On the accuracy of glacier outlines derived from remote-sensing data

F. PAUL,[1] N.E. BARRAND,[2] S. BAUMANN,[3] E. BERTHIER,[4] T. BOLCH,[1] K. CASEY,[5] H. FREY,[1] S.P. JOSHI,[6] V. KONOVALOV,[7] R. LE BRIS,[1] N. MÖLG,[1] G. NOSENKO,[7] C. NUTH,[8] A. POPE,[9] A. RACOVITEANU,[10] P. RASTNER,[1] B. RAUP,[11] K. SCHARRER,[12] S. STEFFEN,[13] S. WINSVOLD[14]

[1]University of Zürich, Zürich, Switzerland
E-mail: frank.paul@geo.uzh.ch
[2]British Antarctic Survey, Natural Environment Research Council, Cambridge, UK
[3]Technical University Munich, Munich, Germany
[4]Legos, Centre National de la Recherche Scientifique, Université de Toulouse, Toulouse, France
[5]Cryospheric Sciences Branch, NASA Goddard Space Flight Center, Greenbelt, MD, USA
[6]International Centre for Integrated Mountain Development, Khumaltar, Kathmandu, Nepal
[7]Russian Academy of Sciences, Moscow, Russia
[8]University of Oslo, Oslo, Norway
[9]Scott Polar Research Institute, University of Cambridge, Cambridge, UK
[10]Laboratoire de Glaciologie et Géophysique de l'Environnment, CNRS/Université Joseph Fourier – Grenoble I, Grenoble, France
[11]National Snow and Ice Data Center, CIRES, University of Colorado, Boulder, CO, USA
[12]Environmental Earth Observation (ENVEO), Innsbruck, Austria
[13]University of Colorado, Boulder, CO, USA
[14]Norwegian Water Resources and Energy Directorate (NVE), Oslo, Norway

ABSTRACT. Deriving glacier outlines from satellite data has become increasingly popular in the past decade. In particular when glacier outlines are used as a base for change assessment, it is important to know how accurate they are. Calculating the accuracy correctly is challenging, as appropriate reference data (e.g. from higher-resolution sensors) are seldom available. Moreover, after the required manual correction of the raw outlines (e.g. for debris cover), such a comparison would only reveal the accuracy of the analyst rather than of the algorithm applied. Here we compare outlines for clean and debris-covered glaciers, as derived from single and multiple digitizing by different or the same analysts on very high- (1 m) and medium-resolution (30 m) remote-sensing data, against each other and to glacier outlines derived from automated classification of Landsat Thematic Mapper data. Results show a high variability in the interpretation of debris-covered glacier parts, largely independent of the spatial resolution (area differences were up to 30%), and an overall good agreement for clean ice with sufficient contrast to the surrounding terrain (differences ~5%). The differences of the automatically derived outlines from a reference value are as small as the standard deviation of the manual digitizations from several analysts. Based on these results, we conclude that automated mapping of clean ice is preferable to manual digitization and recommend using the latter method only for required corrections of incorrectly mapped glacier parts (e.g. debris cover, shadow).

1. INTRODUCTION

How accurately can glacier outlines be mapped? This question is difficult to answer. Yet it is an important question, as glacier area and length changes need to be precisely assessed to be valuable as an indicator of climate change. A wide range of issues have to be considered to delineate glaciers, for example the interpretation of seasonal and/or perennial snowfields, handling of glaciers covered (partly) by clouds, identification of drainage divides in the accumulation region, the correct delineation of debris-covered glaciers, the determination of the terminus for calving glaciers or those with dead ice, or the interpretation of invisible glacier boundaries in cast shadow. Most of these issues were addressed in a workshop organized by the Global Land Ice Measurements from Space (GLIMS) initiative, and are discussed at length in Racoviteanu and

others (2009). The current study summarizes the results of a round-robin experiment with a focus on comparing glacier outlines digitized manually on satellite images with different resolutions and from automated methods.

To be sure that observed glacier changes are related to real changes rather than caused by imprecise determination of the outline, the accuracy of the outlines must be known (i.e. changes should be larger than the accuracy in order to be significant). Furthermore, the definition of the exact perimeter of a glacier as an entity (that can be represented by a polygon in the vector domain) is non-trivial. Two key points apply: First, the entity definition depends on the purpose. For example, hydrologists prefer to separate an ice cap into its main drainage basins, whereas glaciologists tend to keep it as a single unit (Racoviteanu and others, 2009). If there is agreement on the use of drainage divides in general, their

location can be defined in various ways. Applying hydrologic analysis to a digital elevation model (DEM) of the glacier surface for watershed determination is common practice (e.g. Andreassen and others, 2008; Schiefer and others, 2008; Bolch and others, 2010), but it is mandatory to specify the DEM used, as DEMs can differ in quality (e.g. in low-contrast regions). Another issue is the treatment of tributaries. Neglecting for the moment that tributaries change through time (e.g. they might become disconnected during retreat), the required degree of contact to be counted as being part of a glacier can be controversial. To some extent this also applies to ice above the bergschrund, which should be considered as part of the glacier but is often difficult to identify due to snow cover. Hence, large differences in glacier area can be expected due to a different interpretation of a glacier as an entity by an analyst. And second, glaciers are elements of nature with a large variability of shapes and in strong interaction with their surrounding environment. Both points require that rules be applied for a consistent glacier entity definition. For glaciers observed by remote sensing, such guidelines were developed in the framework of the GLIMS initiative in the form of the GLIMS analysis tutorial (Raup and Khalsa, 2007). For example, according to these guidelines all ice above the bergschrund and all debris-covered ice that is connected to the main glacier should be included as a part of the glacier, whereas seasonal snow, lakes with icebergs at the terminus or dead ice are not part of the glacier. But even if strictly followed, these rules lead to some variability in interpretation of the 'true' glacier extent, as the identification of debris-covered glacier parts, seasonal snow, dead ice or ice in cast shadow can be challenging (Racoviteanu and others, 2009). In this regard, finding a 'ground-truth' or reference dataset that can be used to validate glacier outlines is also non-trivial.

It is generally assumed that glacier extents derived from coarser-resolution sensors can be validated by comparing them to results from a higher-resolution sensor (e.g. 30 m Landsat Thematic Mapper (TM) with 1 m Ikonos). Though this might be true in a general sense, several constraints have to be considered to avoid invalid comparisons (cf. Svoboda and Paul, 2009):

1. the images from both sensors should be acquired around the same date (week) within a year to avoid changes due to different snow conditions;

2. at a higher spatial resolution, new details on the surface become visible that require a different interpretation;

3. the common higher-resolution sensors (including aerial photography) lack a spectral band in the shortwave infrared (e.g. Landsat TM5), so the base data for the interpretation are different;

4. a lower contrast between glacier ice and the surrounding rock in the visible and near infrared (VNIR), as available from high-resolution sensors, can degrade the interpretation substantially (i.e. the perimeter can be virtually invisible); and

5. for objects with a shape that is more complex than a simple square (i.e. glaciers), the area also varies with spatial resolution of the image. This has to be considered when glacier sizes derived from high- and lower-resolution sensors are compared (e.g. Paul and others, 2003).

In practice, it is difficult to rule out all the above limitations, and the 'validation' with a higher-resolution dataset can easily be considered scientifically unsound. When the above-mentioned differences in interpretation of glacier extent by different analysts are also taken into account, reporting accuracy correctly is indeed challenging. Another important point is the comparison of automatically derived glacier outlines with manually digitized ones. Here two issues must be considered: (A) apart from a few regions without lakes and debris-covered glaciers, the automatically derived glacier outlines are usually also corrected manually (by visual interpretation of the satellite image), and (B) manually digitized glacier outlines differ in each digitization (even when performed by the same person), as the degree of generalization (e.g. spatial averaging over several pixels, number of vertices used for the line, interpretation of subtle differences in colour) varies each time. Hence, manual digitizing gives inconsistent and generalized results that are difficult to reproduce. This is also a point to consider for change assessment (e.g. Hall and others, 2003). So what is the way forward to assess the accuracy of glacier outlines?

A possible way to analyse this problem is to perform multiple and independent digitizations of a couple of glaciers within a single satellite scene, and, if available, also on high-resolution data for the same glaciers (e.g. Berthier and others, 2009). If glaciers of different size, degree of debris cover and shadow conditions are selected, several conclusions are possible:

a. the multiple digitizations of the high-resolution data provide a mean value for the respective glaciers that can be used as a reference;

b. the multiple digitizations of the lower-resolution data provide (i) a measure for the precision of the analysts digitizing (standard deviation of the derived area values), (ii) a mean value for comparison with an automatically derived extent, and (iii) the possibility of determining whether the automatically or the manually derived extent is more accurate (by comparing the difference from (ii) to the standard deviation (STD) of (i));

c. identification of regions with large differences in interpretation by overlay of outlines; and

d. calculation of (perhaps systematic) omission and commission errors among the different digitizations.

