Answer to the review of the anonymous referee #1:

General comments:

1. The current manuscript introduces the application of PRACTISE, but does not provide a detailed enough description to be able to fully understand how it works. Considering there are already two detailed papers on PRACTISE by the authors in GMD, I would hope citations to them could provide the reader with a satisfactory description. Then this manuscript could be refined to provide more focus on the results of the RCZ / VF comparison, and greater detail on the scaling question. Figures 2 through 5 could be removed to focus more on the results of this study (Fig 6 through 9) and expand analysis around figure 10 – which is of great interest, but under-analyzed in the current manuscript.

Answer to 1. We thank you for the comment and we will revise the paragraphs on the application of PRACTISE in this study to clarify the workflow used in this study by including more references to our previous papers on PRACTISE. However, we also think that the figures 2 to 5 are important here as they outline the workflow in PRACTISE graphically. This makes it simpler for non-experts to understand the processing steps taken in this specific study and it was a recommendation of the editor to our initial submission. Moreover, with these figures, readers are not obliged to read our previous papers where they probably face much more options and detail than they need. The second part of your comment is proposing to expand the analysis on the scaling effects between different NDSI thresholds. We agree with you on this point (see the answer to comment 2).

Manuscript changes to 1.

p.6, l1 to p.7, l5: "In a first step information about the camera location and orientation was needed for georectification of the photography. This information was automatically optimized by using ground control points (GCPs, Fig. 4a; Sect. 3.3 in Härer et al., 2013). The calculated viewpoint and viewing direction were by default used to perform a viewshed analysis (Fig. 4b; Sect. 3.1 in Härer et al., 2013). The viewshed was needed for an identification of areas which were visible from the viewpoint and which were not obscured by topographical features or within a user-specified buffer area around the camera. The respective DEM pixels were then projected to the photo plane (Fig. 4c; Sect. 3.2 in Härer et al., 2013).

Now, the snow classification module was activated to distinguish between snow covered and snow-free DEM pixels (Fig. 4d). Two major procedures were available for classification: a statistical analysis which was using the blue RGB band (Salvatori et al. 2011; Sect. 3.4 in Härer et al., 2013) and a principal component analysis (PCA) based approach (Sect. 3.1 in Härer et al. 2016). The first was used for shadow-free scenes, the second for scenes with shaded areas. Section 3.4 in Härer et al. (2013) gives more insights into a third manual option if none of the two classification routines could be applied successfully. The photograph snow cover maps did have even in the case that an insufficient classification algorithm was used for a specific situation less than 5% misclassified pixels in the worst case region of the photograph in Chapt. 4 in Härer et al. (2013). It was also shown in an earlier publication that the classification of shadow-affected photographs are of the same quality as photographs without shadows (Chapt. 4 in Härer et al., 2016). As for this study, every classified image

was visually inspected and no major snow classification errors comparable to our worst case example in the previous publication were found, we expect a relative misclassification error of 1%. For this example photograph, the snow classification algorithm utilizing a principal component analysis (PCA) was selected to account for the shadow-affected areas in the upper left part of the photograph (Fig. 4d, enlarged view in Fig. 4e).

After the photograph rectification and classification, the remote sensing routine of PRACTISE began with the identification of satellite pixels that spatially overlap with the photograph snow cover map (Sect. 3.2 in Härer et al., 2016). The used photograph and satellite image were thereby recorded within one (RCZ) to three (VF) hours. Moreover, a cloud- and shadow-free satellite image is generated by using fmask (Zhu et al., 2015). The needed NDSI map was calculated in accordance to Eq. (1) by PRACTISE (Fig. 5a).

If both, the NDSI satellite map and the corresponding high resolution photograph snow cover map were processed, an iterative calibration of the NDSI threshold value was started. The Landsat image was thereby resampled to the finer resolution of the photograph in the calibration to avoid losing any information by the aggregation of the photograph snow cover map. The best agreement between the local scale (photograph) and the large scale (Landsat) snow cover map was detected by maximizing the accuracy which is the ratio of identically classified pixels to the overall number of photograph-satellite image pixel pairs *n* (Aronica et al., 2002):

$$F = \frac{(a+ d)}{n}, \quad (2)$$

a represents the number of correctly identified snow pixels and *d* the same for no snow pixels. *F* is between 0 and 1 and becomes 1 for a perfect agreement between the two images.

The calibrated NDSI threshold was finally applied to the Landsat data with 30m pixel size to generate the final Landsat snow cover map. Figure 5b shows the resulting satellite snow cover map superimposed on the photograph snow cover map and a Landsat Look image. A cutout is shown for more detail in Fig. 5c. ”

2. The comparison between RCZ / VF is robust and well quantified (Fig 7). The influence of rock reflectance is well described (Fig 8) and provides a valuable process basis to the quadratic relationships (Fig 9). However, the scaling question, while well illustrated by the example of 16 September (Fig 10a), suggests that all data have been used to create relationships presented in (Fig 10c). How was Fig 10c constructed and what does the ‘cumulated probability’ mean in this context? Can you show that the increase in the identical nature of snow cover maps is not simply a function of decreasing number of snow maps pairs (again this links to greater clarification as to what is meant by cumulated probability). If this relationship is statistically robust, could more be made of this message, as understanding the influence of measurement resolution by satellite imagery is important to understanding the fate of snow and ice in small glacierized basins.

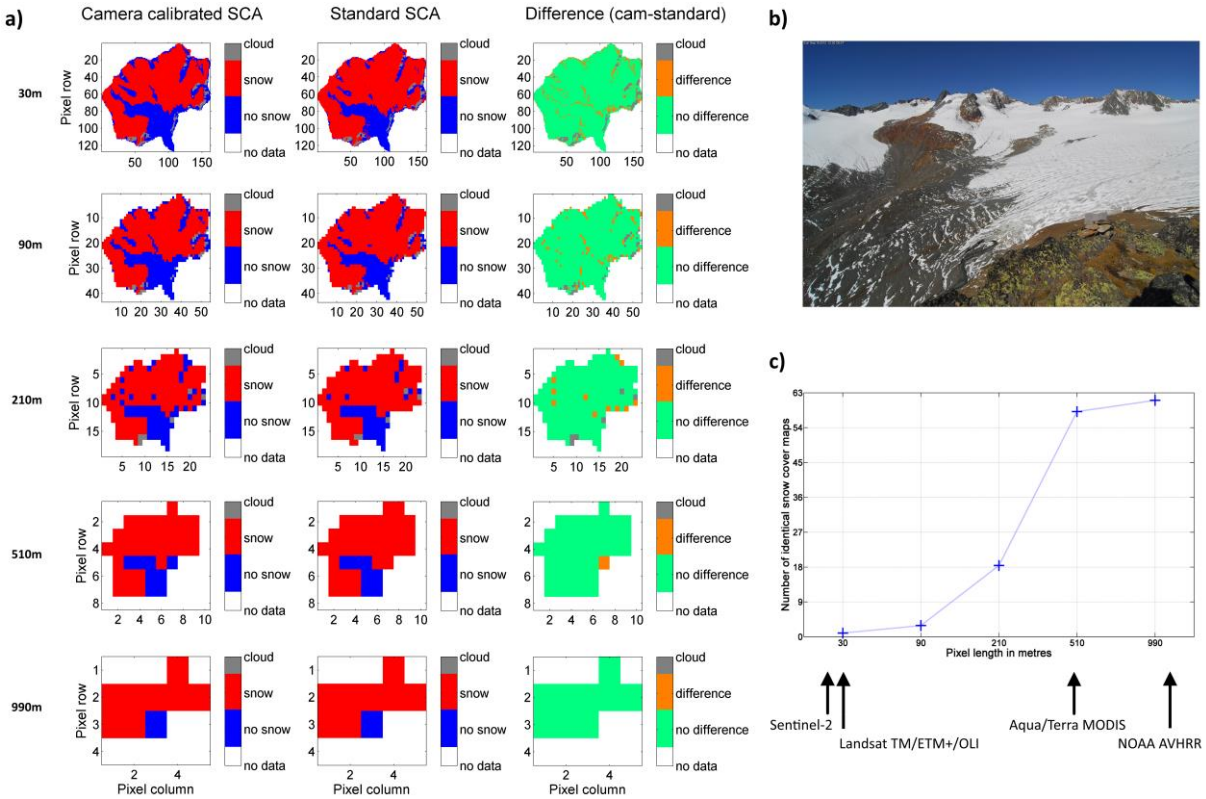
Answer to 2. We really appreciate that you value our work presented using the figures 7 to 10. With respect to figure 10c, you are firstly right that all analysed data at Vernagtferner and Zugspitze is used to generate the graph. The term ‘cumulated probability’ in figure 10c might be unclear, we will change it to the ‘number of identical snow cover maps’ where the total number is 63. The figure hence answers

the question how many of the snow maps generated with the calibrated and the standard threshold are identical in the complete catchment areas for the different pixel sizes between 30 m and 990 m. For example, 58 of the 63 snow maps (over 90%) are completely identical at a pixel size of 510 m.

Now, we come to the final part of your question where you asked if the identical nature of snow cover maps is not simply a function of decreasing number of snow map pairs and if this relationship is statistically robust. We can neither agree nor deny this statement from our data. We thought about this effect and our data shows often increase but also at other dates decrease in agreement for coarser resolutions. The simple reason is that one pixel being different has a stronger relative effect with coarser resolution as our catchments are small and the number of pixels becomes low. We therefore decided to not look at the relative increase in agreement of the calibrated and standard snow cover maps with larger pixel sizes but to focus on the pixel size at which the snow cover maps become completely identical as this is independent of relative changes with aggregation. Fig. 10 c outlines when total agreement of the snow cover maps for the different NDSI thresholds was found.

Manuscript changes to 2.

Figure 10



p.11, l1 to 16: "Our aggregation experiment of the Landsat snow cover maps for the different $NDSI$ thresholds shows that the SCA deviation between standard and calibrated snow cover maps diminishes for coarser resolution data in most cases. Figure 10 a outlines this error reduction with spatial aggregation for a Landsat 7 scene of Vernagtferner catchment on 16 September 2011. Figure 10 b shows the simultaneously captured photograph used for calibration. We however cannot draw an absolute conclusion from fig. 10 a that the difference in snow cover maps between the different thresholds is always reduced with a coarser resolution. The simple reason is that with larger pixel sizes,

the number of pixels in the catchment becomes lower and the relative weight of a pixel being different for different thresholds has a larger relative weight. Therefore, we decided to investigate at which spatial resolution the standard and calibrated snow cover maps become identical for the 63 cases investigated in the two catchments. This variable is absolute and thus independent of relative weights and changes with spatial aggregation. The aggregation step to 510m is thereby of major importance as more than 90% (58 of 63) of SCA maps become identical at this pixel size. Thus, using the standard threshold of 0.4 instead of the higher $NDSI$ thresholds at VF and the lower $NDSI$ values at RCZ seems to be accurate in most cases with a pixel size of 500m. For applications at this scale, the positive effect of using camera calibrated data diminishes and might rarely justify the effort."

Specific comments:

Note: We do not show each of the manuscript changes for the specific comments here as the changes are obvious by the answer. Nonetheless, the changes will be denoted in the new manuscript.

Pg1, ln19: Quantify how different the statistically insignificant correlation was to the standard threshold.

Answer: We add the correlation coefficient here.

Pg1, ln 20: what is the 'another literature value'? State it here.

Answer: The other literature value is the locally optimized 0.7 threshold value of Maher et al. (2012) which was also found for single events at Vernagtferner. We add this to the abstract.

Pg1, ln 21: replace 'case' with 'cases where'.

Answer: Thank you for the correction.

Pg2, ln 5: 'precipitation water' – just say 'precipitation'?

Answer:, We delete the word 'water'.

Pg2, ln 10-13: avoid single sentence paragraphs. Change this throughout the manuscript.

Answer: You are right, we change the single sentence paragraphs in our manuscript.

Pg2, ln28: 'In this context' is superfluous and could be removed.

Answer: We remove it, thank you.

Pg3, Ln 23: 'built up' is poor terminology for geological composition

Answer: We agree and change the terminology.

Pg3, ln 25: ‘pending’ is strange terminology. Do you mean ‘underlying rock’ or ‘substrate’

Answer: Thank you, we use ‘underlying rock’ now.

