network" using image segmentation from the DTM. The need for this became apparent to us as we found the models were in some cases flagging road cut slopes as RIL polygons. This is not a new sub-project, only a lesson learned from this sub-project that we can now apply moving forward.	
Reviewer 2	**6.4**	Would help to restate Objective 2, since the introduction of the report was many pages ago.	Reorganized Sections 5 and 6, made clear what objective 2 is.	
Reviewer 2	**6.4**	Why not use aerial imagery?	See response to previous comments.	
Reviewer 2	**6.4**	What part of the landslide are you proposing to measure? The source area?	We are focused on shallow landslides with this study. These are typically small features, widths of 10 meters or less, so we will seek to extract the entire failure zone using lidar differencing.	
Reviewer 2	**6.4**	Some figures would be really helpful here. In the landslide mapping section, written as though the map will just have "landslide polygons" not discussing scarp, deposit, etc.	We are focused on shallow landslides. All we can map are "landslide polygons" and runout tracks. We may not be able to resolve deposits, particularly for landslides that transition into debris flows. There is a separate project examining larger deep-seated landslides.	
Reviewer 2	**7.0**	What do you mean by "machine-learning-based sensitivity analysis algorithms? Without specifics, this sounds buzzy.	This is buzzy, but not sure that we want to go into any more detail here because we're not sure yet what will work. This is a big topic. Examples for how to do it can be found in Molnar, 2025. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, christophm.github.io/interpretable-ml-book/	
Reviewer 2	Appendices	Because all analysis is based on elevation derivatives, suggest putting discussion of these derivatives (appendix E) first.	Appendices are organized in order of reference in report.	
Reviewer 2	**Appendix C**	Citation?	This sentence describes an observation from this study, it does not require a citation.	
Reviewer 2	**Appendix E**	This is less about "going wrong" than making decisions.	Unsure how to address this comment.	
Reviewer 2	**Appendix E**	Add references. See for example Grieve et al., 2016 and (Struble & Roering, 2021)	Added references.	
Reviewer 2	**Appendix E**	Other options include LSD TopoTools and TopoToolbox .	There are lots of open-source options, SAGA, WhiteBoxTools, QGIS, MultiscaleDTM, LSD TopoTools, but I am not aware of any that calculate gradient from a point looking upslope or downslope or that use an algorithm that is linear in runtime for elevation derivatives measured over any spatial scale. We are using this because we have complete control over the algorithm and the implementation, and it is licensed open source (GPLv3).	
Reviewer 2	**Appendix E**	Is there a better reference than a blog post available?	Added citation	
Reviewer 2	**Appendix F**	Also in Appendix C.	Added text to point that out.	