It has to be noted that (iii) in (b) of the list above is a comparison of accuracy (difference from a reference value) with precision (internal variability or random uncertainty). The underlying assumption is that the precision of (i) must be higher than the accuracy from (ii) to perform better. As the above assessments will not be performed and reported routinely with every study related to the determination of glacier area and its changes, in this study we present such a comparison for a couple of selected glaciers from different regions of the world. The general idea behind this comparison experiment with many participants (named 'round robin') is based on previous comparisons performed in the framework of GLIMS (GLIMS analysis comparison experiment, GLACE). They revealed large differences in the interpretation of glacier entities in general, and led to the guidelines presented in the GLIMS analysis tutorial (Raup and Khalsa, 2007). In this study, the focus is on the multiple digitizations of mostly well-defined glacier entities for
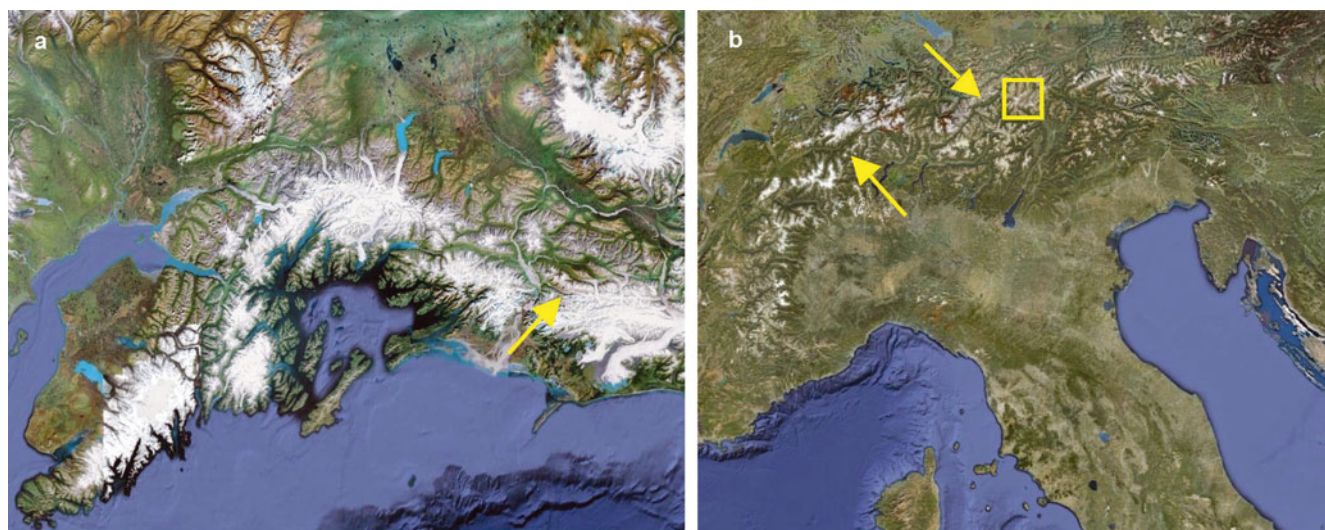
Fig. 1. Location of the test sites in Alaska (a) and the European Alps (b). The yellow box denotes the location of the test site in the Ötztal Alps. More detailed images are shown in subsequent figures. Images: screenshots from Google Maps™, © 2013 TerraMetrics.

accuracy assessment rather than investigating the above-mentioned methodological issues.

This study is performed within the framework of the European Space Agency (ESA) project Glaciers_cci. The CCI (Climate Change Initiative) projects have a focus on error characterization of operationally derived essential climate variables (ECVs) following the climate-monitoring principles of the Global Climate Observing System (GCOS) (2003). The mapping of the spatial extent of the ECV 'glaciers and ice caps' as defined by GCOS (2004) is already in quite a mature state, as well-established, fast, accurate and thus frequently applied algorithms exist. However, accuracy assessment of mapped glacier outlines is still somewhat unspecific in the related studies, mainly due to the limitations discussed above. The results presented here are part of the required product validation for the Glaciers_cci project and belong to a larger round robin that also includes the determination of a 'best' glacier-mapping algorithm for a test region in the Himalaya. In this study, 21 glaciers in five satellite images from two regions (Alaska and the European Alps) were digitized one to three times by up to 20 participants (Table 1). As a validation measure, we compare glacier extents derived automatically from Landsat TM scenes, with the results of the round robin and outlines digitized from high-resolution satellite images and aerial photography.

2. STUDY REGIONS AND DATASET DESCRIPTIONS
The two study regions are located in the eastern Chugach Mountains, Alaska, USA, and in various parts of the European Alps (Switzerland, Austria, Italy), as shown in Figure 1. The major selection criteria for the regions and the selected glaciers were:

> variability with regard to glacier size and mapping challenges (debris, shadow, snowfields),
>
> a feasible number of sites and glaciers to facilitate wide participation,
>
> scenes from the end of the ablation period with seasonal snow close to a minimum,
>
> availability of high-resolution scenes with good contrast in Google Maps™,
>
> a good temporal coincidence or at least comparable snow conditions of the high- and lower-resolution images, and
>
> well-defined glacier entities in all images used for multiple digitization.

Of course, many other regions fulfil these criteria, and the finally selected regions are only examples.

The test region in Alaska (centre 60°43′ N, 144°9′ W) uses a QuickBird scene from 27 August 2003 (Fig. 2) and includes eight glaciers of different sizes with a varying degree of debris cover, shadow and some seasonal snow (mostly off-glacier). For this region, the focus was on comparing outlines derived by different participants (one-time digitization) using high-resolution (2.4 m in this case) satellite data. The test region in the Alps includes one glacier in the Gotthard (centre

Table 1. Overview of the input datasets used for the round robin (rr). Set No. 3 is processed by the team at University of Zürich and used for validation (val). The 'Scene' column gives the path-row for the related Landsat scene

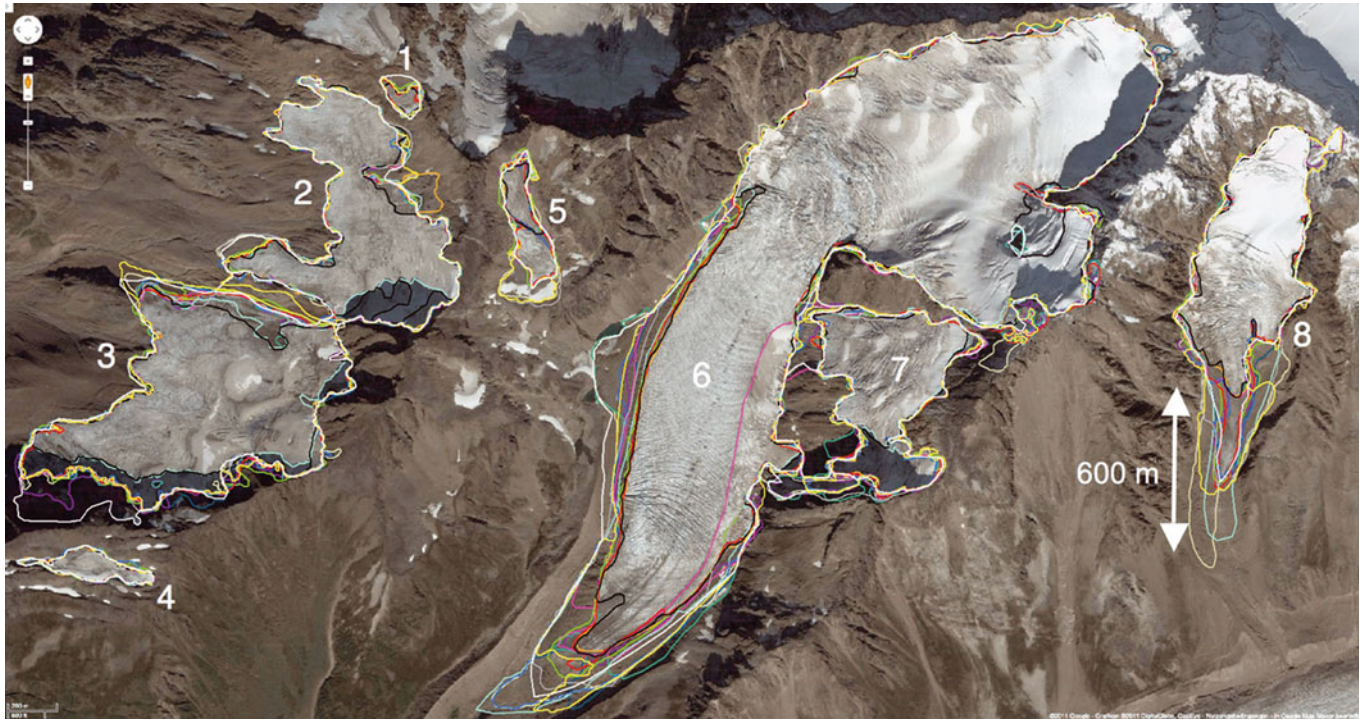| No.	Region	Scene	Sensor	Date	Task
1	Alaska (rr)	066-017	QuickBird	27 Aug 2003	Digitize 8 glaciers once
2	Ötztal Alps (rr)	193-027	Landsat TM	30 Jul 2003	Digitize 10 glaciers 3 times
3	Ötztal Alps (val)	193-027	Ikonos	12 Aug 2003	3 glaciers digitized 5 times (internal)
4	Switzerland (rr)	194-027	Aerial photography	4 Oct 2009	Digitize 3 glaciers at least once