Pg3, ln 29: no need for ‘for’ in the statement ‘guarantees for comparable’

Answer: Thank you for the correction.

Pg4, ln 1: do you mean ‘dates’ rather than ‘cases’? Stick to constant terminology.

Answer: We totally agree, we change it.

Pg4, ln 4: is ‘rectifaciton’ as spelling error?

Answer: Yes, you are right, it should be ‘rectification’.

Pg5, ln 1: no need for ‘It has to be mentioned that’

Answer: We agree and remove it accordingly.

Pg5, ln 22: – remove ‘an’

Answer: Thank you for the correction.

Pg5, ln 30 & 32: remove ‘used’

Answer: We remove ‘used’ in both sentences.

Pg5, ln 32: ‘misclassificied’ is misspelt

Answer: Thank you for finding this spelling error.

P6, ln 4-6: tenses are used interchangeably. Suggest sticking to past tense consistently throughout the methods section

Answer: You are right, we change the tense in the methods section to past tense where appropriate.

Pg6, ln 9: remove ‘thereby’

Answer: We correct it.

P6, ln 13: no need for ‘It has to be mentioned that’

Answer: You are right, we remove it.

Pg7, ln 10: what does ‘underline’ mean in this context, I think the wrong word is being used here.

Answer: This is true, we wanted to clarify that the minimal differences between the Otsu method and the standard threshold does not justify the additional effort needed for the Otsu method. We use ‘justify’ now.”

Pg7, ln 20: – if statistically significant, then present the stats here (r-value and p-value).

Answer: We change ‘significantly weaker’ to ‘weaker’ as the expression was not meant in a statistically quantitative way here. It is a qualitative statement.

Pg8. Ln 17: – remove ‘in percents’

Answer: Thank you for the correction.

Pg8, ln24: – what chose a threshold of 0.7? Provide some justification.

Answer: We have chosen the 0.7 threshold as it is in the range of plausible NDSI threshold values (0.35 to 0.7) that might result from a single date calibration at VF. Moreover, Maher et al. stated this value in their study area when they also calibrated the NDSI threshold only for one date due to the lack of additional data. We simply want to show here that a NDSI threshold calibrated at a single date does not give too much insight how the NDSI threshold might look like at other dates but at the same time gives investigators a false sense of confidence. We clarify it.

Pg10, ln 11: – what is a ‘date by date transfer’?

Answer: The term was misleading. We wanted to state that we tested if the NDSI threshold calibrated at one catchment can be used at another catchment in the same Landsat scene. We rephrase the sentence.

Pg10, ln 23: – ‘jeopardous’ is probably not the correct term to use here. ‘inappropriate’ or something similar may be better.

Answer: Thank for the suggestion, we change it.

Figures: Use of titles within figures and sub-figures is unnecessary (e.g. above each sub-figure in Fig 7). Instead use the associated caption to clearly describe each figure. Often current captions are a mix of methods and results, rather than sticking to the bare minimum need to adequately describe what is presented. The main body of the text should instead be used to explain methodological procedures and results.

Answer: Thank you, we will change the figure captions and the respective text in the main body of the text.

Answer to the review of S. Cascoin

General comments:

1) The authors choose not to apply atmospheric correction to the Landsat scenes. Absorption and scattering by the air molecules and aerosols can have an impact on the NDSI, although atmospheric absorbance is typically low in the SWIR and Green wavelengths, which are used to compute the NDSI (especially in high elevation areas). However, this shall be better discussed since the authors justify their study by the fact that the 0.4 NDSI threshold is used for the MODIS snow products ("This is of special interest as MODIS snow cover products are today the most frequently applied satellite snow cover maps"). If the authors referred to NASA's MOD10/MYD10 snow products, the atmospheric correction is applied before computing the NDSI. In addition, the NDSI threshold is not applied anymore in the latest collection 6. More importantly, the lack of atmospheric correction cast doubts on the significance and the transferability to other sites of the "newly developed calibrated quadratic polynomial model which is accounting for seasonal threshold dynamics". The authors should clarify this to avoid confusing readers who are not familiar with satellite imagery processing.

Answer to 1) Thank you for this useful comment. You are right that we did not state the atmospheric correction of the MODIS snow cover product (MOD10/MYD10). And it is also true that the recently updated MODIS snow cover product collection does not use the fixed threshold of 0.4 anymore but uses a flag system in combination with a NDSI value of 0 as threshold. However, an own NDSI threshold can be used. Hence, the new algorithm gives more freedom to the users. This will probably lead to the situation that many users that cannot assess the value of the flag system or the best NDSI threshold for their scene might simply use the standard 0.4 threshold value again. In any case, we will add this information in the introduction to clarify the new situation for MODIS snow products to users unexperienced in this field of research as well as that we agree to add the statement to the manuscript that the developed quadratic approach might be only transferable to other high elevation areas at the moment and that further tests and probably an atmospheric correction of the data are needed if an application in lowland areas is planned.

Manuscript changes to 1)

p.2, l.29 to 34: "Accuracies in this range even though for the atmospherically corrected MODIS snow cover product (MOD10/MYD10) make $NDSI$ based snow cover products well accepted for global scale applications, but uncertainties have to be expected at the local scale (Härer et al. 2016). Moreover, the snow detection algorithm for the MODIS snow cover product changed in the latest collection 6. The algorithm now uses a NDSI threshold of zero together with a flag system to detect snow cover and users are encouraged to use their own NDSI threshold in the MODIS Snow Products Collection 6 User Guide if a binary snow cover map is wanted."

p.10, l.29 to 32: “However, the detected $NDSI$ threshold dependency automatically leads to the question if the need for threshold adaption is also necessary for coarser resolution satellite snow cover maps, for example, for a spatial resolution of 500 m or 1 km.”
p.10, l.23 to 27: “This assumption and the transferability of the model is probably only true for high elevation areas. Even though that the $NDSI$ is an index which reduces the dependence on atmospheric conditions, an atmospheric correction might be necessary as well as more dynamic approaches that reflect the vegetation growth and senescence over the year for lowland areas. Hence, the approach needs to be further evaluated and developed in future studies with more test sites.”

2) The authors give too much details about the PRACTISE software, which was used to rectify the photographs from the time lapse cameras, whereas it was already described in another journal (for instance Fig. 2 was already published in Härer et al. 2016; Fig. 3 and Fig. 4 further illustrate the PRACTISE workflow and are not useful in my opinion). An important step for this study is rather how these camera snow maps were resampled to the Landsat resolution and it is missing. Indeed, camera snow maps have a submeter ground sampling distance. As a result, it is likely that some Landsat pixels were classified as "snow" from the camera images, while they were actually not 100% snow covered in the camera images.

Answer to 2) We were encouraged to extend the details about the PRACTISE software and its application in its study by the editor before the discussion was opened. And we thank the editor now for this recommendation as we see the benefit that users interested in the approach but not in each detail of the algorithm of PRACTISE can easily follow the processing steps needed to calibrate the NDSI threshold of a photograph with this description. You are however right that we should add the resampling strategy used for the different spatial resolutions of the georectified photographs (1 and 5m) and of the Landsat satellite image (30 m). Moreover, we added more citations to our earlier publications on PRACTISE to clarify the workflow (see comment1 of the anonymous reviewer). To avoid losing any information, we used the finer resolution for calibration by resampling the Landsat resolution to the photograph resolution. The calibrated NDSI threshold is then finally applied to the Landsat pixels at their original resolution of 30m to generate the Landsat snow cover map which indeed will have mixed pixels.

Manuscript changes to 2)

p.6, l.27 to p.7, l4: “The Landsat image was thereby resampled to the finer resolution of the photograph in the calibration to avoid losing any information by the aggregation of the photograph snow cover map. The best agreement between the local scale (photograph) and the large scale (Landsat) snow cover map was detected by maximizing the accuracy which is the ratio of identically classified pixels to the overall number of photograph-satellite image pixel pairs *n* (Aronica et al., 2002):

$$F = \frac{(a+ d)}{n}, \quad (2)$$

a represents the number of correctly identified snow pixels and *d* the same for no snow pixels. *F* is between 0 and 1 and becomes 1 for a perfect agreement between the two images.

The calibrated $NDSI$ threshold was finally applied to the original Landsat data with 30m pixel size to generate the final Landsat snow cover map."

3) The literature review in the introduction was a bit overlooked. The authors state that "The used snow-cover mapping approaches can be grouped into three categories: manual interpretation, classification-based, and index-based methods" but there are other approaches based on spectral unmixing. The proposed "geology dependent offset" is the result of a well-known phenomena (e.g. Kaufman et al. 2002, GRL), and is similar to the NDSI_0 method developed by Chaponnière et al. (2005, IJRS); it can be also seen as an extreme simplification of a spectral mixture analysis used in other MODIS snow products (e.g., Sirguey et al., 2009 RSE; Painter et al., 2009 RSE).

Answer to 3) We agree that our literature review in the introduction benefits from the references that you proposed. We thus include them in the revised manuscript.

Manuscript changes to 3)

p.2, l.13 to 19: "The used snow-cover mapping approaches can be grouped into four categories: manual interpretation, classification-based and index-based methods, and spectral mixture analysis. Manual interpretation as well as classification-based approaches are often used in local snow cover mapping studies. Both are out of the scope of this study as a need for expert knowledge and a high time-demand limit their applicability for large time series data. Spectral Mixture Analysis are also not in the focus of this study as they need an extensive spectral database for the different land surface components (Sirguey et al., 2009; Painter et al., 2009). These databases are usually not commonly available and only the final snow cover product can be downloaded (TMSCAG for Landsat and MODSCAG for MODIS)."

p.3, l.23 to 25: "Moreover, we present a seasonal model calibrated with the $NDSI$ threshold time series. The quadratic polynomial model can also be locally adapted by including a geology dependent offset which is comparable to earlier findings of Chaponnière et al. (2005)."

4) Overlap with Härer et al. (2016 GMD). In a previous paper, the authors already showed the results of the NDSI threshold calibration on three Landsat scenes using the same method. Here, the authors extend this approach to a time series of Landsat images, which is a good idea I think. The authors obtain a (weak) seasonal cycle in the calibrated NDSI threshold value. Given that an important insight of this TCD paper was already introduced by Härer et al. 2016 ("A spatial and temporal adjustment of NDSI thresholds is therefore important to ensure optimum results in the snow cover mapping"), I do not think that the "investigation of the reasons of this effect is beyond this study". The authors could test if the calibrated NDSI_tr is correlated to the solar zenith and azimuth angles. In addition, the authors did not consider the hypothesis that this seasonal cycle may be due to inaccurate snow detection in the camera images. What is the bit depth of the camera images? Snow detection from terrestrial camera imagery is difficult in shaded slopes especially from 8-bits RGB pictures. The reported accuracy (below 5% misclassified pixels from visual inspection) can lead to significant changes in the NDSI, which are probably within the range of the calibrated NDSI threshold variability? This could

be tested by excluding shaded areas before computing the accuracy of the Landsat snow masks. Another source of error that was not discussed is the one due to the geometric distortion between oblique images and nadir-looking satellite images.

Answer to 4) Thank you for this comment. We want to mention here that we used two cameras but only at a single catchment and 3 dates in our case study publication from 2016. We therefore were not sure what the spatial representativity of the calibrated NDSI threshold is within the same Landsat scene. Another question additional to the time series analysis that you mention was if we could use a single location for calibration and then use it for this Landsat scene? The study is therefore also completely new as a second catchment within the same Landsat scene is used for testing the spatial representativity within the scene. The systematic offset that was found, analysed and interpreted is thus also a major finding of the publication. Your thoughts on the reasons for variability are the same that we have. We also want to know what is driving the variability. However, this opens a really huge field of options that could be tested (e.g. albedo, snow grain size, snow age, …) which might need an extended experimental setup and testing all of these options would fill a complete publication on its own. So, it will be a task of our future work. We nevertheless agree that a correlation test for solar zenith and azimuth angles might be helpful here as we see this weak seasonal behaviour. We therefore include it.

