Review of the revised ms by Smith et al.

General comments

The revised ms by Smith et al. is rather different from what has been presented at first. It has now a clear focus on the step-by-step description of the method used to map debris-covered glaciers and compared to the figure presented in the rebuttal it now seems that also the spectral mapping method works correctly. The authors now also show a direct comparison with outlines from other datasets and discuss shortcomings of the method in a more transparent way. In this regard I think it was worth giving the study another chance. Due to the major changes of the ms, there are now new points that should be improved in my opinion. They are related to smaller issues such as terminology and clarity of the figures, but also to larger issues such as a missing comparison to results from other, more advanced algorithms of debris classification. My main issue here is that the study basically compares results to either zero (only spectral mapping) or one (full manual delineation) with the obvious results shown in Figs. 9 to 12 on an aggregated level. However, a large number of more or less sophisticated methods have been developed in the meantime for mapping debris-covered glaciers (thermal bands, decision tree classifiers, object based image analysis, use of coherence images, etc.). Apart from one very simple method that is at least mentioned as a base for this study (but also not used for comparison), none of these other methods are mentioned or compared to.

I think it is beyond the purpose of this study to also apply all these methods and compare the results of the say 'velocity - river network approach' presented here to the outcome of these methods. But I expect at least a tabular summary of these other methods with columns listing what they need as an input, where they have been tested, and what their pros and cons are (e.g. regarding processing time, data availability, quality, required post-processing), as well as a direct comparison to at least one of the other methods (e.g. the method described in the study by Paul et al. 2004) that do not require extensive additional processing (like the velocity fields for this study) and can thus be easily implemented. This would also help to much better see what the advantages of the more complex method presented here is over what is already available. In short, I like the idea with the velocity fields and the river seed points but I need to see how results improve by the extra-effort required to get this information, and what this extra effort is (there is only a random note on this in L 358/9).

There are two further important points: one is the missing presentation of results (e.g. the total glacier area mapped) and one is the critical discussion of the method in view of recently published datasets (GAMDAM and the new Chinese glacier inventory). Why is it worth applying this method despite these new datasets and considering the intention that the method might not provide accurate outlines for individual glaciers? As a smaller point, I still do not understand why the TM and ETM+ data have to be co-registered to a master image. The level 1T product from USGS is in general highly accurate over glaciers (positional variability < 1 pixel), in particular the ETM+ scenes used to generate the GLS2000 global reference dataset. This extra effort seems distracting for others to apply the method and I think this is a bad idea. Maybe the authors can check what the effect of NOT additionally geocoding (with AROP) the already orthorectified level 1T product on the results is and remove this part from the pre-processing description if the effects are small.

I have given below some more specific and further minor comments that should be considered in a revised version of the ms. I hope these are sensible and can be implemented by the authors along with the more general points mentioned above. I leave to the editor whether the revised ms needs to be reviewed again or not.

S**pecific comments**

Abstract

I suggest rewriting the abstract with some care to better motivate the study (what is the key shortcoming that was responsible for this study? Certainly not that there are two (not several) inventories that provide only a one point in time snapshot). Show that the method presented here is not only a significant improvement over purely spectral-based classification (this is not a big deal and applies to several other algorithms as well) but that it is better than (in terms of accuracy, time required or whatever) than the current state of the art.

It not only limits longitudinal studies, it limits all studies (so just remove longitudinal).
Please use debris-covered glacier tongues instead of 'glacier debris tongues' and remove the sentence breaks.
'The relationship between Landsat band …' should be better specified (e.g. 'such as the band ratio with Landsat using bands …')

Introduction

L2: No, its changes in glacier length. We do not yet have any useful relation between area changes and climate change.

L3: I would not say more recent studies and although. There is neither a temporal preference nor is it related to remote sensing techniques (which would include DEMs derived from aerial photography. You might say that studies on volume changes have increased with free the availability of DEMs (e.g. SRTM) and altimetry (e.g. ICESat) data.

L5: Instead of Stocker et al. (2013), I recommend citing Vaughan et al. (2013), which is the Cryosphere Chapter.