Fig. 2. Overlay of the outlines for the eight glaciers from the test region in Alaska. Background image: screenshot from Google Maps[TM], © 2013 DigitalGlobe, GeoEye.

46°31.5′ N, 8°28′ E) and two glaciers in the Silvretta region (centre 46°51/52′ N, 10°11/12′ E) of Switzerland based on aerial photography, and a Landsat TM scene (centre 46°50′ N, 10°25′ E) of the Ötztal Alps (Fig. 3) including three glaciers covered by high-resolution Ikonos images. In this case, the TM scene was acquired only 14 days before the Ikonos scene, so a validation of the results obtained with Landsat is possible. Ten glaciers on the Landsat scene were selected for multiple digitizing (three times), and contrast-enhanced red, green, blue (RGB) composites (TM bands 3 2 1, 4 3 2 and 5 4 3 as RGB, respectively) were provided to the participants for this purpose. The three glaciers depicted on aerial photography were digitized at least once, and the three glaciers selected for validation were not made available to the participants but digitized by the participants from the University of Zürich. For the Landsat TM scene, glacier outlines were also derived automatically.

The screen shots from Google Maps[TM] are already orthorectified, as they have to match with an underlying DEM. To use them for the digitization of glacier outlines, their upper left corners were iteratively geocoded by changing the coordinates in the tfw (tif world) file until they matched with the respective (also orthorectified) Landsat scene (level 1T product from the United States Geological Survey). The pixel size was derived from the scale bar on each image. Compared to the pixel size of Landsat and at the scale of individual glaciers, we found that the resulting geolocation accuracy was sufficient (i.e. no systematic shifts were observed). The images for the three regions were provided in Geotif format along with a tfw file. The screen shots did not always have the original resolution of the high-resolution datasets and included local artifacts (e.g. from jpg compression). However, this does not impact on the value of the round robin, as all participants used the same datasets and a variety of spatial resolutions and challenging mapping conditions should be investigated in the round robin anyway.

3. METHODS

3.1. Instructions for the participants

To guide the digitizations of the participants, detailed instructions were provided in a short document describing the reasons for the round robin, the datasets, how to perform the work and a feedback form. Internal tests revealed that this part of the round robin would take ~4–6 hours for one round of digitization. The participants were free in working with their specific software, the magnification applied during digitizing and the level of detail (the number of vertices digitized per polygon). They were also allowed to check for details with Google Maps[TM] or other sources of high-resolution imagery and no time limit was given. All participants were asked to report meta-information (e.g. time taken, additional sources used) in a feedback form.

The interpretation of the glacier extent as seen on the respective image was the core of the exercise, so no additional advice was provided for that. For the multiple digitizations of the same glaciers by the same analyst, it was not permitted to refer to the previous digitization (using that outline in the background) and we asked the analysts to wait 24 hours for the next digitizing round to not remember the details of the previous digitization and obtain independent interpretations. The participants were free to provide results for only some of the test regions, but most participants provided results for all.

3.2. Analysis of the results

All shapefiles with the glacier outlines were superimposed with the respective images in a GIS software, and glacier area (m^2) was extracted for each glacier and participant. For each digitization we determined (1) the total area of all glaciers in a region per participant, (2) the mean size and STD of all digitizations for each glacier and (3) the relative difference of the automatically derived glacier area from the
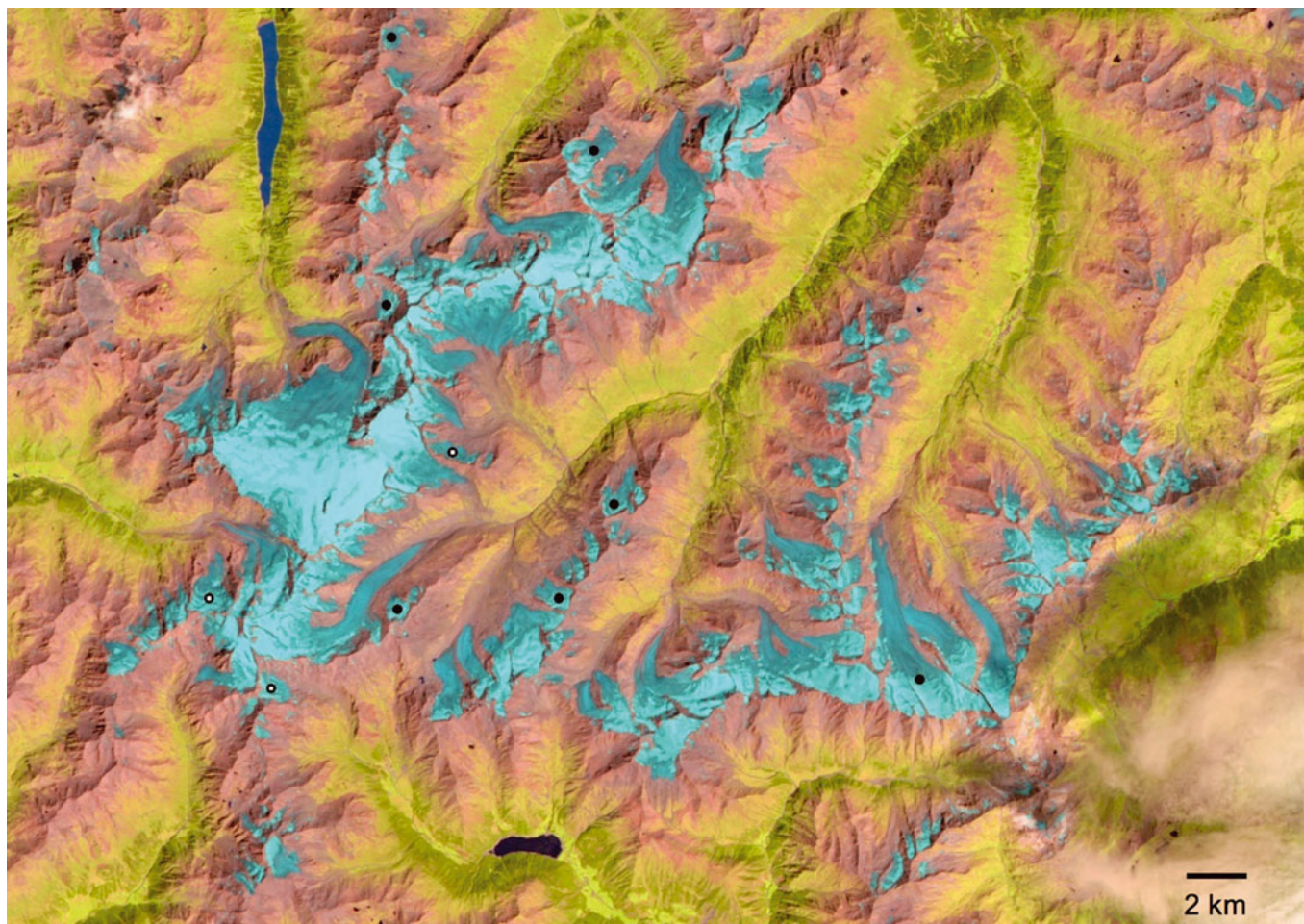
Fig. 3. Location of the glaciers in the Ötztal Alps that were selected for the multiple digitizations (marked with a black circle) on a Landsat 5 TM band 543 composite. The three glaciers with a white dot in the black circle are selected as reference glaciers for the digitizing with Ikonos.

mean of all manual digitizations. This was also done for (4) the total glacier area per participant. Where a second (and third) set of digitizations was provided, the same analysis as described above was performed. Furthermore, overlays of all outlines were created to obtain a visual impression of the digitizations and to identify the most critical regions (i.e. with the largest variability in interpretation). This qualitative comparison is also an important complement to the quantitative comparisons described above, as the same glacier area could result from outlines at very different locations (i.e. wrongly interpreted). Based on these overlays, we decided not to calculate omission and commission errors explicitly, as the variability of the outline locations was more or less random, i.e. without a preference for including or excluding certain glacier parts.