The second part of your comment aims at the uncertainty existing in the photograph snow cover maps. You are right, shadows have been a problem when using RGB photography. Therefore, we tackled this issue by developing our method for shadow-affected 8-bit photographs, presented in Härer et al. (2016). We have shown in Härer et al. (2016) that the classifications using this approach has the same quality as the classification using the standard method for sunny photographs without shadows. And we carefully checked all images to ensure that the quality is as high as possible for each photograph. We mention here the highly conservative estimate of 5%. We checked each camera image visually and the value of 5% is the absolute maximum of error that we could think of in one of our classified images (Chapt. 4 in Härer et al., 2013). Usually, the classification error is below 1% if no major classification errors are obvious and thus not an issue. However, we will add these statements for clarification.

Manuscript changes to 4)

Removed p.10, l1 to l3: “This temporal development is potentially related to the sun angle, snow age, grain size or albedo development or other effects. A detailed investigation of the reasons of this effect is beyond this study but will be subject of future studies.”

p.10, l6 to 12: “These results are promising and it needs to be investigated if the seasonal behaviour of the calibrated $NDSI$ thresholds can be attributed to the elevation and azimuth angles of the sun. The correlation r between azimuth angle and $NDSI$ is 0.75 for RCZ and 0.42 for VF. For sun elevation, r is 0.77 for RCZ and 0.54 for VF. The sun angles thus are correlated to the seasonal development but do not fully explain the behaviour. The temporal development is thus potentially also related to snow age, grain size, albedo development or other effects. These might also explain the observed variability

within the seasons. A detailed investigation of this variability is however beyond this study but will be subject of future studies."
p.4, l4 to 5: "The photographs are recorded as 8-bit data with three colour channels (red, green and blue; RGB) on an hourly basis for RCZ and three times a day for VF."
p.6, l12 to 18: "The photograph snow cover maps did have even in the case that an insufficient classification algorithm was used for a specific situation less than 5% misclassified pixels in the worst case region of the photograph in Chapt. 4 in Härer et al. (2013). It was also shown that the classification of shadow-affected photographs are of the same quality as sunny photographs (Chapt. 4 in Härer et al., 2016). As for this study, every classified image was visually inspected and no major snow classification errors comparable to our worst case example in the previous publication were found, we expect a relative misclassification error of 1%."

Specific comments:

Note: We do not show each of the manuscript changes for the specific comments here as the changes are obvious by the answer. Nonetheless, the changes will be denoted in the new manuscript.

P1L12: Earth not earth

Answer: Thank you for the correction, we will revise it.

P3L22: glacierized not glaciered

Answer: You are right, we will change it.

P3L23-25: It could be useful to show the spectral profiles of the snow-free substratum (limestone is more reflective in the visible range than gneiss).

Answer: We think that adding a figure is a bit too much here as only two bands are interesting for the NDSI, however we will add a paragraph describing the general spectral behaviour of limestone and gneiss with respect to the NDSI calculation in the results and discussion section to explain the different mean NDSI values.

P4L13: I do not think that this statement is true "no atmospheric correction is applied (..) the majority of studies that apply the NDSI for snow cover mapping". Many studies use the MOD10 snow products, or TMSCAG for Landsat, which use surface reflectances.

Answer: We will rephrase this sentence as there are also many studies that use atmospheric correction. We will clarify this.

P3L17: NDSI and NDSI_thr are written in equation mode, sometimes in plain text.

Answer: Thank you, we will change it.

P4L9: photographs not photographies. What is the acquisition time of the camera? is it synchronous to Landsat overpass time?

Answer: The photographs at RCZ are taken on an hourly basis and at VF in the morning, at noon and in the afternoon and thus the Landsat image is calibrated with a photograph that is recorded within the same hour at RCZ and within three hours at VF. We will add a statement for clarification.

P4L31: did you find a difference in the results between Landsat 8 and the Landsat 5/7? Landsat 8 instrument has higher radiometric resolution which improves snow classification in mountains (less saturation, higher SNR in shaded slopes).

Answer: We also had this thought at the beginning. However, we would need more acquisitions for this investigation. And the strong variability as described in Table 1 and Figure 7b superimposed on the seasonal threshold behaviour probably hides the signal between different sensor systems.

P6L8: Note that this metric is usually referred to as accuracy and may not be a robust performance measure when the number of a class is much greater than the number of the other one.

Answer: Thank you, we will use the term 'accuracy'. In general, each performance measure has a weakness. In our case however, both investigated catchments are partially glacierized. A minimum area of snow or ice is thus left in summer. It thus should not be a major problem in our catchments. Moreover, a prerequisite for our calibration method is that there are snow covered areas as well as areas free of snow in the photograph and thus also in the Landsat scene (p.4, l.13).

P7L29: if vertical, these rock faces are not visible in images captured by nadir-looking sensors like Landsat 7.

Answer: Thank you for bringing up this mistake. We are investigating summer dates here. Low and flat areas are also snow-free in this time of the year, the sentence was deleted.

P7L31: "NDSI reflectances" does not make sense

Answer: You are right, we will correct this in the complete manuscript.

P8L10-12: this sentence is not clear to me.

Answer: We will clarify the sentence. The 'uncertainty' term is maybe again inexact. The percentages outline the differences in snow cover between the standard and the calibrated threshold.

P8L31: fitted against what? Day of year I think.

Answer: Yes, this is true. We will clarify this.

P9L18: why not using real MODIS images instead?

Answer: We focus on Landsat and the scale of 30m in our study. The aggregation up to 990m is only an experimental setup to outline if the calibration of NDSI thresholds is needed for larger pixel sizes than 30m and we want to analyse different scales here. It is thus not the objective to evaluate a single snow cover product like MOD10/MYD10 which in addition is atmospherically corrected and does not use the 0.4 threshold anymore (see your general comment 1). Moreover, our small catchments are not the best experimental setup for MODIS.

P9L25: if I understand well, the resampling to 500m has increased the optimal threshold value. Can you think of an explanation?

Answer: Sorry, there is a misunderstanding. We simply say here that the NDSI threshold of 0.4 seems to be a really good estimate at a pixel size of 500m. And the NDSI threshold increases for RCZ and decreases for VF so we do not have any trend here. We add a sentence to underline this.

Fig. 5: rainbow colormaps are not recommanded (Borland 2007). I am also surprised by the choice of the projection (plate carrée? non-equidistant projections are not recommended for this kind of maps)

Answer: These figures are equidistant. It is the standard Matlab output using latitude and longitude in m. We will clarify this by adding the unit and we will also follow your suggestion to adapt the colormap.

On the need of a time and location dependent estimation of the NDSI threshold value for reducing existing uncertainties in snow cover maps at different scales

Stefan Härer[1], Matthias Bernhardt[1], Matthias Siebers[2], Karsten Schulz[1]

[1]Institute for Water Management, Hydrology and Hydraulic Engineering (IWHW), University for life sciences (BOKU), 1190 Vienna, Austria
[2]Commision for Glaciology, Bavarian Academy of Sciences and Humanities, 80539 Munich, Germany

Correspondence to: Matthias Bernhardt (matthias.bernhardt@boku.ac.at)

Abstract. Knowledge about the current snow cover extent is essential for characteri~~z~~sing energy and moisture fluxes at the ~~earth~~ Earth surface. The snow~~-~~ covered area (SCA) is often estimated by using optical satellite information in combination with the normalized-difference snow index ($NDSI$~~NDSI~~) .The $NDSI$~~NDSI~~ thereby uses a threshold for the definition if a satellite pixel is assumed to be snow covered or snow free. The spatio-temporal representativeness of the standard threshold of 0.4 is however questionable at the local scale. Here, we use local snow cover maps derived from ground-based photography to continuously calibrate the $NDSI$~~NDSI~~ threshold values ($NDSI_{thr}$) of Landsat satellite images at two European mountain sites of the period from 2010 to 2015. Both sites, the Research Catchment Zugspitzplatt (RCZ, Germany) and the Vernagtferner area (VF, Austria), are located within a single Landsat scene. Nevertheless, the long-term analysis of the $NDSI_{thr}$ demonstrated that the $NDSI_{thr}$ at these sites are not correlated (r=0.17) and different to the standard threshold of 0.4. For further comparison, a dynamic and locally optimized $NDSI$~~NDSI~~ threshold was used as well as another locally optimized literature threshold value (0.7). It was shown that large uncertainties in the prediction of the SCA of up to 24.1% exist in satellite snow cover maps in cases where the standard threshold of 0.4 is used, but a newly developed calibrated quadratic polynomial model which is accounting for seasonal threshold dynamics can reduce this error. The model ~~-~~minimizes the SCA uncertainties at the calibration site VF by 50% in the evaluation period and was also able to improve the results at RCZ in a significant way. Additionally, a scaling experiment has shown that the positive effect of a locally adapted threshold diminishes from a pixel size of 500m and more which underlines the general applicability of the standard threshold at larger scales.

1 Introduction

Numerous studies ranging from the local to the global scale have underlined the influence of snow cover on e.g. air temperature, runoff generation, soil temperature and soil moisture (Bernhardt et al., 2012; Deb et al., 2015; Dutra et al., 2012; Dyurgerov, 2003; Liston, 2004; Mankin and Diffenbaugh, 2015; Santini and di Paola, 2015; Tennant et al., 2015). Hence, an accurate

estimation of the spatial extent of the snow pack is fundamental for a suite of applications (Pomeroy et al., 2015). The accuracy of weather and climate models heavily depends on this information, as the range of surface temperatures is instantly limited to a maximum of 0° C in existence of snow and the surface albedo becomes typically significantly enhanced (Agosta et al., 2015; Liston, 2004; Rangwala et al., 2010; Takata et al., 2003; Vavrus et al., 2011). From a hydrological point of view, the formation of a snow pack has a buffering effect and thus often leads to a transfer of precipitation ~~water~~ from the cold to the warm season of the year (Bernhardt et al., 2014; Viviroli et al., 2011). This leads to a support of summer runoff needed e.g. in agriculture or for sanitary water supply, but can also lead to an intensification of flood events e.g. in case of rain on snow events (Viviroli et al., 2011). With this in mind, information on the current snow distribution is elementary for water resources management (Thirel et al., 2013) and weather forecasting model systems (Dee et al. 2011).

Snow cover distribution is often derived from satellite data and then either used as input for operational models (Butt and Bilal, 2011; Dee et al., 2011; Homan et al., 2011; Tekeli et al., 2005) or for the offline evaluation of modelled snow cover (Bernhardt and Schulz, 2010; Warscher et al., 2013) and snow fall patterns (Maussion et al., 2011).

The used snow-cover mapping approaches can be grouped into ~~three~~ four categories: manual interpretation, classification-based~~,~~ and index-based methods, and spectral mixture analysis. Manual ~~classification~~ interpretation as well as classification-based approaches~~,~~ are often used in local snow cover mapping studies. Both are out of the scope of this study as a need for expert knowledge and a high time-demand limit their applicability for large time series data. Spectral Mixture Analysis are also not in the focus of this study as they need an extensive spectral database for the different land surface components (Sirguey et al., 2009; Painter et al., 2009). These databases are usually not commonly available and only the final snow cover product can be downloaded (TMSCAG for Landsat and MODSCAG for MODIS). ~~Hence, w~~We focus on the automatic normalized-difference snow index ($NDSI$~~NDSI~~) approach here. It was developed by Dozier (1989) and is a simple and established method to identify snow cover in optical satellite images. NOAA/NESDIS which is assimilated into ERA/Interim (Dee et al., 2011; Drusch et al., 2004), or the widely used MODIS snow cover products (Hall and Riggs, 2007; Hall et al., 2002) make use of the $NDSI$~~NDSI~~.

The $NDSI$~~NDSI~~ traces back to band rationing techniques (Kyle et al., 1978; Dozier, 1984) related to the NDVI (Rouse et al., 1974; Tucker, 1979) and is based on the physical principle that snow reflection is significantly higher in the visible range of the spectrum than in mid-infrared. The index ranges between -1 and 1 and a differentiation between snow and no snow is based on a $NDSI$~~NDSI~~ threshold value ($NDSI_{thr}$) which is commonly assumed to be 0.4 (Dozier, 1989; Hall and Riggs, 2007; Sankey et al., 2015). According to Hall et al. (2001) the accuracy for monthly snow detection using the MODIS product with its standard threshold is about 95% in non-forested and about 85% in forested areas. Accuracies in this range even though for the atmospherically corrected MODIS snow cover product (MOD10/MYD10) make $NDSI$~~NDSI~~ based snow cover products well accepted for global scale applications, but uncertainties have to be expected at the local scale (Härer et al. 2016). Moreover, the snow detection algorithm for the MODIS snow cover product changed in the latest collection 6. The algorithm now uses a NDSI threshold of zero together with a flag system to detect snow cover and users are encouraged to use their own NDSI threshold in the MODIS Snow Products Collection 6 User Guide if a binary snow cover map is wanted.