L8: notably the one from GLIMS (Global Land Ice …) and … the Randolph Glacier Inventory (RGI) …

L10: I think it would be important to have a short comment also on the now available GAMDAM and new Chinese Glacier inventories. They both have delineated all the debris-covered parts manually and might be more accurate than the outlines provided from the algorithm presented here. To this end, it should be better described what the further benefits of the method presented here are.

L14: The cited study has a focus on technical challenges for glacier mapping. Maybe add another citation for the societal impacts.

L30: When the method presented here builds upon the methods developed by Paul et al. (2004), I think it would make much sense to directly compare the results to the outcome of this earlier method (outline overlay, quality, workload, issues for post processing, etc.). An improvement over the pure spectral mapping is rather easy to achieve.

L32: I am uncertain if these goals really apply to this study. The method is rather complex (requires different software packages and intense pre-processing) and not really tested globally (e.g. to the stagnant debris-covered glacier in the Mt. Everest region).

Study area

L43/44: This is all fine, but what about the variability in surface velocities? I think a critical part of the presented method is its applicability to glaciers flowing very slowly. Are there examples in the sample?

L59: It could also be possible to have a high diversity of glaciers in only one climatic zone. Or does this statement relate to differing mass balance gradients / temperature regimes?

L66: Which version of the SRTM DEM has been used, the one with voids or the interpolated version from CGIAR?
L68: What method was used for down-sampling?

Methods
L76/99: Naming two different sections "Data preparation" seems unfortunate. I suggest using a different name for section 3.1 (pre-processing?).
L103/4: Ok, but what has been done in these cases? Where snow/cloud-covered images just excluded or somehow corrected?
L106-113: I do not understand why this is required? Is the Level 1T orthorectification by USGS so poor? For a region to the south (Karakoram), we found some shaking mountain crests in a 15 year time series of Landsat TM, ETM+ and OLI images, but everything else was precise within a pixel. I would never touch this. As this step with AROP is related to additional workload, making the method presented here less attractive, I strongly recommend checking if it can be removed. Less is more!
L114: Maybe add what is calculated from the DEM
L121: This is a fine goal, but try to reduce the complexity of the method to the extent possible to make it more attractive.
L123: Debris-covered glacier tongues tend to …
L123: I think this statement needs to be more substantiated. What does 'tend to' mean? Is it the majority (say 90%) or only a few? As these lakes play an important (?) role for the algorithm, the question is what happens when they are not present and to how many glacier tongues this applies. Are the results without lakes as good as with lakes? I think this is an important assessment as global application has been mentioned as a possibility in the introduction and lakes on debris-covered glaciers might not be that common.
L129: All these algorithm-tuning steps performed manually need to be accounted for in the workload budget to allow an honest comparison of the increase in accuracy vs the increase in workload (compared to less complex methods).
L131: far away from any glacier
L134: see comment to L129
L142: Where does this value of >250 come from? The TM1 threshold is designed to improve classification in regions of shadow. Even fresh snow does not have DNs > 250 in shadow? I might be wrong, but from this statement I have to assume that even the simple band ratio method is not properly implemented in the processing workflow (?), thus providing results for the spectral mapping that are not as good as they could be.
L148: please write 'debris-covered glacier tongues'
L149: Why 'at high elevations' and not in shadow? What has the elevation to do with it?
L153: As mentioned in the general comments, when this method is building on this former study, it would be good to compare the results achieved with the methods developed here against it. The comparison against the pure spectral classification is much less interesting as there are meanwhile so many algorithms doing better.
L161: There are too many decisions merged in Fig. 3, the striped pattern selected for illustration is too imprecise, and the area shown is too large. So at first, please zoom in (to 1/5 of the image), second illustrate the effects of the respective binary masks in three subsets (slope threshold, slope variability, elevation range), and finally use a grey-scale background image and semi-transparent colour-coded areas to visualize the effects of each sub-step. The striped regions now also include the very steep headwalls of glaciers, but this cannot be true when regions steeper than 24 degrees are filtered. So please check and revise the figure.

L166ff: It would be good to show a classified velocity map also in the main paper. This is in my opinion the core of the here presented method and results might sensitively depend on the selection of the correct velocity threshold. Without knowing how the resulting velocity ranges over glaciers look like, it is difficult to imagine how Fig 4 was produced.