3.3. Glacier mapping with Landsat TM

The glacier outlines from TM were derived with the well-established band-ratio method (e.g. Albert, 2002; Paul and others, 2002) and have been created for the test region Ötztal Alps in an earlier study (Paul and others, 2011). Glaciers were classified when the raw digital numbers of band TM3 were 1.8 times higher than in TM5 (TM3/ TM5 > 1.8). Apart from one glacier with debris cover in this selection, the outlines were used as is (and not adjusted for the round robin) to have the required independence of the datasets. For the glacier with debris cover, the outline was

corrected afterwards to have a comparable value. Of course, this corrected outline is now different from the fully automatically derived extents, and the difference from the fully manually digitized extents stems more from the variability in interpretation rather than from the accuracy of the automated mapping. This has to be considered when comparing the results.

4. RESULTS

4.1. Overlay of outlines

In Figure 2 an overlay of the digitizations from the different analysts is shown for the test region in Alaska using the QuickBird image with the eight selected glaciers numbered. Although bare ice was digitized highly consistently by all participants (Fig. 4a), some of them missed the shadowed parts of glaciers 2 and 3. Bare rock in the glacier forefield was partly included (glaciers 2 and 7) and the debris cover between glaciers 2 and 3 was not detected. The region with rockfall on glacier 5 was interpreted very differently by the participants. Large differences in interpretation also occurred for the debris-covered tongues of glaciers 6 and 8. The resulting maximum differences in glacier length are 500 m for glacier 6 and 600–700 m for glacier 8. As the close-ups of glaciers 5 and 8 in Figure 4 reveal, higher resolution does not really help to achieve a 'better' or
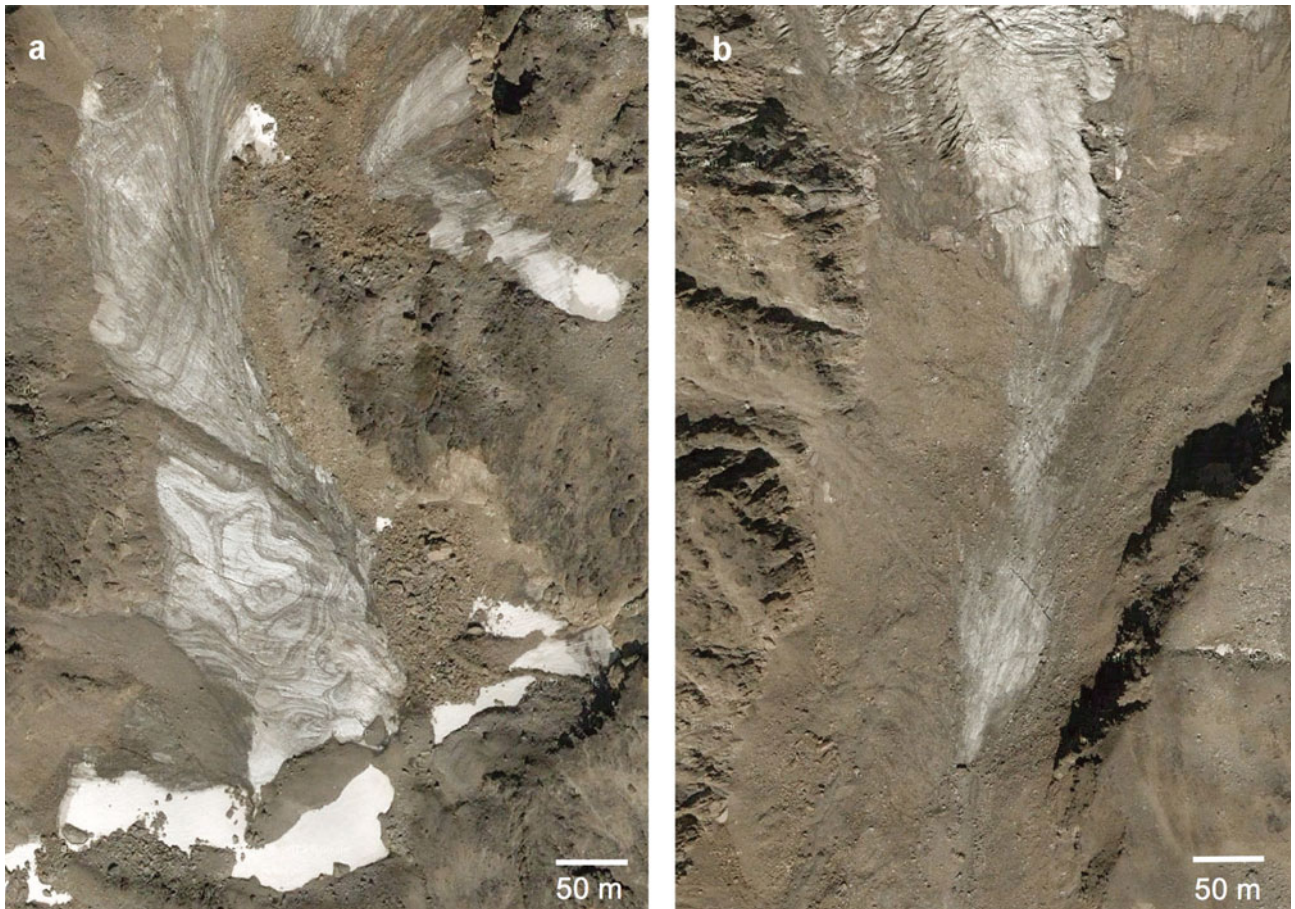
Fig. 4. Close-ups showing glacier 5 (a) and the terminus of glacier 8 (b) from test region in Alaska (see Fig. 2). The boundary of the ice can, despite the high spatial resolution, only be roughly estimated. Images: screenshots from Google Maps™, © 2013 DigitalGlobe, GeoEye.

'clearer' interpretation. In such cases, the location of the outline and the terminus is a subjective issue allowing a wide range of interpretations.

The interpretation of glacier extents on the even higher-resolution aerial photography depicted in Figure 5 confirms the results from the analysis of the QuickBird image described above. For clean ice and with good contrast, all participants digitized the outlines very similarly (i.e. the variability mostly stays within one pixel). When debris covers the surface (Fig. 5b) or the tongue (Fig. 5c), the interpretation is more variable. The outline derived automatically from a TM image acquired in the same year (Fig. 5a) is visually in good agreement for this very small glacier (0.03 km²) apart from two or three TM pixels at the terminus. Apart from a region with rockfall and ice above the bergschrund that were partly excluded, the manually digitized outlines are fairly consistent. This is different for the glacier depicted in Figure 5b, where bare rock in the glacier forefield and above the highest point of the glacier was digitized differently. For the largest glacier in this sample (Fig. 5c) the correct interpretation of the seasonal snow above the accumulation area was not a problem, but the differences in the terminus position are up to 300 m. This is the same value as the retreat of the terminus between 2003 (acquisition date of the TM scene, indicated by the white outline) and 2010 (the year of acquisition of the aerial photography). This implies that a retreat of a glacier tongue over such a distance might not be detected when different analysts provide the outlines for different years.

In Figure 6 we show the overlay of the digitizations by different analysts for the test region in the Ötztal Alps for a selection of six out of ten glaciers (first digitization). In this region the interpretation of the glacier extent is facilitated by the availability of TM bands 5, 4 and 3 (as RGB) composite that shows clean glacier ice and snow in a light-blue to cyan colour and that is not available for high-resolution sensors. On the other hand, pixel size is only 30 m and the interpretation of small details might be more difficult. However, apart from the glaciers depicted in Figure 6a and c, where a snowpatch and debris cover cause some larger variability in interpretation, the outlines are very consistent. In general, their location varies only by about 1 TM pixel (i.e. ±15 m), but one outline is often found outside all others. The outline derived automatically from TM (white) is mostly located within the variability of the manual digitizations. A somewhat larger variability in the outline location when two analysts digitize the same glacier was also found by Berthier and others (2009).