In ~~this context~~line with this, numerous recent studies have questioned the general applicability of a standard $NDSI_{thr}$ in local snow and glacier monitoring. When calibrating the $NDSI_{thr}$ manually or by automated methods against field data for single scenes, large deviations from the standard value of 0.4 have been observed. The published values range from 0.18 to 0.7 (Burns and Nolin, 2014; Härer et al., 2016; Maher et al., 2012; Racoviteanu et al., 2009; Silverio and Jaquet, 2009; Yin et al., 2013). The wide range of values show the spatio-temporal variability of the $NDSI_{thr}$ and raise the question for a valid non-subjective method to define this value.

Maher et al. (2012), for example, assumed a spatially calibrated $NDSI_{thr}$ of 0.7 to be constant over time. The comprehensive work of Yin et al. (2013) compared various automatic entropy-based, clustering-based, and spatial threshold methods to adjust the $NDSI_{thr}$ for specific satellite images. The findings of Yin et al. (2013) are based on single-date comparisons at five sites around the world and were undertaken on a regional scale. The clustering-based image segmentation method developed by Otsu (1979) compared best to the evaluation data sets, which is why the Otsu method is used as comparative data in here.

Härer et al. (2016) have presented a calibration strategy for satellite derived snow cover maps on the basis of local camera systems. The achieved results have shown that $NDSI_{thr}$ can be distinctly different in course of the snow cover period and that there is a need for a temporal adaption of $NDSI_{thr}$ for achieving valid results in view of the local SCA.

The aim of the presented study is to evaluate the variability of $NDSI_{thr}$ in space and time and to test if this variability does lead to significant uncertainties in the existing snow cover maps. A scaling exercise which has investigated up to which scale a locally adapted threshold can improve the classification results shows the limits of the fixed threshold approach at the local scale.

We use the camera-based calibration approach (Härer et al., 2013) as reference as it has shown its low error margins in comparison to high resolution locally derived 1m resolution snow maps at RCZ (Härer et al. 2016). The results achieved by this approach are then compared to the automatic segmentation method of Otsu (1979), which has proven to be one of the most performant snow detection methods available today (Yin et al., 2013) and to the standard threshold of 0.4, as well as to a location specific threshold of 0.7 (Maher et al., 2012). Moreover, we present a seasonal model calibrated with the *NDSI* threshold time series. The quadratic polynomial model can also be locally adapted by including a geology dependent offset which is comparable to earlier findings of Chaponnière et al. (2005). The results will reveal the performance of the different approaches and will clarify for which scales a fixed *NDSI*~~NDSI~~ threshold can be an adequate solution.

2 Study Site and Data

The presented study focuses on two mountain sites in the European Alps, the Research Catchment Zugspitzplatt (RCZ) located in Germany (47°40' N/11°00' E; Bernhardt et al., 2015; Weber et al., 2016) and the Vernagtferner (VF) catchment in Austria (46°52' N/10°49' E; Fig. 1a to c; Abermann et al., 2011). RCZ is a partly glacierized headwater catchment with a spatial extent of about 13.1 km². It stretches from 1371 to 2962 m a.s.l.. ~~and~~ The substrate is ~~is~~ mainly ~~mainly built up by~~ limestone. VF is

also an alpine headwater basin with a size of 11.5 km² and a glaciated part of about 7.9 km² (Mayr et al., 2013). It ranges from 2642 to 3619 m a.s.l. and the ~~pending~~ underlying rock is gneiss.

Both sites are equipped with similar single lens reflex camera systems for monitoring wide parts of the catchments starting from May 2011 at RCZ and from August 2010 at VF. The photographs are recorded as 8-bit data with three colour channels (red, green and blue; RGB) on an hourly basis for RCZ and three times a day for VF. The camera locations at the study sites are depicted in Fig. 1a and b and the camera orientations are Southwest at RCZ and West-Northwest at VF. Both investigation areas are located within a single Landsat scene (Fig. 1c) which guarantees ~~for~~ comparable illumination conditions and allows for a direct comparison of the $NDSI_{thr}$ between both sites.

Overall, 156 Landsat scenes from Landsat 5 TM, 7 ETM+ and 8 OLI were available for the observation period between 18 August 2010 and 31 December 2015. Suitable satellite image-photograph pairs were available at 15 dates for RCZ and VF, at one date for RCZ and in 32 ~~cases~~ dates for VF only. The differences stem from the local weather conditions, from the different lengths of the local photograph time series, and from the restriction that a $NDSI_{thr}$ calibration with PRACTISE or the clustering-based image segmentation from Otsu (1979) can only be applied if there is no full snow coverage in the area.

For the photo rectification part in our study, digital elevation models (DEM) with a horizontal resolution of 1 m of RCZ and VF are used, as well as orthophotos with a sub-meter spatial resolution and topographic maps as additional material to ensure an optimal geometric accuracy.

3 Methods

Our study investigates~~s~~ the differences of automatically derived $NDSI_{thr}$ from a) Landsat satellite imagery and b) terrestrial photography with literature values and displays~~s~~ their effects on the resulting snow cover maps.

Radiometrically and geometrically corrected Landsat Level 1 data was used in combination with the cloud and shadow masking software Fmask of Zhu et al. (2015). Any pixel with a cloud probability exceeding 95% in this analysis was excluded with a surrounding buffer of three pixels (Härer et al., 2016). The top of atmosphere reflectance values were calculated according to the Landsat user handbook but ~~N~~o atmospheric correction ~~i~~was applied to the Landsat data to facilitate a direct comparison to ~~the majority of~~many studies that apply the $NDSI$~~NDSI~~ for snow cover mapping, especially in high elevation areas where atmospheric influence is known to be low (Bernhardt and Schulz, 2010; ~~Maussion et al., 2011;~~ Maher et al., 2012; Warscher et al., 2013~~; Sankey et al., 2015~~).

The normalized-difference snow index ($NDSI$~~NDSI~~) wa~~i~~s calculated in accordance to Dozier (1989) by using green (~0.55 µm) and mid-infrared (MIR, ~1.6 µm) reflectance values:

$$NDSI = \frac{\rho_{\text{green}} - \rho_{\text{MIR}}}{\rho_{\text{green}} + \rho_{\text{MIR}}}, \quad (1)$$

$NDSI$~~NDSI~~ values can range between -1 and 1 and the $NDSI_{thr}$ defines the $NDSI$~~NDSI~~ value from which on the satellite pixel is assumed as snow covered. We used fixed $NDSI_{thr}$ values and dynamically derived $NDSI_{thr}$ values in course of this. In case

of the fixed values, the standard of 0.4 and a literature value of 0.7 (Maher et al., 2012) were used. For the dynamic approaches, the clustering-based image segmentation approach from Otsu (1979) and a terrestrial camera-based calibration approach of Härer et al. (2016) were applied.

By using Otsu (1979), the $NDSI_{thr}$ wa~~i~~s calibrated by maximizing the between-class variance of the two classes snow and no snow:

$$\max_{-1 \le NDSI_{thr} \le 1} \{\sigma_O^2\} = \max_{-1 \le NDSI_{thr} \le 1} \{P_s(NDSI_{thr})\, P_{ns}(NDSI_{thr})\, [\mu_s(NDSI_{thr}) - \mu_{ns}(NDSI_{thr})]\}, \quad (2)$$

where P_s and P_{ns} are the probabilities of the classes snow and no snow with respect to the $NDSI_{thr}$, and μ_s and μ_{ns} are the mean values of these two classes. The probability of P_s ~~i~~is~~s~~ thereby calculated as the number of pixels with $NDSI$~~NDSI~~ values above the $NDSI_{thr}$ divided through the total number of pixels in the image. P_{ns} calculates the absolute difference of P_s to 1.

~~It has to be mentioned that w~~We further restricted the satellite image area used for deriving $NDSI_{thr}$ in accordance to Otsu (1979) to the catchment area of RCZ and VF to allow for a spatio-temporal variable $NDSI$~~NDSI~~ threshold value within the investigated satellite scenes. Moreover, this definition facilitated ~~and to allow for a~~ to directly compar~~ison e of~~ the locally derived thresholds using the Otsu method and our own method presented in the next paragraph.~~.~~

The second dynamic method to calibrate the $NDSI_{thr}$ of the Landsat data for RCZ and VF use~~s~~d ground-based photographs as baseline.

The Matlab software PRACTISE (version 2.1; Härer et al., 2013 and 2016) wa~~i~~s utilized first to georectify the available terrestrial photographs and secondly to calibrate the $NDSI_{thr}$. For doing so, overlapping areas in the photograph-satellite image pairs we~~a~~re used. For further understanding, Figure 2 gives a~~n~~ general overview of the needed input, the internal processing steps and the generated output data of PRACTISE 2.1. The first program part georectifie~~ds~~ the photograph~~ie~~s and differences between areas with and without snow. This results in a high resolution photography-based snow cover map (Fig. 2, left column). The second part calibrate~~ds~~ the $NDSI_{thr}$ for the satellite scene of interest and use~~s~~d the achieved value to calculate a $NDSI$~~NDSI~~ based satellite snow cover map (Fig. 2, right column).

The photo georectification is based on the assumption that the recorded two-dimensional photograph (Fig. 3, blue colour) is geometrically connected to the three-dimensional real world (Fig. 3, black colour). Knowing the camera type, its lens and sensor system, as well as the camera location and orientation, a georectification becomes possible if a high resolution digital elevation model (DEM) is available as well.

Having this theoretical background in mind, we outlined the single processing steps for a photograph and a Landsat 7 scene of VF on 17 November 2011 (Figures 4 a to e, 5 a to c).

Before the PRACTISE program wa~~i~~s used, any possible distortion effects of the photograph caused by the camera lens ~~a~~were removed by utili~~s~~zing the freely available Darktable software (http://www.darktable.org/) and LensFun parameters (http://lensfun.sourceforge.net/). Now that all data wa~~i~~s available and ready, the PRACTISE program evaluation ~~can~~ could start.

In a first step information about the camera location and orientation wa~~i~~s needed for ~~an~~ georectification of the photography. This information wa~~i~~s automatically optimi~~z~~s~~e~~d by using ground control points (GCPs, Fig. 4a; Sect. 3.3 in Härer et al., 2013). The calculated viewpoint and viewing direction wea~~a~~re by default used to perform a viewshed analysis (Fig. 4b; Sect. 3.1 in Härer et al., 2013). The viewshed wa~~i~~s needed for an identification of areas which ~~a~~were visible from the viewpoint and which wea~~a~~re not obscured by topographical features or within a user-specified buffer area around the camera. The respective DEM pixels wea~~a~~re then projected to the photo plane (Fig. 4c; Sect. 3.2 in Härer et al., 2013).

Now, the snow classification module wa~~i~~s activated to distinguish between snow covered and snow-free DEM pixels (Fig. 4d). Two major procedures ~~are~~ were available for classification~~:~~. ~~A~~a statistical analysis which wa~~i~~s using the blue RGB band (Salvatori et al. 2011; Sect. 3.4 in Härer et al., 2013) and a principal component analysis (PCA) based approach (Sect. 3.1 in Härer et al. 2016). The first ~~i~~was used for shadow-free scenes, the second for scenes with shaded areas. Section 3.4 in Härer et al. (2013) ~~and (2016)~~ give~~s~~ more insights into t~~he used classification algorithms and their performance as well as on~~ a third manual option if none of the two classification routines ~~can~~ could be applied successfully. The ~~used~~ photograph snow cover maps d~~o~~id have even in the case that an insufficient classification algorithm was used for a specific situation less than 5% misclassifi~~c~~ied pixels in the worst case region of the photograph~~, which was proven by visual inspection~~ in Chapt. 4 in Härer et al. (2013). It was also shown in an earlier publication that the classification of shadow-affected photographs are of the same quality as photographs without shadows (Chapt. 4 in Härer et al., 2016). As for this study, every classified image was visually inspected and no major snow classification errors comparable to our worst case example in the previous publication were found, we expect a relative misclassification error of 1%. For this example photograph, the snow classification algorithm utili~~s~~zing a principal component analysis (PCA) was selected to account for the shadow-affected areas in the upper left part of the photograph (Fig. 4d, enlarged view in Fig. 4e).