L177: I do not understand how it was possible to select snow free images for the correlation. The glaciers shown in Fig. 4 are heavily snow-covered and show deep shadows. How was it possible to derive meaningful correlations (from optical images) and henceforth velocities larger than a given threshold in these regions? I do even not see any noise in these regions, which is hard to believe. Please clarify.

L190: Please add how long the velocity processing normally takes. Without this information it is impossible to see if the extra-effort is worth the improved result.

L195ff: The distance filtering seems to be similar to the 8-point neighbourhood filter applied in the 2004 study, basically removing everything that is not connected to glaciers. Again, can an estimate of the required workload for these steps been added to get an impression on the required extra effort.

L213: As mentioned before, I only see here the river seed points (in Fig. 5) and the effects of everything in Fig. 6. But I do not see the map with the distances (for the various datasets) that has been used as a base for the removal. What are the distances that have been used as a threshold? Please add this information (parts of it was already shown in the rebuttal).

L216-226: As above, it would be nice to see the effects of the filtering steps on the binary masks (for a set of close-ups, not at the scale of Figs. 6 or 7). The filtering seems to be rather massive and I assume there is also some impact on smaller clean ice glacier extents. The effect can be quantified in the results or discussion, but it should be shown that it is a minor one (at least I hope so).

L227: Please compare results with what can be achieved (I assume much faster) with the Paul et al. (2004) method. The comparison against the pure spectral classification is has a very limited meaning. Also: Place the spectral outlines on top of the final outlines to better see the real difference.

L230: Can a little bit of statistics be provided for this manually digitized dataset, for example the size class distribution and debris cover percentages for each class?

Results

L248: The results section now contains a discussion of algorithm errors, a presentation of a further method (vertex distance matching), and a statistical comparison to a random sampling of glacier areas. There are no numbers about the derived glacier areas and how they compare to other datasets (e.g. the RGI, GAMDAM or new Chinese datasets). Is there a chance to add some results that look more like results? I suggest describing the vertex distance matching method in the methods section.

L250: I do not understand why the 750 glaciers are now reduced to 215 for comparison. What was wrong with the others? Has this sample still the same size class distribution?

L254: I recognize that the method has not been developed to provide accurate outlines for individual glaciers. However, I find the comparison to elevation distributions too aggregated. Is there a possibility to add a scatter plot (size vs. relative size difference) showing how individual glaciers compare (maybe marking the heavily debris-covered glaciers with a different symbol)? Such a comparison would also justify the performed separation into individual glaciers (which is not required when only hypsometry is compared).

L276/7: I recommend a comparison with a dataset derived from a more sophisticated method.

L291: we used 465 …

Discussion

300: It would be nice if this part (5.1) of the discussion could expand a little bit on other methods that have been tested previously. Currently it is largely centred on what has been tested in this study. However, there are also object-based classification approaches, neural networks have been tested, and pattern recognition (also: thermal bands, decision tree classifiers, coherence images, and hybrid approaches). Maybe put these into context.

303: "algorithm, but neither provided" (although sounds as it was clear from the beginning that they will not work)

L314: Please note that it is recommended to use glacierized rather than glaciated when referring to contemporary glaciers.

L329: Please show that it also moves a step forward compared to more complex approaches, (such as Paul et al., 2004), that can be easily implemented here. That inclusion of additional measures (slope, vegetation, neighbourhood analysis) improves classification of debris-covered glaciers over pure spectral approaches is known since 2004.

L341: Please add here a short discussion on the results achieved when directly comparing individual glaciers (see comment at L254).

L345: This might be well the case, but is this something anybody would really do? As far as I know, scientists tend to always want to have the best possible dataset, independent of the scale of their application. And if a freely available source does not satisfy there quality needs, they digitize everything by themselves. Please also note that we now have the GAMDAM inventory and the new Chinese inventory for that region (both not yet in RGI 4.0). Maybe it would be sensible to demonstrate that the dataset created here is not obsolete in this regard.

L346: 'powerful tool': this statement requires some information on the required workload for data processing, considering all steps that are really important (a high workload would imply that the tool is not that powerful).