Most of the participants digitized the glaciers in this test region (Ötztal Alps) three times. This allows us to also determine the analysts' internal variability in interpretation for the same glaciers. The outline variability in the second and third digitization when comparing all analysts' results has the same characteristics as for the first digitization shown in Figure 6, so we do not show again the related cross-analyst comparison. Instead, we show in Figure 7 for three analysts and four glaciers (Fig. 6a–d) overlays of the multiple digitizations. These overlays reveal that the
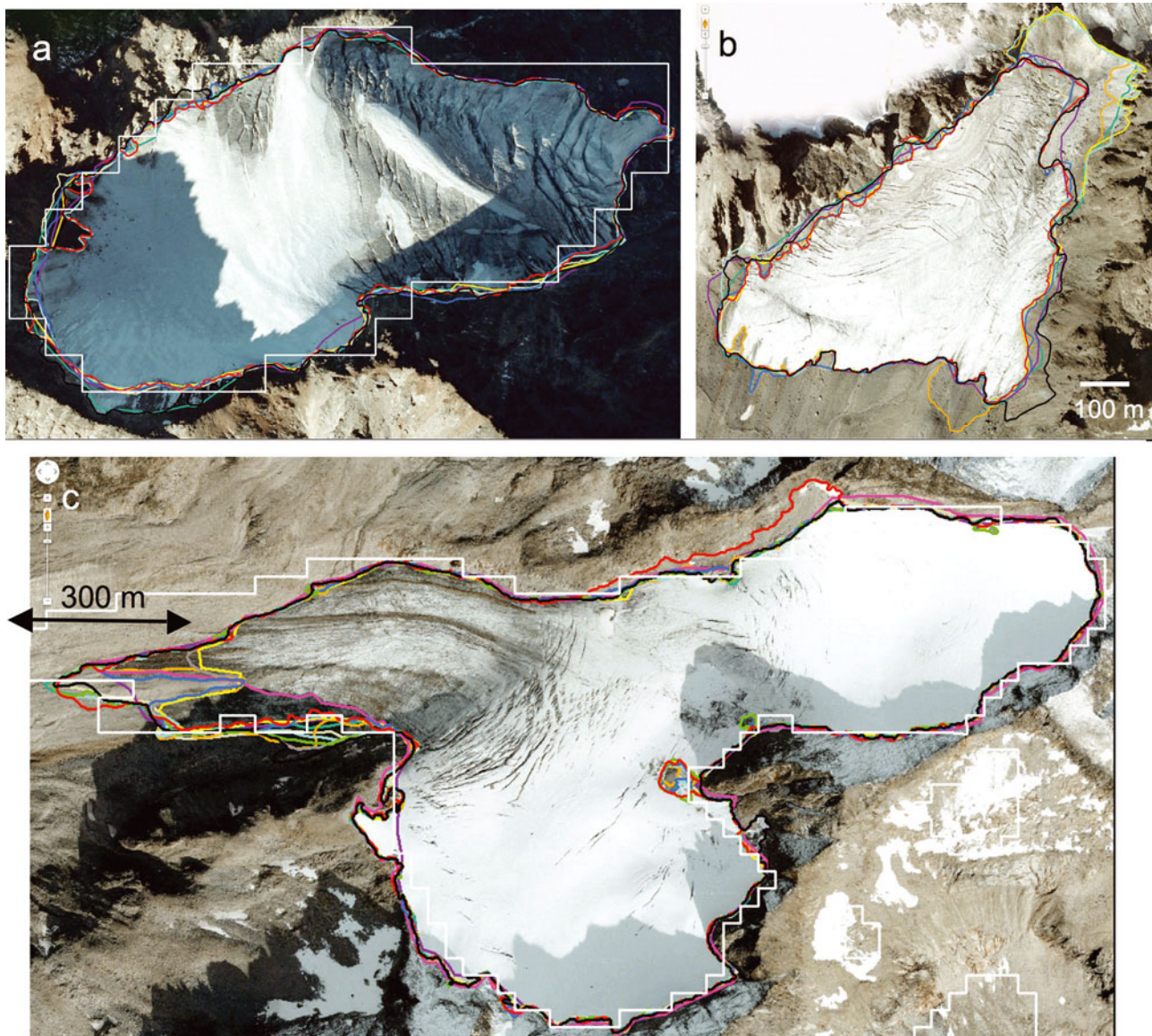
Fig. 5. Overlay of manually digitized glacier extents for the three test glaciers in the Swiss Alps: (a) Vadret Futschöl, (b) Vadret d'Urezzas and (c) Geren glacier. One raster cell of the white line is 30 m in length. Images: screenshots from Google Maps[TM], © 2013, Geoimage Austria, Flotron/Perrinjaquet.

variability of the digitizing by the same analyst is similar to the multi-analyst variability.

Considering the high variability seen in the interpretation of the high-resolution datasets, we decided to also derive the reference size for three glaciers from Ikonos (indicated in Fig. 3) using multiple digitizations. The overlay of the outlines depicted in Figure 8 illustrates the variability in interpretation and indicates that the creation of a reference dataset (that can be used as the 'truth') is challenging. Consequently, one has to admit that without an appropriate validation dataset the accuracy of other datasets cannot be calculated. However, the mean value of the four digitizations is likely close to the 'truth' and was thus taken as a reference value.

4.2. Statistical analysis

Statistical results for the three test regions are summarized in Table 2 and visualized in Figure 9. For the Alaska test region (QuickBird scene), the mean STD of glacier areas (all glaciers, all analysts) is ~6%. For glacier 5 (0.06 km^2) with

the difficult interpretation (cf. Figs 2 and 4a) the mean STD is much higher (31%); for the very small glacier 1 (0.01 km^2) the STD is 18%, and for the much larger (0.4 km^2) but heavily debris-covered glacier 8 it is 13%. The time taken to digitize all eight glaciers varied between 40 and 200 min. For the glaciers digitized on aerial photography, the STD is 1.6–8.8% (mean 3.6%), with the largest value for the partly debris-covered glacier depicted in Figure 5b. The digitization of all three glaciers took 10–30 min.

The first digitization on the TM scene had STD of 2.7–14.6%, with a mean value of 3.5%. The second and third digitizations had a very similar variability and mean. The digitization of all ten glaciers took 10–50 min. The mean difference of the automatically derived outlines and the reference value (from the manual digitizations) is only −1.3% (range 0.2 to −5.7%). However, this value is biased by the good agreement with the largest glacier (No. 10 in Table 2). When that glacier is excluded, the mean difference is −3.1%, very similar to the STD of the manual digitizations. Compared with the reference dataset from Ikonos (Fig. 8),
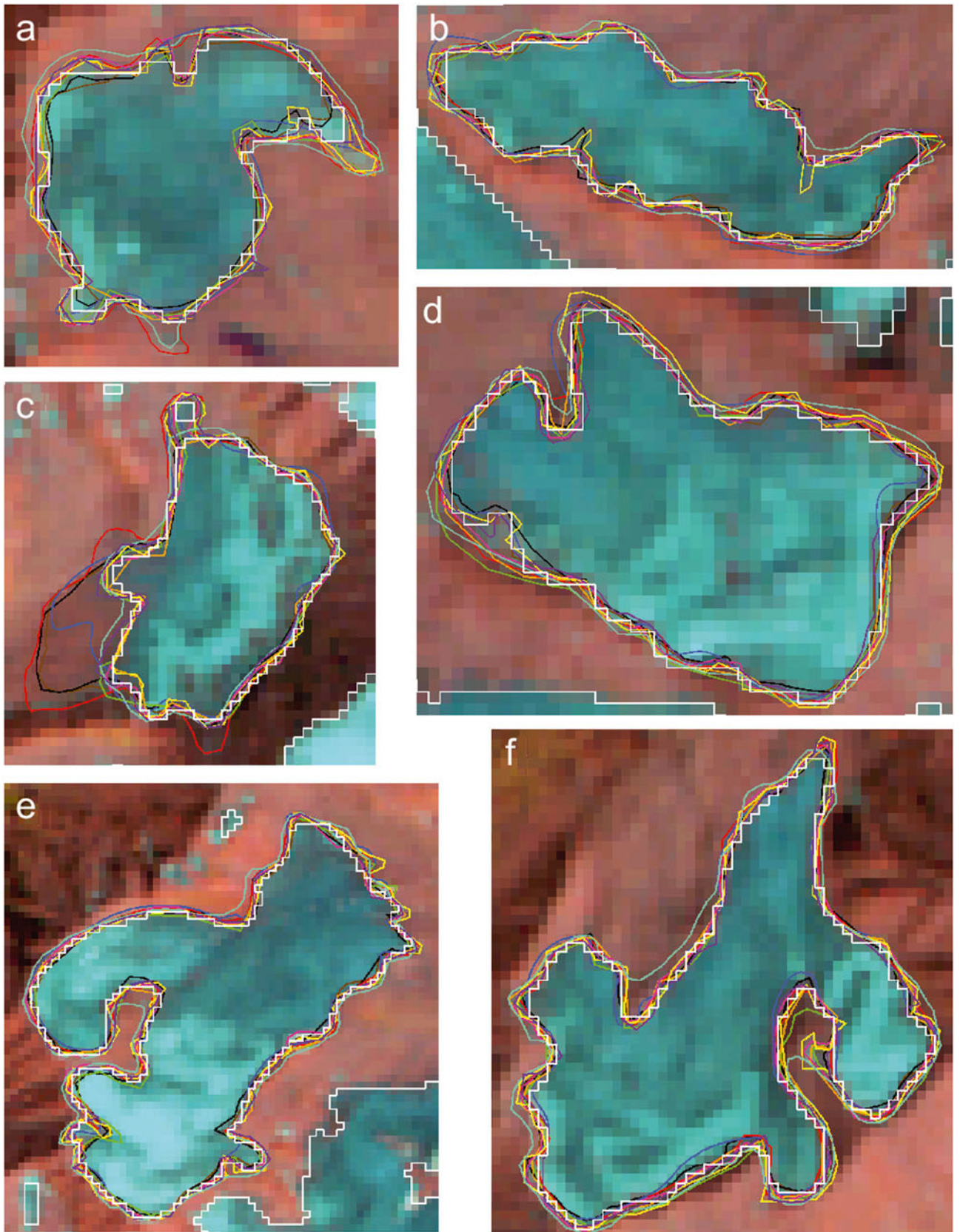
Fig. 6. Six examples from the multiple digitizations of glaciers (bands 543 as RGB) using the TM scene shown in Figure 3 as performed by different analysts (coloured lines). White outlines refer to the automatically derived extents; one image pixel is 30 m in length.