After the~~-~~ photograph rectification and classification, the remote sensing routine of PRACTISE beg~~i~~an~~s~~ with the identification of satellite pixels that spatially overlap with the photograph snow cover map (Sect. 3.2 in Härer et al., 2016). The used photograph and satellite image were thereby recorded within one (RCZ) to three (VF) hours. ~~It also generates~~ Moreover, a cloud- and shadow-free satellite image is generated by using fmask (Zhu et al., 2015). The needed $NDSI$~~NDSI~~ map wa~~i~~s calculated in accordance to Eq. (1) by PRACTISE (Fig. 5a).

If both, the $NDSI$~~NDSI~~ satellite map and the corresponding high resolution photograph snow cover map were processed, an iterative calibration of the $NDSI$~~NDSI~~ threshold value wa~~i~~s started. The Landsat image was thereby resampled to the finer resolution of the photograph in the calibration to avoid losing any information by the aggregation of the photograph snow cover map. ~~to acquire the~~The best agreement between the local scale (photograph) and the large scale (Landsat) snow cover map was detected by maximi~~s~~zing the accuracy which is the ratio of identically classified pixels to the overall number of photograph-satellite image pixel pairs n (Aronica et al., 2002):

$$F = \frac{(a+d)}{n}, \quad (2)$$

a ~~thereby~~ represents the number of correctly identified snow pixels and *d* the same for no snow pixels. *F* is between 0 and 1 and becomes 1 for a perfect agreement between the two images.

The calibrated $NDSI$ threshold was finally applied to the original Landsat data with 30m pixel size to generate the final Landsat snow cover map. Figure 5b shows the resulting satellite snow cover map superimposed on the photograph snow cover map and a Landsat Look image. A cutout is shown for more detail in Fig. 5c.

~~It has to be mentioned that t~~The glacier retreat between DEM production years (2007, 2010) and analysis period 2010-2015 has resulted in a discrepancy between real world elevations and the available DEMs, especially in the last years of the observation period. Figure 6 exemplarily depicts the glacier retreat between 2007 and 2010 by superimposing the ice mass loss on an orthophoto of VF from 2010.

This loss in elevation leads to inaccuracies in the georectification results of the photographs. And a test for the photograph of 28 August 2010 applying the DEM of 2007 and 2010 showed that these georectification issues in turn affect the $NDSI_{thr}$ calibration results. For the DEM from 2007, the calibrated $NDSI_{thr}$ is 0.47 while the correct threshold for the up-to-date DEM from 2010 is 0.52. As a consequence, we limited the analysis to higher elevated and thus colder areas of the catchment where glacier retreat is marginal (areas north-west of the green line in Fig. 5b and Fig. 6).

To ensure that reducing the spatial overlap between photograph snow cover map and $NDSI$~~NDSI~~ satellite map does not have any negative effect on the calibrated $NDSI_{thr}$, we firstly calibrated the $NDSI_{thr}$ for the three investigated Landsat scenes in 2010 for the complete and the upper area only. Moreover, we calibrated the $NDSI_{thr}$ for the 44 remaining scenes between 2011 and 2015 using the upper area DEM from 2007 and 2010 to test for a $NDSI_{thr}$ sensitivity in the longer time series. For both approaches, the differences between the calibrated $NDSI_{thr}$ never become larger than 0.01. Hence, we assume that our calibration approach of using the higher elevated areas at VF which is incorporated in PRACTISE by excluding a radius of 1800 m around the camera from the analysis (green line in Fig. 5b and Fig. 6) is valid for the complete analysed time series between 2010 and 2015.

As ~~W~~we did not find a similar effect on the $NDSI_{thr}$ calibration in our tests at RCZ~~. Hence~~, there was no need to remove the glacier areas at RCZ from the analysis.

4 Results and Discussion

The $NDSI$~~NDSI~~ thresholds derived by the two dynamic methods are now discussed and related to static thresholds.

The $NDSI_{thr}$ predicted by the Otsu method are densely grouped around 0.4. This is underlined by a mean of 0.36 and a standard deviation of 0.04 at RCZ and a mean of 0.41 with a corresponding standard derivation of 0.04 at VF (Tbl. 1). The statistics do not include two dates at VF as no separating $NDSI_{thr}$ could be found by using the Otsu method here (squares in Fig. 7a). This stands in contradiction to the real situation as the photographs do show that there was no full snow coverage at the respective dates which would generally allow for an prediction of $NDSI_{thr}$. This shows that the application of the Otsu method is potentionally uncertain in nearly fully snow covered situations. Furthermore, ~~a tendency to slightly higher mean $NDSI_{thr}$ at~~

~~VF and slightly lower thresholds at RCZ~~ the small differences to the standard of 0.4 with a tendency to slightly higher mean $NDSI_{thr}$ at VF and slightly lower thresholds at RCZ ~~could be detected and the very small observed differences to the standard of 0.4 would~~do not ~~underline~~ justify the ~~need~~ additional effort ~~for~~ of the Otsu method for a location dependent threshold prediction. Additionaly, the weak seasonal dynamics which can be found at VF would also not require a time dependent calculation of the threshold.

The camera-based method leads in general to a more dynamic $NDSI_{thr}$ in time and to a higher systematic difference of $NDSI_{thr}$ between the two sites. The archived 16 $NDSI_{thr}$ at RCZ and 47 $NDSI_{thr}$ at VF are compared in a first step. The presumption of a comparable $NDSI_{thr}$ for both sites could not be confirmed in this case. Significant differences were detected despite the fact that both sites are high alpine and are located within a single Landsat scene. Moreover, the calibrated $NDSI_{thr}$ were in large parts significantly different to the standard value of 0.4. Figure 7b and Table 1 illustrate the variability and the range of $NDSI_{thr}$ at both sites. The minimum value at RCZ is 0.15 while the maximum value is 0.39. The values at VF are in general on a higher level and range between 0.35 and 0.74. Both sites thus strongly scatter around their catchment-specific mean value (0.28 at RCZ, 0.57 at VF) but show a characteristic development over the year ~~-~~(Fig. 8~~9~~) which is also detected in a ~~significantly~~ weaker form for the Otsu method at VF. Independent of the fact that this seasonal dynamic is comparable for both sites using the camera-based method. Fig. 7b highlights that the correlation coefficient between $NDSI_{thr}$ at RCZ and VF is very low when they are compared on a date by date basis (r=0.17). By contrast, a correlation between the Otsu method and the terrestrial camera-based method at VF of -0.56 is found which however cannot be observed at RCZ between the two methods (r=0.10, Fig. 7a and b).

The results of the camera-based methods require a deeper investigation to analyse if such different $NDSI_{thr}$ are justifiable. Despite the strong scatter and the resulting low correlation, the differences in the catchment-specific mean $NDSI_{thr}$ levels seem to be systematic (Tbl. 1). Topographic characteristics could be a possible reason. These are similar with respect to elevation, slope and aspect but different for the ~~pending~~ underlying rock being limestone at RCZ and gneiss at VF. We hence investigated the $NDSI$~~NDSI~~ ~~reflectance~~ values for the snow-free bare rock areas within each catchment. ~~This is valid for the complete time series as the steepest almost vertical rock faces in the catchment are snow free in all used scenes.~~ Figure 9~~8~~ presents frequency histograms of the~~se~~ $NDSI$~~NDSI~~ ~~reflectances~~ for five summer dates. Other seasons were excluded due to the increased probability of fractional snow cover in the Landsat pixels. The tests show that the maximum frequencies after smoothing the histogram are stable for these dates for each catchment. The mean maximum frequency is about -0.34 at RCZ and 0.01 at VF. This corresponds to the spectral behaviour of limestone and gneiss. The typical limestone is lighter than gneiss in the visible range but the reflectance further increases for wavelengths up to 2 µm while it stays similar for a typical gneiss. As the NDSI calculates the difference between the green (0.55 µm) and the mid-infrared wavelength (1.55 µm) in the numerator and uses the sum of these two bands in the denominator, limestone has therefore a negative value and gneiss is around zero. The mean $NDSI$~~NDSI~~ ~~reflectance~~ difference of the rocks at RCZ and VF amounts to about 0.34. This difference is comparable to the mean systematic difference of 0.26 found for the mean calibrated $NDSI_{thr}$ at both sites. It is therefore

probable that the different rock types and therewith the background radiation triggers the catchment-specific mean $NDSI_{thr}$ levels which in turn supports the idea of adapting $NDSI_{thr}$ locally.

Next, the effect of the calibrated $NDSI_{thr}$ on the predicted snow covered area (SCA) at RCZ and VF is analysed. The differences between the SCA predicted with the standard threshold of 0.4 and with the Otsu method are in principle small. This can be related to the minor differences between standard $NDSI_{thr}$ and the threshold predicted over Otsu. The absolute differences are 0.05 km² in average for VF and 0.15 km² for RCZ. The effects achieved with the photographic method instead are on a level which questions the applicability of the standard threshold for local investigations. The differences in SCA inbetween the products using the calibrated $NDSI_{thr}$ and the standard threshold of 0.4 are calculated using the camera-calibrated SCA as baseline which has shown the highest accuracy of the derived snow cover products when compared to the available photoclassifications of PRACTISE (Härer et al. 2016):

$$SCA_{diff\%} = \frac{100\,(SCA_{0.4} - SCA_{cam})}{SCA_{cam}} \quad (3).$$

The values are between -24.1% at RCZ and +17.2% at VF (Fig. 7c) and reveal how ~~much uncertainty currently exists in~~different standard and calibrated $NDSI$ ~~NDSI~~ based snow cover maps are on the small scale. The deviations are in general larger at RCZ where the calibrated $NDSI$ ~~NDSI~~ threshold values are mainly below 0.4. This means that the SCA is systematically underestimated when using the standard of 0.4. The lower error ~~in percents~~ at VF compared to the error percentages at RCZ can be related to the generally higher snow covered area in the VF catchment. These relative differences result in turn in significantly different absolute SCA~~-~~ (standard threshold versus calibrated threshold). Here, the highest differences are 1.09 km² at RCZ and 1.67 km² at VF. This is a relevant error margin especially if the small catchment sizes of only 13.1 km² (RCZ) and 11.5 km² (VF) are taken into account.

Given this finding and the large variability observed in calibrated $NDSI_{thr}$ it is obvious that ~~widely used~~ methods ~~(e.g. Maher et al., 2012)~~ which locally calibrate the $NDSI_{thr}$ for a single date and then apply this threshold at multiple dates are ~~also~~ probably no solution. ~~and can even deteriorate the accuracy compared to the standard threshold method.~~ An example is the application of a calibrated threshold of 0.7 ~~of 0.7~~ at VF to the complete time series in this catchment. We use 0.7 here as~~.~~ Maher et al. (2012) state this value in their study and as it is in the plausible range of the observed $NDSI$ thresholds at VF (0.35 to 0.74). However, when applied to the complete time series,~~T~~ this approach results in a mean absolute error in SCA of 1.26 km² compared to an average deviation of 0.41 km² for the standard threshold method. This approach thus might help in some studies where, by luck, a $NDSI$ threshold is found for the calibration date that also describes the other analysis dates well. However, our example shows that the chances are also high that it deteriorates the accuracy compared to the standard threshold method when applied to other dates.

An alternative to the temporally constant threshold methods is a statistical modelling approach fitted to the calibrated $NDSI_{thr}$. This however requires a solid set of calibration data to adjust the model to the observations at multiple dates. VF hence serves as an example for this approach because of its higher data availability. As stated before a seasonal dynamic in the calibrated

$NDSI_{thr}$ could be observed at both sites. ~~This temporal development is potentially related to the sun angle, snow age, grain size or albedo development or other effects. A detailed investigation of the reasons of this effect is beyond this study but will be subject of future studies.~~ A quadratic polynomial model was fitted against the day of year for~~to~~ the calibrated $NDSI_{thr}$ ~~for~~ of the years 2010 to 2013 at VF ($NDSI_{vf}$, Fig. 8~~9~~). $NDSI_{vf}$ might not exactly reproduce the calibrated thresholds at any time step (r^2=0.45; RMSE=0.06) but the evaluation of this simple model for 2014 and 2015 at VF shows a remarkable reduction in the average SCA error from 0.35 km² when applying the standard threshold of 0.4 down to 0.17 km². These results are promising and it needs to be investigated if the seasonal behaviour of the calibrated $NDSI$ thresholds can be attributed to the elevation and azimuth angles of the sun. The correlation r between azimuth angle and $NDSI$ is 0.75 for RCZ and 0.42 for VF. For sun elevation, r is 0.77 for RCZ and 0.54 for VF. The sun angles thus are correlated to the seasonal development but do not fully explain the behaviour. The temporal development is thus potentially also related to snow age, grain size, albedo development or other effects. These might also explain the observed variability within the seasons. A detailed investigation of this variability is however beyond this study but will be subject of future studies.