L351: 'mostly static': At least in this part of the world, try it in the Karakoram …

L354: As mentioned above, demonstrate that the method presented here can be a substitute for algorithms such as Paul et al. (2004). Being better than pure spectral mapping is easy.

L358/9: Here it is, a statement on processing time. Please expand on this to be transparent.

L364: Indeed, this is now the problem. Please demonstrate why the method presented here is worth to apply anyway.

L371-379: This sounds a little bit theoretical. I hope it can be replaced with a discussion of more practically relevant topics after revision.

L372: The glacier ice detection is mostly based on the SWIR band (where reflectance is very low) and TM1 in shadow, where saturation is also not a problem. Maybe think of a better explanation.

Conclusion

L383/4: Please see above, it needs to outperform simple hybrid classifiers to be worth testing.

L393/4: Please test this in the Mt. Everest range before stating it. Flow velocities are close to zero in this region for many of the heavily debris-covered glaciers.

L396: What is now the recommendation when I need accurate outlines for a larger region? Use first this algorithm and then manually correct the remaining errors or a native digitization of these parts?

Figures

General remarks

In my review above, I have suggested to add several further figures. These should have the form of a regular square to show at least too images side by side. Compared to Fig. 2, I al-

so suggest showing a close-up (maybe 2/3 of the current Fig. 2). The regions currently shown in the various figures differ. I think this is not a good idea to trace the effects of the different processing steps. I suggest showing the processing steps in only one or maybe two different regions.

Fig. 1: Caption: "and location of eight Landsat image footprints … along with their path/row combinations." By the way, why has scene 146/30 & 31 not been used to close the gap?

Fig. 2: Is there also an image available with less seasonal snow? It also seems as if the glaciers in shadow are not properly mapped. I might be wrong, but I think the band ratio can do better (by changing the TM 1 threshold).

Fig. 3: See comments to L161. Please show the same region as in Fig. 4 or 5.

Fig. 4: The red on red is difficult to see. Please use a different colour and add the velocity map the mask is based on. Caption: This is not really a binary mask. It is an overlay of the velocity mask with an RGB composite image showing included regions transparent and excluded regions in red. Please be precise with the caption. Is this really a 7/5/3 composite in the background? It more looks like a 7/4/3 or 5/4/3 composite.

Fig. 5: I suggest to also show the other seed points (e.g. lakes etc.) in this image. Can the image also illustrate what is distant, i.e. which regions will not be considered? Caption: 'The blue lines illustrate the presence of …'

Fig. 6: Legend: The 'unfiltered outlines' look more like 'unfiltered areas' (i.e. the polygons are filled). Legend: 'Velocity threshold' is unclear. Are the black regions those that are still included after a threshold is applied? In this case I am not quite sure what the benefit of the velocity calculation is. I see velocity noise all over the image.

Fig. 7: The blue outlines surrounding the blue glaciers are difficult to see, maybe use yellow? Maybe label some main glaciers in Fig. 2 to indicate where they are (Inylcheck, Tomur). Legend: Instead of 'unfiltered outlines' I would label the black polygons 'removed after filtering'.

Fig. 8: As above: maybe use yellow lines instead of blue and place the red ones on top to better see what has been added by the algorithm.

Fig. 9 to 12: Please remove the title from each of the plots and introduce some minor tick marks on the x and y axis on all plots. Please also consider showing two of these plots side-by-side (naming them Fig. 9a and 9 and 10a and b). All captions: I think the interpretation of what the figures show (this indicates) is not required in the caption but should be properly explained in the main text.

Fig. 13: The orange lines are difficult to see, I suggest using yellow instead.

Fig. 14: I think this is a bad example for illustrating glacier changes. The Landsat scene from Oct. 5, 2002 suffers from intense seasonal snow. In consequence, most of the 'area' changes visible in the image are due to the reduction in seasonal snow. The only real changes can be found at the terminus of the five clean-ice glaciers along the middle part of the image. At the 'debris-tongues' there is no change at all as far as I can see it. Maybe use a different example. The 'commonly misclassified river sand' needs to be marked (why has this type of misclassification not been removed with a vegetation filter?). The red outlines on the reddish background image are difficult to see, I suggest using yellow instead.