the TM-derived glacier outlines in the Ötztal test region are 4–5% smaller. However, the STD of the multiple digitization of this reference dataset is of the same order of magnitude (up to 3%) so that differences in the interpretation and in the mapping of mixed pixels have about the same effects. Only the tendency to underestimate 'real' glacier area by automated mapping with TM is robust. The scatter plot in Figure 9 shows a decreasing STD with increasing glacier
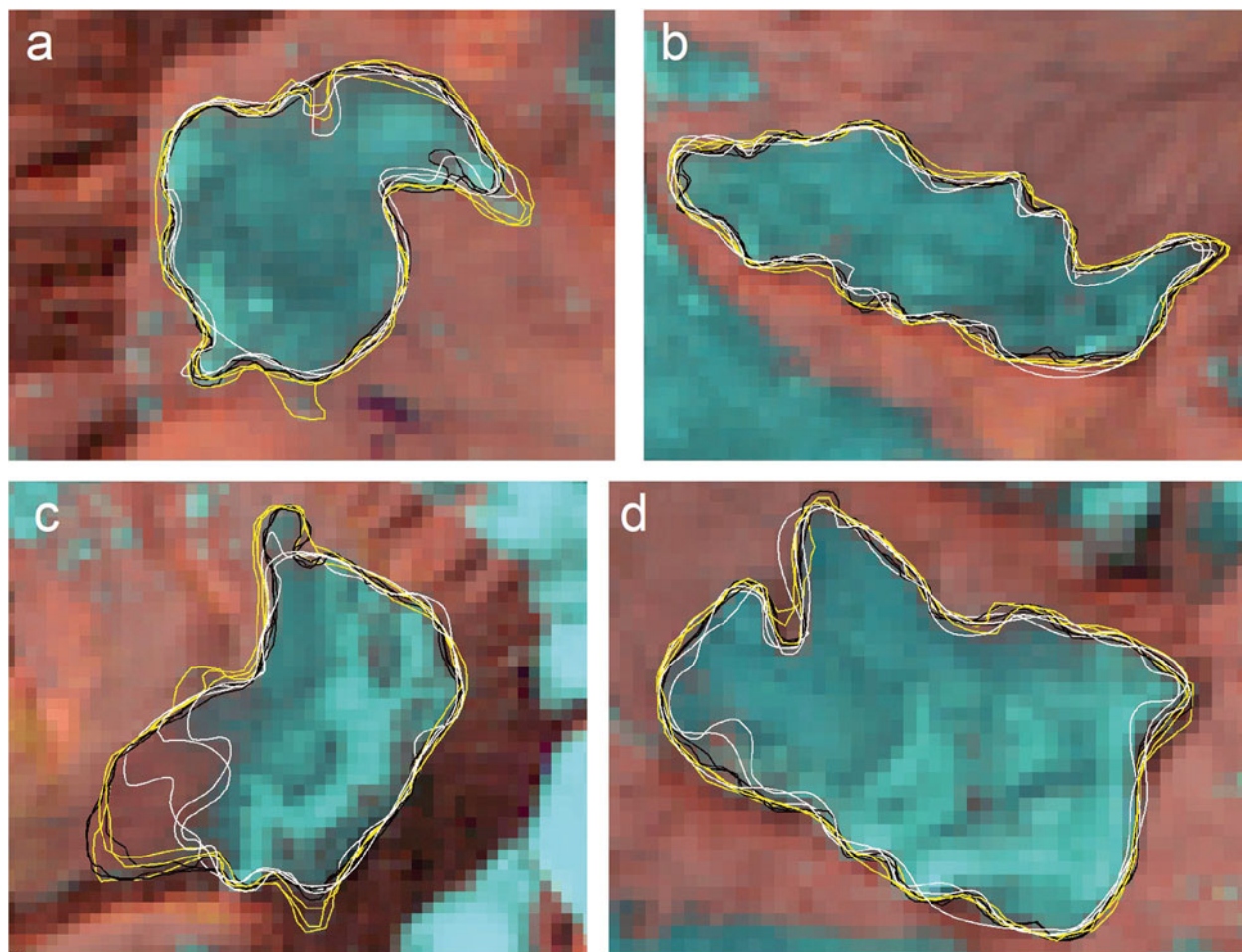
Fig. 7. Overlay of manually digitized glacier extents from three analysts (depicted by yellow, white and black lines) and four glaciers (Fig. 6a–d). For each analyst the three digitizations are shown in the same colour; one image pixel is 30 m in length.
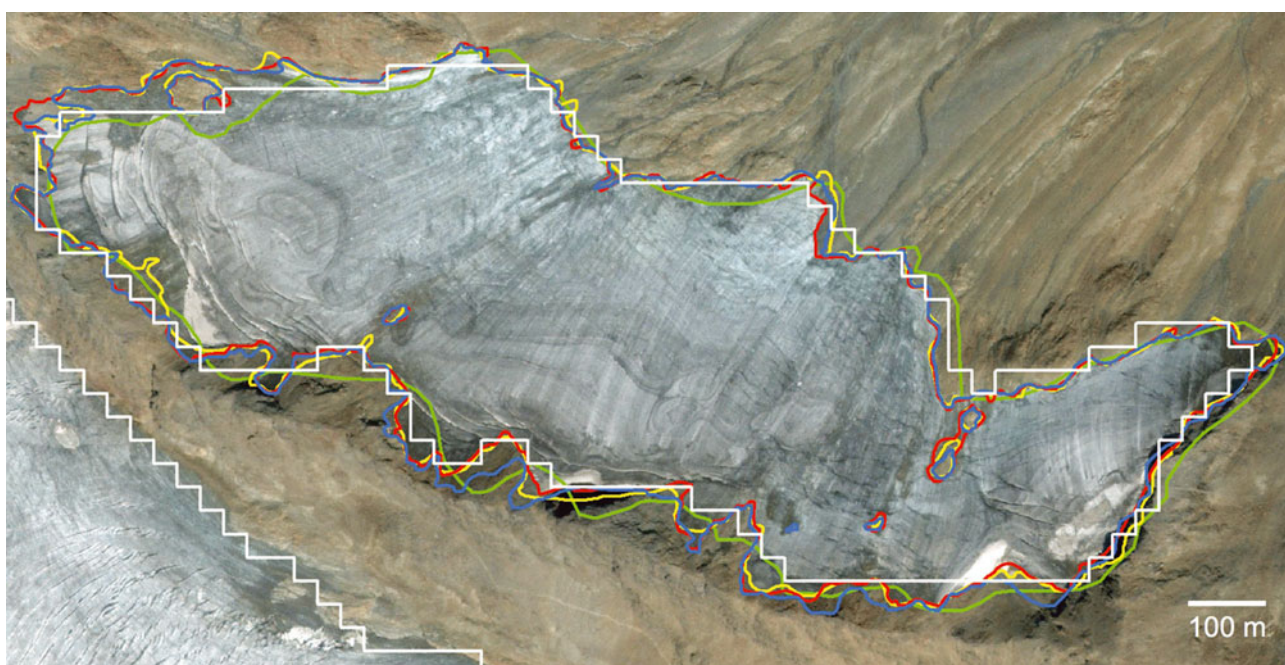