As not any site is equipped with camera infrastructure, it was also tested if the achieved regression model can be transferred to RCZ while including information about the geology dependent offset between the average $NDSI_{thr}$ values. Hence, the model is fitted to the complete calibrated $NDSI_{thr}$ time series at VF (r^2=0.36; RMSE=0.07) and a term (-0.34) for the systematic mean $NDSI$~~NDSI~~ ~~reflectance~~ difference of the rocks at RCZ and VF is added ($NDSI_{rcz}$, Fig. 9~~8~~). The evaluation of $NDSI_{rcz}$ seems to slightly underestimate the calibrated $NDSI_{thr}$ at RCZ. Nevertheless, the quadratic polynomial model accounting for the reflectance differences at different sites results in a significant reduction of snow cover mapping uncertainties of 40% as the mean SCA error amounts to 0.18 km² while the application of the standard threshold method causes an average deviation in snow cover of 0.31 km² in RCZ. Given the assumption that the seasonal dynamic and the correction factor are generally applicable, the presented seasonal model derived from the multi-year use of PRACTISE at a single site is hence not only temporally but by using information about the spectral properties of the pending rock types without the need for other camera systems also spatially transferrable. This assumption and the transferability of the model is probably only true for high elevation areas. Even though that the $NDSI$ is an index which reduces the dependence on atmospheric conditions, an atmospheric correction might be necessary as well as more dynamic approaches that reflect the vegetation growth and senescence over the year for lowland areas. Hence, the approach needs to~~will~~ be further evaluated and developed in future studies with more test sites.

We have now underlined the importance of a locally adapted $NDSI$~~NDSI~~ threshold calibration for Landsat snow cover maps at the two presented catchments. However, the detected $NDSI$~~NDSI~~ threshold dependency automatically leads to the question if the need for threshold adaption is also necessary for coarser resolution satellite snow cover maps~~. This is of special interest as~~, for example, for ~~MODIS snow cover products are today the most frequently applied satellite snow cover maps. They are based on the NDSI technique and the 0.4 threshold and have~~ a spatial resolution of 500 m or 1 km. Hence, we aggregated the Landsat snow cover maps using calibrated and standard $NDSI$~~NDSI~~ threshold values from 30 m to 90 m, 210 m, 510 m, and

990 m resolution. Our aggregation experiment of the Landsat snow cover maps for the different $NDSI$ thresholds shows ~~It can be seen~~ that the SCA deviation between standard and calibrated snow cover maps diminishes for coarser resolution data in most cases. Figure 10 a outlines this error reduction with spatial aggregation for a Landsat 7 scene of Vernagtferner catchment on 16 September 2011. Figure 10 b shows the simultaneously captured photograph used for calibration. We however cannot draw an absolute conclusion from fig. 10 a that the difference in snow cover maps between the different thresholds is always reduced with a coarser resolution. The simple reason is that with larger pixel sizes, the number of pixels in the catchment becomes lower and the relative weight of a pixel being different for different thresholds has a larger relative weight. Therefore, we decided to investigate at which spatial resolution the standard and calibrated snow cover maps become identical for the 63 cases investigated in the two catchments. ~~Figure 10 c underlines this finding by depicting the spatial resolution at which standard and calibrated snow cover maps become identical for the 65 cases investigated in the two catchments.~~ This variable is absolute and thus independent of relative weights and changes with spatial aggregation. The aggregation step to 510m is thereby of major importance as more than 90% (58 of 63) of SCA maps ~~for our investigation period and study~~ become identical at this pixel size. Thus, using the standard threshold of 0.4 instead of the higher $NDSI$ thresholds at VF and the lower $NDSI$ values at RCZ seems to be accurate in most cases ~~case of the MODIS snow cover product~~ with a pixel size of 500m. For applications at this scale, the ~~additional effort~~positive effect of using camera calibrated data ~~only provides slight improvements~~diminishes and might rarely justify the effort. However, our new method using camera-calibrated data ~~allows to~~focuses on setting in value the higher resolution satellite data of the Landsat series and of the new Sentinel -2.

5 Conclusions

The study has revealed that using the standard threshold of 0.4 is adequate for satellite products with a pixel size of 500 meters and more. For higher resolution snow cover mapping, significant improvements in the quality of the snow cover maps can be achieved if a threshold is used which is variable in space and time. The clustering-based segmentation technique of Otsu is producing results which are only slightly different from those of the standard threshold of 0.4 and do not indicate a need for a further adaption. However when compared to local images, the resulting differences are becoming obvious and could only be reduced by the presented camera-based technique. The long-term analysis of calibrated $NDSI_{thr}$ at two comparable high elevation sites has shown that large deviations from the 0.4 standard threshold exist. The calibrated optimal threshold values span a range from 0.15 to 0.74 over the complete time series and can reach a difference of 0.45 between both observation sites at a single date. It was also shown that these differences in $NDSI_{thr}$ lead to significantly different SCA when compared to the standard of 0.4.

The $NDSI_{thr}$ at both sites have similar seasonal dynamics while scattering around different site-specific average values (0.28 at RCZ, 0.57 at VF). The difference between the average threshold values at the two sites could be related to the different reflection properties of the rock types in the investigation areas (limestone at RCZ and gneiss at VF). The overall correlation

coefficient between $NDSI_{thr}$ of both sites is low (r=0.17). This ~~which~~ prohibits to generally ~~a date by date transfer of~~use the calibrated threshold values from one catchment ~~to~~in an~~the~~ other catchment of the same satellite scene.

In view of the validity of the standard threshold of 0.4 at the local scale it was found that relative SCA error margins of up to 24.1% were found for the standard threshold method when using 30m Landsat products. This is critical for any snow cover mapping application and especially for model evaluation studies. We hence conclude that the application of a fixed $NDSI$~~NDSI~~ threshold can lead to large uncertainties in the resulting snow cover products at least at the local scale. Consequently, local studies strongly need to account for the $NDSI_{thr}$ variability in space and time in order to guarantee high accuracy snow cover products. But, in case studies are carried out with sensors having a pixel size of 500 meters and more the advantage of a location dependent $NDSI_{thr}$ vanishes.

It was shown that site-specific single-date adaptations of the $NDSI_{thr}$ also do not lead to resilient results. The uncertainty introduced by a single measurement is not quantifiable and can lead to results worse than that achieved by using the standard value of 0.4. A quantitative calibration or visual derivation of the $NDSI_{thr}$ for a single date and its application to other dates is therefore ~~jeopardous~~inappropriate.

The approximation of the $NDSI_{thr}$ over a simple seasonal model fitted to the calibrated $NDSI_{thr}$ at the respective site has shown improvements instead. The achieved model was able to reduce the error in the SCA prediction by 50% when compared to the standard threshold method. Nevertheless, a fundamental data pool of in situ information covering the dynamic over the year as well as the range of possible $NDSI_{thr}$ within a season is needed for calculating this relation. Finally, it was shown that the fitted model parameters are also spatially transferable if an additional term accounts for the background radiation of the different rock types. This is possible without in situ measurements by utili~~z~~s~~s~~ing the constant $NDSI$~~NDSI~~ ~~reflectance~~ differences of the rock surface in the respective catchments. However, this needs to be further tested at more sites. Future studies will hence use the existent webcam infrastructure in the European Alps as well as camera systems installed worldwide at the INARCH network sites (Pomeroy et al., 2015) for the generation of numerous calibrated $NDSI_{thr}$. The observed threshold values will serve as operational source for applicable $NDSI_{thr}$ and will allow to evaluate the presented temporally and spatially variable prediction approach of $NDSI_{thr}$. In case of a successful evaluation, the presented scheme allows for an objective and reproducible derivation of the $NDSI_{thr}$ value for any given satellite scene. This is a large advantage as the threshold is up to now often set intuitively or assumed as constant which does neither conform to the complexity of the models evaluated on basis of $NDSI$~~NDSI~~ based snow cover maps nor to the needs of the models which are assimilating these maps.

Acknowledgements

This work was funded by the Austrian Science Fund (I 2142-N29), the doctoral scholarship program of the German Federal Environmental Foundation (DBU), the Helmholtz Research School Mechanisms and Interactions of Climate Change in Mountain Regions (MICMoR) and has additionally received a fundamental support of the Environmental Research Station Schneefernerhaus (UFS) in course of the Virtual Alpine Observatory (VAO). The Commission for Glaciology of the Bavarian

Academy of Sciences and Humanities has kindly provided data of Vernagtferner. We want to thank the crew of the UFS (Markus Neumann, Dr. Till Rehm and Hannes Hiergeist) for supporting this peace of research by hosting the authors and maintaining the camera system. Thomas Werz and Michael Weber have also supported the research by temporally maintaining the camera system. Relevant data can be made available by the authors.

References

Abermann, J., Kuhn, M., and Fischer, A.: A reconstruction of annual mass balances of Austria's glaciers from 1969 to 1998, Ann. Glaciol., 52(59), 127–134, doi:10.3189/172756411799096259, 2011.

Agosta, C., Fettweis, X., and Datta, R.: Evaluation of the CMIP5 models in the aim of regional modelling of the Antarctic surface mass balance, The Cryosphere, 9(6), 2311–2321, doi:10.5194/tc-9-2311-2015, 2015.

Aronica, G., Bates, P. D., and Horritt, M. S.: Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE, Hydrol. Process., 16(10), 2001-2016, doi: 10.1002/hyp.398, 2002.

Bernhardt, M., and Schulz, K.: SnowSlide: A simple routine for calculating gravitational snow transport, Geophys. Res. Lett., 37, L11502, doi:10.1029/2010GL043086, 2010.

Bernhardt, M., Schulz, K., and Pomeroy, J.: The International Network for Alpine Research Catchment Hydrology. A new GEWEX crosscutting Project, Hydrol. Wasserbewirts., 59(4), 190-191, 2015.

Bernhardt, M., Schulz, K., Liston, G. E., and Zängl, G.: The influence of lateral snow redistribution processes on snow melt and sublimation in alpine regions, J. Hydrol., 424-425, 196-206, doi:10.1016/j.jhydrol.2012.01.001, 2012.

Bernhardt, M., Härer, S., Jacobeit, J., Wetzel, K. F., and Schulz, K.: The Virtual Alpine Observatory - research focus Alpine hydrology, Hydrol. Wasserbewirts., 58(4), 241-243, 2014.

Burns, P., and Nolin, A.: Using atmospherically-corrected Landsat imagery to measure glacier area change in the Cordillera Blanca, Peru from 1987 to 2010, Remote Sens. Environ., 140, 165-178, doi:10.1016/j.rse.2013.08.026, 2014.

Butt, M. J., and Bilal, M.: Application of snowmelt runoff model for water resource management, Hydrol. Process., 25(24), 3735-3747, doi:10.1002/hyp.8099, 2011.

Chaponnière, A., Maisongradne, P., Duchemin, B., Hanich, L., Boulet, G., Escadal, R., Elouaddat, S.: A combined high and low spatial resolution approach for mapping snow covered areas in the Atlas mountains, International Journal of Remote Sensing, 26(13), 2755-2777, doi: 10.1080/01431160500117758, 2005.

Deb, D., Butcher, J., and Srinivasan, R.: Projected Hydrologic Changes Under Mid-21st Century Climatic Conditions in a Sub-arctic Watershed, Water Resour. Manag., 29(5), 1467-1487, doi:10.1007/s11269-014-0887-5, 2015.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Holm, E. V., Isaksen, L., Kallberg, P., Koehler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thepaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. Roy. Meteor. Soc., 137(656), 553-597, doi:10.1002/qj.828, 2011.