Fig. 8. Overlay of three manually digitized glacier extents (blue, yellow, red) for the upper Guslarferner, Austrian Alps (cf. Fig. 6b and 7b). The white outline is derived automatically from TM; the green line is manually digitized from TM (randomly selected). Image: screenshot from Google Maps[TM], © 2013 European Space Imaging.
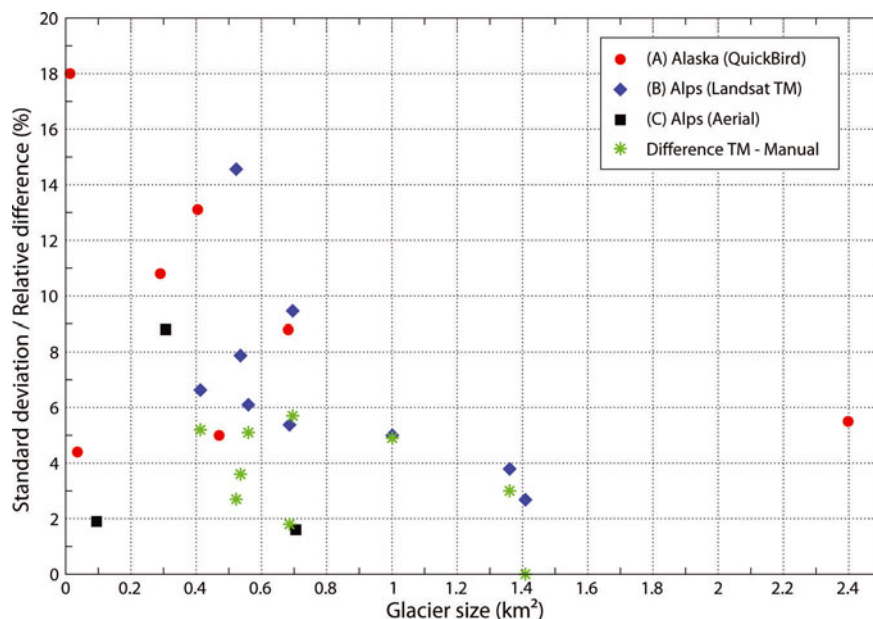
Fig. 9. Glacier size vs standard deviation of the manually digitized area values for seven of the eight glaciers from the test region in Alaska (A), nine of the ten glaciers from the test region in the Ötztal Alps (B) and three glaciers from the test regions in Switzerland (C). The differences of outlines derived automatically from TM to the mean of manual digitizations (Ötztal Alps) are shown as green stars.

Table 2. Comparison of glacier area values as derived for the three test regions by all analysts. For the regions where multiple digitizations were performed (Ötztal Alps), the mean value of the first digitization from each analyst is taken for the comparison. The numbers in italics indicate values that are not used for further analysis, as the dates of the images compared are different. n is sample size, t is time used for the manual digitizing (min), STD is standard deviation and Diff. is difference between the manually and automatically derived area

Region	Glacier	Area			STD	Automated with TM	
		mean	min	max		Area	Diff.
		m²	m²	m²	%	m²	%
	1	13 161	10 343	19 557	18.0	–	–
	2	470 309	391 475	474 862	5.0	–	–
	3	681 970	544 314	772 594	8.8	–	–
Alaska ($n = 13$,	4	35 890	33 795	38 286	4.4	–	–
$t = 40$–200)	5	39 294	26 372	58 448	30.1	–	–
	6	2 397 957	2 196 792	2 588 353	5.5	–	–
	7	289 665	218 184	326 143	10.8	–	–
	8	404 490	344 254	503 332	13.1	–	–
	All	**4 322 091**	**4 057 342**	**4 652 627**	**5.7**	**–**	**–**
	1	413 025	358 260	470 325	6.6	391 500	−5.2
	2	1 423 065	1 328 811	1 556 307	3.8	1 380 600	−3.0
	3	528 487	430 703	659 533	14.6	514 400	−2.7
	4	559 764	511 161	609 834	6.1	531 000	−5.1
Ötztal Alps	5	1 000 940	932 074	1 128 614	5.0	952 200	−4.9
($n = 13$, $t = 40$–	6	686 207	629 529	738 677	5.4	674 100	−1.8
150)	7	535 773	474 604	649 233	7.9	516 600	−3.6
	8	1 409 021	1 326 903	1 470 863	2.7	1 409 337	0.0
	9	695 791	602 448	835 277	9.5	656 100	−5.7
	10	8 918 710	8 558 375	9 397 447	2.9	8 936 010	0.2
	All	**16 170 784**	**15 520 158**	**17 150 047**	**3.4**	**15 961 847**	**−1.3**
	4	546 858	539 848	556 046	1.3	531 000	−2.9
Ötztal Alps ($n = 5$,	8	1 511 081	1 460 441	1 570 260	3.2	1 409 337	−6.7
$t = 60$)	9	663 232	644 903	689 287	2.8	656 100	−1.1
	All	**2 721 171**	**2 656 784**	**2 815 593**	**2.6**	**2 596 437**	**−4.6**
	1	94 683	92 611	97 425	1.9	*106 256*	*12.2*
Switzerland	2	306 672	271 571	357 903	8.8	*284 339*	*−7.3*
($n = 12$, $t = 10$–60)	3	705 369	680 312	720 165	1.6	*754 200*	*6.9*
	All	**1 106 724**	**1 074 279**	**1 155 421**	**3.6**	*1 144 795*	*3.4*

size, but the distribution of values can also be summarized differently: it is <5% for glaciers larger than $1 km^2$; 1–15% (excluding two outliers) for smaller glaciers; and 2–6% difference with the automated mapping. A systematic trend towards larger differences for faster digitizations of outlines was not found.

5. DISCUSSION

The results presented above confirm previous accuracy assessments (e.g. Paul and Kääb, 2005; Bolch and Kamp, 2006; Andreassen and others, 2008; Paul and others, 2011) that reported relative area differences of the automatically derived outlines from manually digitized outlines on higher-resolution datasets between 2% and 5%. These values also apply to the glaciers investigated here with sizes of a few km^2. The relative area differences tend to be larger for even smaller glaciers and smaller for larger glaciers (Paul and others, 2003), at least when they are mostly debris-free. For debris-covered glaciers the accuracy is more a function of the visibility of the debris cover rather than of glacier size.

The problems of interpreting debris cover on glaciers have also been reported previously (e.g. Paul and others, 2002; Racoviteanu and others, 2009; Bolch and others, 2010). That the location of terminus positions can differ by several hundred metres indicates the magnitude of un-certainty when length changes of glaciers are derived from outlines that have been digitized by different analysts. This also applies to outlines that were digitized from topographic maps and derived by cartographers (e.g. Hall and others, 2003; Bhambri and Bolch, 2009; Bolch and others, 2010). Hence, length changes derived from glacier outlines of two epochs must either be large, refer to debris-free glaciers or be assessed by the same analyst to be reliable.

In a qualitative sense, the automatically derived glacier outlines have the advantage of being reproducible, not generalized (i.e. they follow exactly the pixel outline) and rapidly determined. For example, ~2 s per glacier are required for a TM scene with, for example, 600 glaciers and 15–20 min processing time (including threshold selection by visual comparison of the derived outlines in shadow regions). Manual digitization takes on average 3 min per glacier using TM and 7–10 min per glacier using the higher-resolution datasets, which is about 100–300 times slower. Of course, the automatically derived outlines need to be edited and this extra work is already included in the full manual digitization. However, depending on the region, only a small percentage of glaciers might require editing, so automated mapping will still be much faster in these regions.

In a quantitative sense, automated mapping is at least as accurate as manual digitization, as the differences from a reference value are very similar to the precision of the manual digitizations (as expressed by their standard devia-tions). However, glacier area is systematically a few per cent smaller in most cases, hinting at mixed pixels along the glacier perimeter that are omitted in the automated mapping but are considered during manual digitization. Of course, the reference value can also be too large compared to reality, but the comparison with the validation dataset from Ikonos also reveals slightly smaller areas with TM. We thus recommend selecting a threshold for the band ratio as low as possible (i.e. before the number of misclassified individual pixels increases) for the automated techniques to properly include most of the often slightly dirty glacier ice around the perimeter of a glacier, and also carefully checking the boundary of seemingly clean glaciers.