Dozier, J.: Snow reflectance from LANDSAT-4 Thematic Mapper, IEEE Trans. Geosci. Remote Sens., GE-22(3), 323-328, doi:10.1109/TGRS.1984.350628, 1984.

Dozier, J.: Spectral Signature of Alpine Snow Cover from the Landsat Thematic Mapper, Remote Sens. Environ., 28, 9-22, doi:10.1016/0034-4257(89)90101-6, 1989.

Drusch, M., Vasiljevic, D., and Viterbo, P.: ECMWF's global snow analysis: Assessment and revision based on satellite observations, J. Appl. Meteorol., 43(9), 1282-1294, doi:10.1175/1520-0450(2004)043<1282:EGSAAA>2.0.CO;2, 2004.

Dutra, E., Viterbo, P., Miranda, P. M. A., and Balsamo, G.: Complexity of Snow Schemes in a Climate Model and Its Impact on Surface Energy and Hydrology, J. Hydrometeorol., 13(2), 521-538, doi:10.1175/JHM-D-11-072.1, 2012.

Dyurgerov, M.: Mountain and subpolar glaciers show an increase in sensitivity to climate warming and intensification of the water cycle, J. Hydrol., 282(1-4), 164-176, doi:10.1016/S0022-1694(03)00254-3, 2003.

Hall, D. K., and Riggs, G. A.: Accuracy assessment of the MODIS snow products, Hydrol. Process., 21(12), 1534-1547, doi:10.1002/hyp.6715, 2007.

Hall, D. K., Foster, J. L., Salomonson, V. V., Klein, A. G., and Chien, J. Y. L.: Development of a technique to assess snow-cover mapping errors from space, IEEE Transactions on Geoscience and Remote Sensing, 39(2), 432-438, doi: 10.1109/36.905251, 2001.

Hall, D. K., Riggs, G. A., Salomonson, V. V., DiGirolamo, N. E., and Bayr, K. J.: MODIS snow-cover products, Remote Sens. Environ., 83(1-2), 181-194, doi:10.1016/S0034-4257(02)00095-0, 2002.

Härer, S., Bernhardt, M., and Schulz, K.: PRACTISE – Photo Rectification And ClassificaTIon SoftwarE (V.2.1), Geosci. Model Dev., 9(1), 307-321, doi:10.5194/gmd-9-307-2016, 2016.

Härer, S., Bernhardt, M., Corripio, J. G., and Schulz, K.: PRACTISE – Photo Rectification And ClassificaTIon SoftwarE (V.1.0), Geosci. Model Dev., 6(3), 837-848, doi:10.5194/gmd-6-837-2013, 2013.

Homan, J. W., Luce, C. H., McNamara, J. P., and Glenn, N. F.: Improvement of distributed snowmelt energy balance modeling with MODIS-based NDSI-derived fractional snow-covered area data, Hydrol. Process., 25(4), 650-660, doi:10.1002/hyp.7857, 2011.

Kyle, H. L., Curran, R. J., Barnes, W. L., and Escoe, D.: A cloud physics radiometer, 3rd Conference on Atmospheric Radiation, American Meteorological Society, 28-30 June 1978, Davis, Calif., p. 107, 1978.

Liston, G. E.: Representing subgrid snow cover heterogeneities in regional and global models, J. Climate, 17(6), 1381-1397, doi:10.1175/1520-0442(2004)017<1381:RSSCHI>2.0.CO;2, 2004.

Maher, A. I., Treitz, P. M., and Ferguson, M. A. D.: Can Landsat data detect variations in snow cover within habitats of arctic ungulates?, Wildlife Biol., 18(1), 75-87, doi:10.2981/11-055, 2012.

Mankin, J. S., and Diffenbaugh, N. S.: Influence of temperature and precipitation variability on near-term snow trends, Clim. Dynam., 45(3-4), 1099-1116, doi:10.1007/s00382-014-2357-4, 2015.

Maussion, F., Scherer, D., Finkelnburg, R., Richters, J., Yang, W., and Yao, T.: WRF simulation of a precipitation event over the Tibetan Plateau, China - an assessment using remote sensing and ground observations, Hydrol. Earth Syst. Sc., 15(6), 1795-1817, doi:10.5194/hess-15-1795-2011, 2011.

Mayr, E., Hagg, W., Mayer, C., and Braun, L.: Calibrating a spatially distributed conceptual hydrological model using runoff, annual mass balance and winter mass balance, J. Hydrol., 478, 40-49, doi:10.1016/j.jhydrol.2012.11.035, 2013.

Painter, T. H., Rittger, K., McKenzie, C., Slaughter, P., Davis R. E., and Dozier, J.: Retrieval of subpixel snow covered area, grain size, and albedo from MODIS, Remote Sens. Environ., 113(4), doi: 10.1016/j.rse.2009.01.001.

Pomeroy, J., Bernhardt, M., and Marks, D.: Research network to track alpine water, Nature, 521(7550), 32-32, doi:10.1038/521032c, 2015.

Racoviteanu, A. E., Paul, F., Raup, B., Khalsa, S. J. S., and Armstrong, R.: 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, 2009.

Rangwala, I., Miller, J. R., Russell, G. L., and Xu, M.: Using a global climate model to evaluate the influences of water vapor, snow cover and atmospheric aerosol on warming in the Tibetan Plateau during the twenty-first century, Clim. Dynam., 34(6), 859-872, doi:10.1007/s00382-009-0564-1, 2010.

Rouse, J. Jr, Haas, R. H., Schell, J. A., and Deering, D. W.: Monitoring vegetation systems in the Great Plains with ERTS. 3rd Earth Resources Technology Satellite-1 Symposium - Volume I: Technical Presentations. NASA SP-351, 309-317, 1974.

Sankey, T., Donald, J., Mcvay, J., Ashley, M., O'Donnell, F., Lopez, S. M., and Springer, A.: Multi-scale analysis of snow dynamics at the southern margin of the North American continental snow distribution, Remote Sens. Environ., 169, 307-319, doi:10.1016/j.rse.2015.08.028, 2015.

Santini, M., and di Paola, A.: Changes in the world rivers' discharge projected from an updated high resolution dataset of current and future climate zones, J. Hydrol., 531, 768-780, doi:10.1016/j.jhydrol.2015.10.050, 2015.

Silverio, W., and Jaquet, J. M.: Prototype land-cover mapping of the Huascaran Biosphere Reserve (Peru) using a digital elevation model, and the NDSI and NDVI indices, J. Appl. Remote Sens., 3, 033516, doi:10.1117/1.3106599, 2009.

Sirguey, P., Mathieu, R., and Arnaud, Y.: Subpixel monitoring of the seasonal snow cover with MODIS at 250 m spatial resolution in the Southern Alps of New Zealand: Methodology and accuracy assessment, Remote Sens. Environ., 113(1), doi:10.1016/j.rse.2008.09.008, 2009.

Takata, K., Emori, S., and Watanabe, T.: Development of the minimal advanced treatments of surface interaction and runoff, Global Planet. Change, 38(1-2), 209-222, doi:10.1016/S0921-8181(03)00030-4, 2003.

Tekeli, A. E., Akyurek, Z., Sorman, A. A., Sensoy, A., and Sorman, A. U.: Using MODIS snow cover maps in modeling snowmelt runoff process in the eastern part of Turkey, Remote Sens. Environ., 97(2), 216-230, doi:10.1016/j.rse.2005.03.013, 2005.

Tennant, C. J., Crosby, B. T., and Godsey, S. E.: Elevation-dependent responses of streamflow to climate warming, Hydrol. Process., 29(6), 991-1001, doi:10.1002/hyp.10203, 2015.

Thirel, G., Salamon, P., Burek, P., and Kalas, M.: Assimilation of MODIS Snow Cover Area Data in a Distributed Hydrological Model Using the Particle Filter, Remote Sens., 5(11), 5825-5850, doi:10.3390/rs5115825, 2013.

Tucker, C. J. :Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens. Environ.,

8(2), 127-150, doi:10.1016/0034-4257(79)90013-0 , 1979.

Vavrus, S., Philippon-Berthier, G., Kutzbach, J. E., and Ruddiman, W. F.: The role of GCM resolution in simulating glacial inception, Holocene, 21(5), 819-830, doi:10.1177/0959683610394882,2011.

Viviroli, D., Archer, D. R., Buytaert, W., Fowler, H. J., Greenwood, G. B., Hamlet, A. F., Huang, Y., Koboltschnig, G., Litaor, M. I., Lopez-Moreno, J. I., Lorentz, S., Schaedler, B., Schreier, H., Schwaiger, K., Vuille, M., and Woods, R.: Climate change and mountain water resources: overview and recommendations for research, management and policy, Hydrol. Earth Syst. Sc., 15(2), 471-504, doi:10.5194/hess-15-471-2011, 2011.

Warscher, M., Strasser, U., Kraller, G., Marke, T., Franz, H., and Kunstmann, H.: Performance of complex snow cover descriptions in a distributed hydrological model system: A case study for the high Alpine terrain of the Berchtesgaden Alps, Water Resour. Res., 49(5), 2619-2637, doi:10.1002/wrcr.20219, 2013.

Weber, M., Bernhardt, M., Pomeroy, J. W., Fang, X., Härer, S., and Schulz, K.: Description of current and future snow processes in a small basin in the Bavarian Alps, Environ. Earth Sci., 75(17), 1223, doi:10.1007/s12665-016-6027-1, 2016.

Yin, D., Gao, X., Chen, X., Shao, Y., and Chen, J.: Comparison of automatic thresholding methods for snow-cover mapping using Landsat TM imagery, International Journal of Remote Sensing, 34(19), 6529-6538, doi:10.1080/01431161.2013.803631, 2013.

Zhu, Z., Wang, S., and Woodcock, C. E.: Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images, Remote Sens. Environ., 159, 269-277, doi:10.1016/j.rse.2014.12.014, 2015.

Table 1. Basic statistic measures of the automatically derived ***NDSI*** ~~NDSI~~ threshold time series at RCZ and VF using the Otsu segmentation method and the camera-based calibration method.

Site	Automatically derived NDSI threshold values									
	Mean		Standard Deviation		Max		Min		Spread	
	camera	Otsu	camera	Otsu	camera	Otsu	camera	Otsu	camera	Otsu
RCZ	0.28	0.36	0.07	0.04	0.39	0.45	0.15	0.29	0.24	0.16
VF	0.57	0.41	0.09	0.04	0.74	0.47	0.35	0.33	0.39	0.14

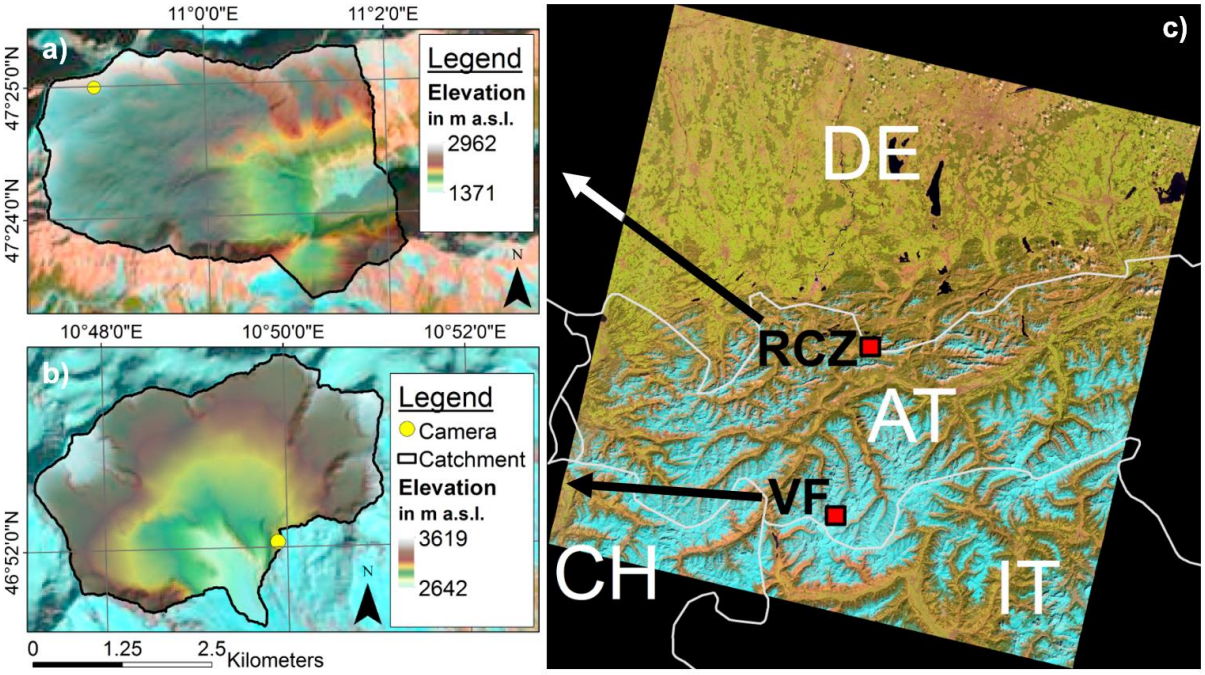


Figure 1: The figure shows the two test sites used in this study as well as their location within a Landsat scene. Both have indicated the camera location in yellow, the catchment area outlined in black and the digital elevation model (DEM) superimposed on a Landsat Look image. **a)** Research Catchment Zugspitzplatt (Germany), **b)** Vernagtferner catchment (Austria), **c)** Landsat scene (Landsat Look image, WRS2 path 193, row 27) which contains both sites.