For an improved interpretation of debris-covered glacier parts (beyond what can be achieved from the original image), we recommend checking with potentially available high-resolution remote-sensing data as available in Google Earth™ or similar tools providing access to these data. Even if these images have not been acquired at the same date, they might assist the visual interpretation. Several semi-automated techniques for mapping debris-covered glaciers have been proposed (e.g. Paul and others, 2004; Shukla and others, 2010; Bhambri and others, 2011), but they all require more complex processing, an accurate DEM and final manual editing. Ultimately, coherence images from micro-wave sensors taken in summer seem to provide the possibility of accurately identifying glacier parts under debris cover (e.g. Atwood and others, 2010; Frey and others, 2012; Strozzi and others, 2012). As a more general recommendation we suggest mapping glaciers automatic-ally with one of the well-established methods (e.g. a simple band ratio with a threshold) whenever possible, and restricting manual digitization to the careful correction of errors (e.g. debris cover, shadow).

It is likely not feasible to perform a multiple digitization experiment as illustrated here for each glacier change study using manually or automatically derived glacier outlines from remote-sensing data. However, given the limited availability of reference datasets, the lack of contrast in natural colour images (from high-resolution sensors), and based on the results of this study, we recommend that each study on that topic should select a few (five to ten) differently sized glaciers (with and without debris cover) and digitize them three to five times. The standard deviation of such an assessment provides the analysts' internal precision that can be used as a measure of accuracy in any analysis related to glacier size, including determination of the significance of measured changes. Considering the one-pixel variability of the outline position for clean ice, a rough estimate of the precision can be calculated by buffering the area by one pixel and calculating the relative change in glacier size.

6. CONCLUSIONS

We have presented a comparison of manually and auto-matically derived glacier outlines using high- (~1 m) and medium-resolution (30 m) datasets. Results were provided by all co-authors of this study and are seen as a representative assessment of the common challenges in glacier delineation and the accuracy and precision that can be achieved. Our main conclusion is that automated mapping of clean glacier ice is at least as accurate as manual digitization, but glacier sizes tend to be a few per cent smaller than the reference datasets. Automated mapping has the clear advantages of being much faster, not generalized and generating reprodu-cible results, i.e. the same threshold values always generate the same outlines. Manual digitization should thus focus on the correction of automatically derived outlines to cope with the typically problematic issues such as debris cover or ice in shadow. High-resolution data should be used whenever possible to aid in the interpretation of debris-covered areas and other critical regions. However, when using such high-resolution data as a base for a digitization, the precision of the derived glacier area is not necessarily higher (e.g. due to low contrast or difficult interpretation). As a measure of the

accuracy of manually corrected glacier outlines in the absence of reference data, we recommend performing the multiple digitizing of a couple of glaciers (different size, with and without debris) and calculating the precision from the standard deviation of the digitizations. This measure is also appropriate to assess the significance of any derived relative changes in glacier size.

AUTHOR CONTRIBUTION STATEMENT

F.P. led the study and the writing of the paper. All co-authors contributed their round-robin results and helped in writing the paper.

ACKNOWLEDGEMENTS

REFERENCES

Albert TH (2002) Evaluation of remote sensing techniques for ice-area classification applied to the tropical Quelccaya ice cap, Peru. *Polar Geogr.,* **26**(3), 210–226

Andreassen LM, Paul F, Kääb A and Hausberg JE (2008) Landsat-derived glacier inventory for Jotunheimen, Norway, and deduced glacier changes since the 1930s. *Cryosphere,* **2**(2), 131–145 (doi: 10.5194/tc-2-131-2008)

Atwood DK, Meyer F and Arendt A (2010) Using L-band SAR coherence to delineate glacier extent. *Can. J. Remote Sens.,* **36**(1), 186–195

Berthier E, Le Bris R, Mabileau L, Testut L and Rémy F (2009) Ice wastage on the Kerguelen Islands (49S, 69E) between 1963 and 2006. *J. Geophys. Res.,* **114**(F3), F03005 (doi: 10.1029/2008JF001192)

Bhambri R and Bolch T (2009) Glacier mapping: a review with special reference to the Indian Himalayas. *Progr. Phys. Geogr.,* **33**(5), 672–704

Bhambri R, Bolch T and Chaujar RK (2011) Automated mapping of debris-covered glaciers in the Garhwal Himalayas using ASTER DEMs and multi-spectral data. *Int. J. Remote Sens.,* **32**(23), 8095–8119 (doi: 10.1080/01431161.2010.532821)

Bolch T and Kamp U (2006) Glacier mapping in high mountains using DEMs, Landsat and ASTER data. In Kaufmann V and Sulzer W eds. *Proceedings of the 8th International Symposium on High Mountain Remote Sensing Cartography, 21–27 March 2005, La Paz, Bolivia.* Karl Franzens University, Graz, 13–24 (Grazer Schriften der Geographie und Raumforschung 41)

Bolch T, Menounos B and Wheate R (2010) Landsat-based inventory of glaciers in western Canada, 1985–2005. *Remote Sens. Environ.,* **114**(1), 127–137 (doi: 10.1016/j.rse.2009.08.015)

Frey H, Paul F and Strozzi T (2012) Compilation of a glacier inventory for the western Himalayas from satellite data: methods, challenges and results. *Remote Sens. Environ.,* **124**, 832–843 (doi: 10.1016/j.rse.2012.06.020)

Global Climate Observing System (GCOS) (2003) *The second report on the adequacy of the Global Observing Systems for Climate in support of the UNFCCC, GCOS-82.* World Meteorological Organization, Geneva (WMO/TD No. 1143)

GCOS (2004) *Implementation plan for the Global Observing System for Climate in support of the UNFCCC, GCOS-92.* World Meteorological Organization, Geneva (WMO/TD No. 1219)

Hall DK, Bayr KJ, Schöner W, Bindschadler RA and Chien JYL (2003) Consideration of the errors inherent in mapping historical glacier positions in Austria from ground and space (1893–2001). *Remote Sens. Environ.,* **86**(4), 566–577 (doi: 10.1016/S0034-4257(03)00134-2)

Paul F and Kääb A (2005) Perspectives on the production of a glacier inventory from multispectral satellite data in Arctic Canada: Cumberland Peninsula, Baffin Island. *Ann. Glaciol.,* **42**, 59–66 (doi: 10.3189/172756405781813087)

Paul F, Kääb A, Maisch M, Kellenberger T and Haeberli W (2002) The new remote-sensing-derived Swiss glacier inventory: I. Methods. *Ann. Glaciol.,* **34**, 355–361 (doi: 10.3189/172756402781817941)

Paul F, Huggel C, Kääb A, Kellenberger T and Maisch M (2003) Comparison of TM-derived glacier areas with higher resolution data sets. *EARSeL eProc.,* **2**(1), 15–21

Paul F, Huggel C and Kääb A (2004) Combining satellite multi-spectral image data and a digital elevation model for mapping debris-covered glaciers. *Remote Sens. Environ.,* **89**(4), 510–518 (doi: 10.1016/j.rse.2003.11.007)

Paul F, Frey H and Le Bris R (2011) A new glacier inventory for the European Alps from Landsat TM scenes of 2003: challenges and results. *Ann. Glaciol.,* **52**(59), 144–152 (doi: 10.3189/172756411799096295)

Racoviteanu AE, Paul F, Raup B, Khalsa SJS and Armstrong R (2009) Challenges and recommendations in mapping of glacier parameters from space: results of the 2008 Global Land Ice Measurements from Space (GLIMS) workshop, Boulder, Colorado, USA. *Ann. Glaciol.,* **50**(53), 53–69 (doi: 10.3189/172756410790595804)

Raup B and Khalsa SJS (2007) *GLIMS data analysis tutorial.* http://www.glims.org/MapsAndDocs/assets/GLIMS_Analysis_Tutorial_a4.pdf

Schiefer E, Menounos B and Wheate R (2008) An inventory of morphometric analysis of British Columbia glaciers, Canada. *J. Glaciol.,* **54**(186), 551–560 (doi: 10.3189/002214308785836995)

Shukla A, Arora MK and Gupta RP (2010) Synergistic approach for mapping debris-covered glaciers using optical–thermal remote sensing data with inputs from geomorphometric parameters. *Remote Sens. Environ.,* **114**(7), 1378–1387

Strozzi T, Paul F and Kääb A (2010) Glacier mapping with ALOS PALSAR data within the ESA GlobGlacier project. In *Proceedings of the ESA Living Planet Symposium, 28 June–2 July 2010, Bergen, Norway.* Noordwijk, European Space Agency (SP-686)

Svoboda F and Paul F (2009) A new glacier inventory on southern Baffin Island, Canada, from ASTER data: I. Applied methods, challenges and solutions. *Ann. Glaciol.,* **50**(53), 11–21 (doi: 10.3189/172756410790595912)