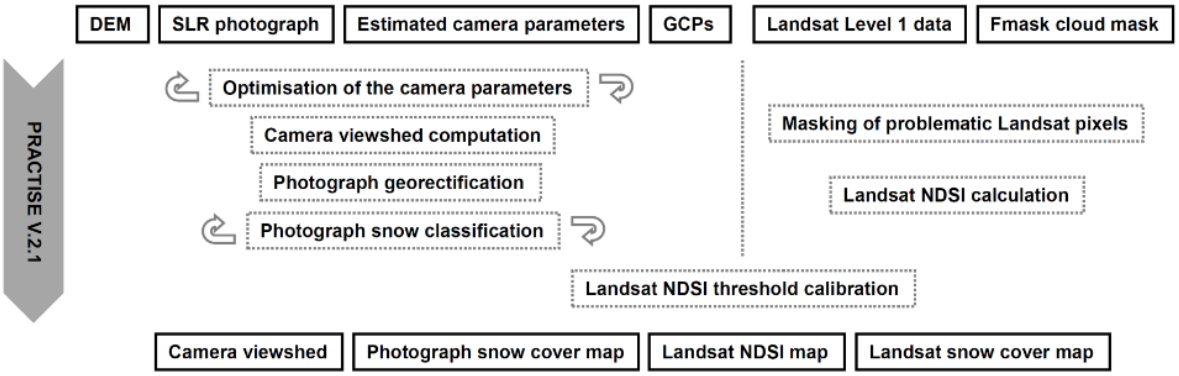


Figure 2: Input and output data as well as the workflow of PRACTISE (version 2.1) to generate the calibrated ***NDSI*** ~~NDSI~~ snow cover maps from Landsat data are depicted here (from Härer et al., 2016).

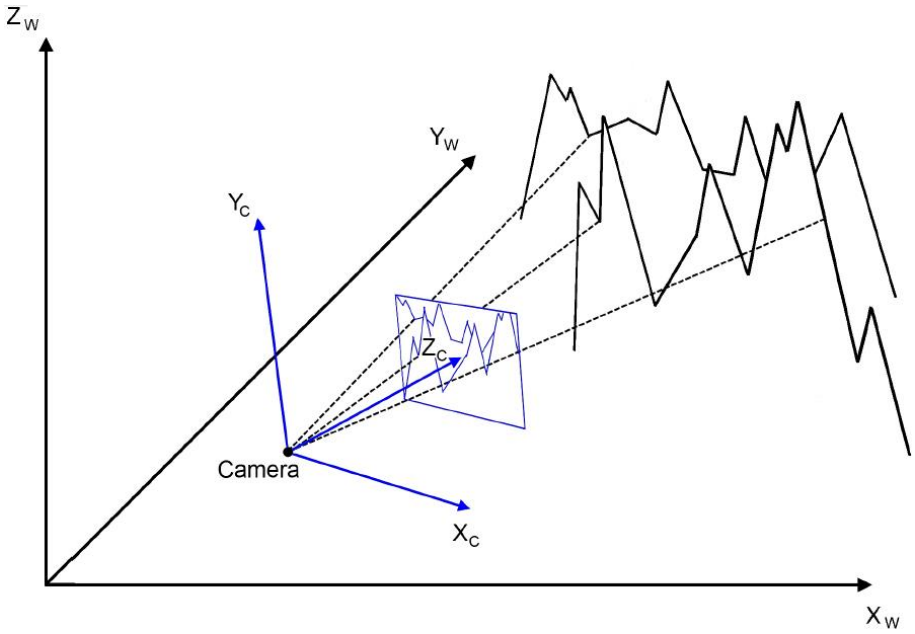
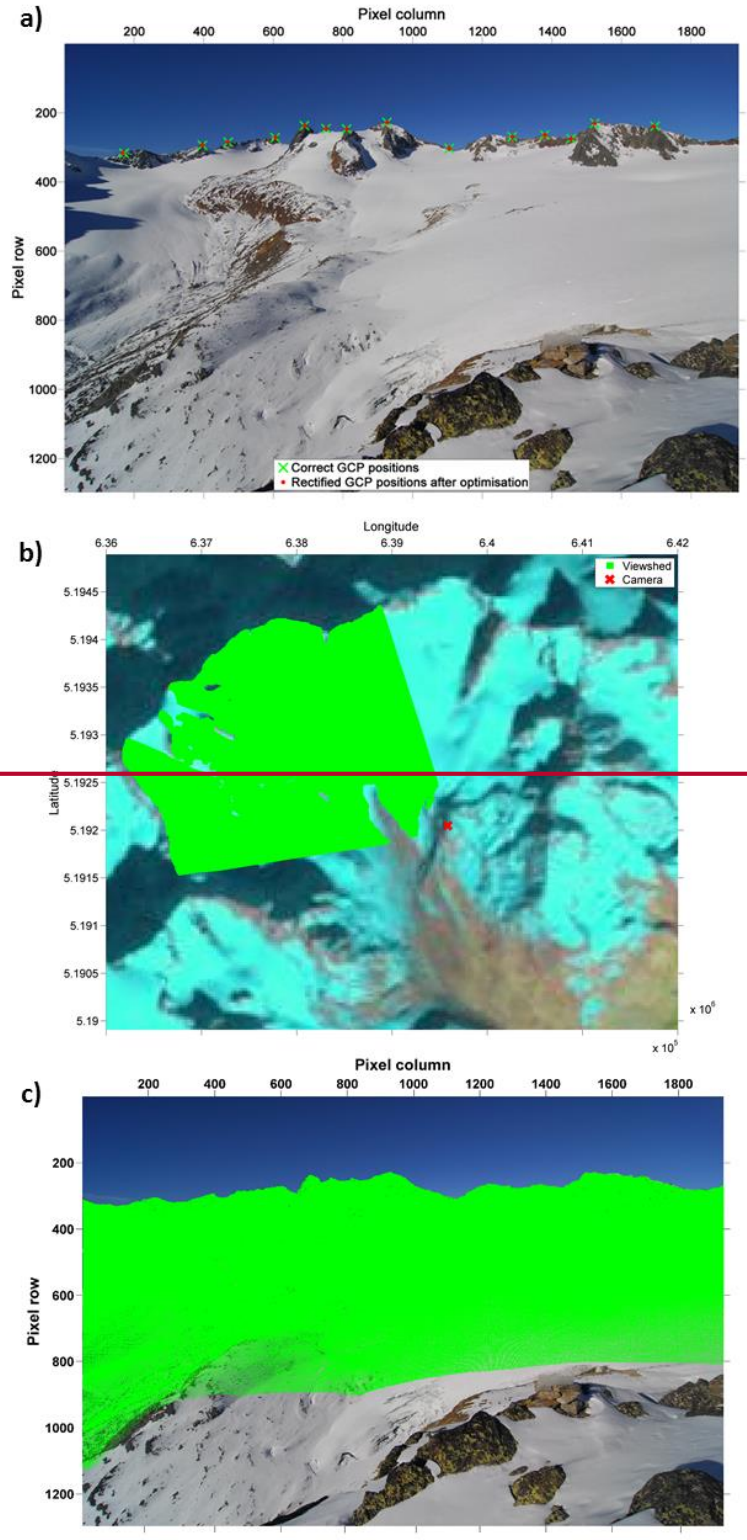
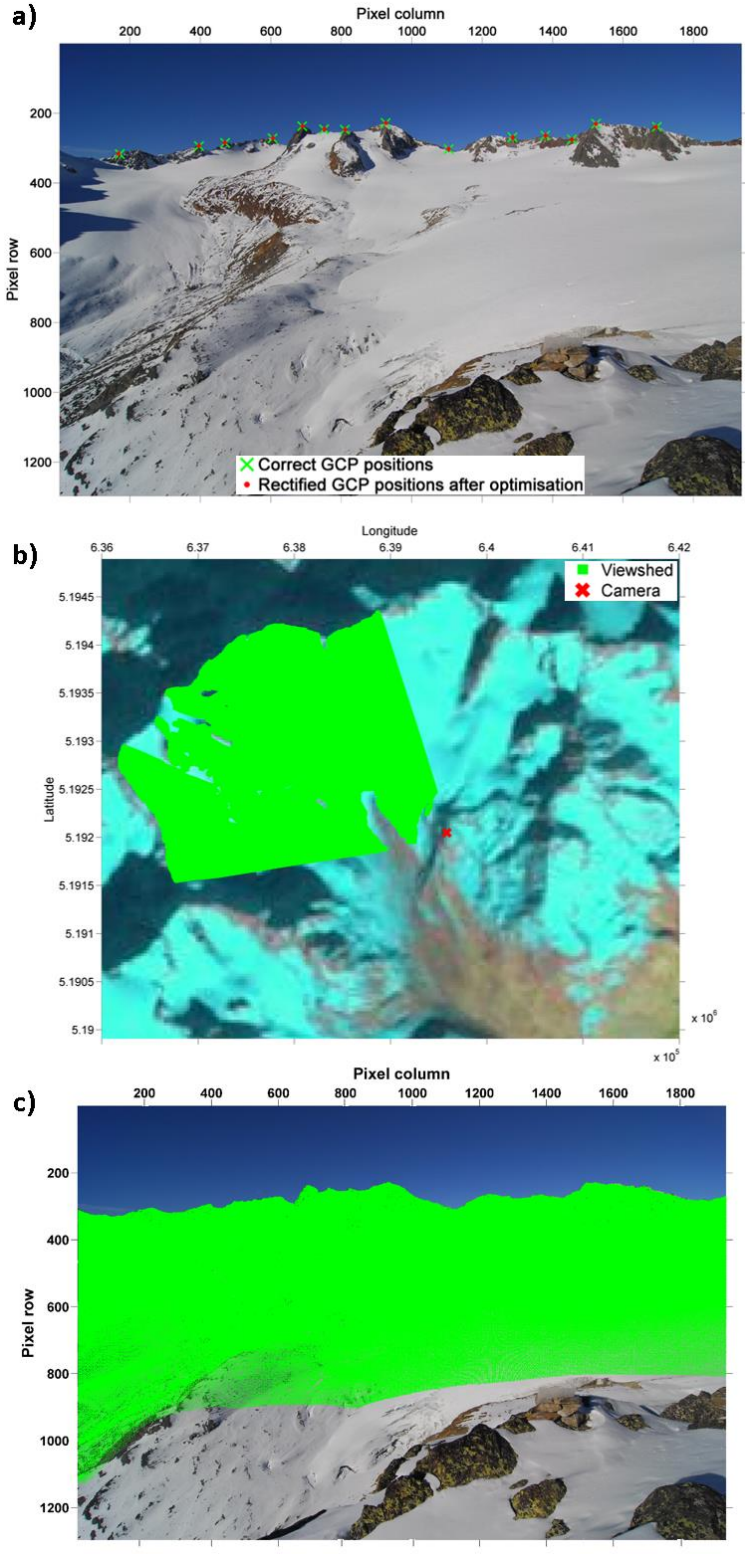
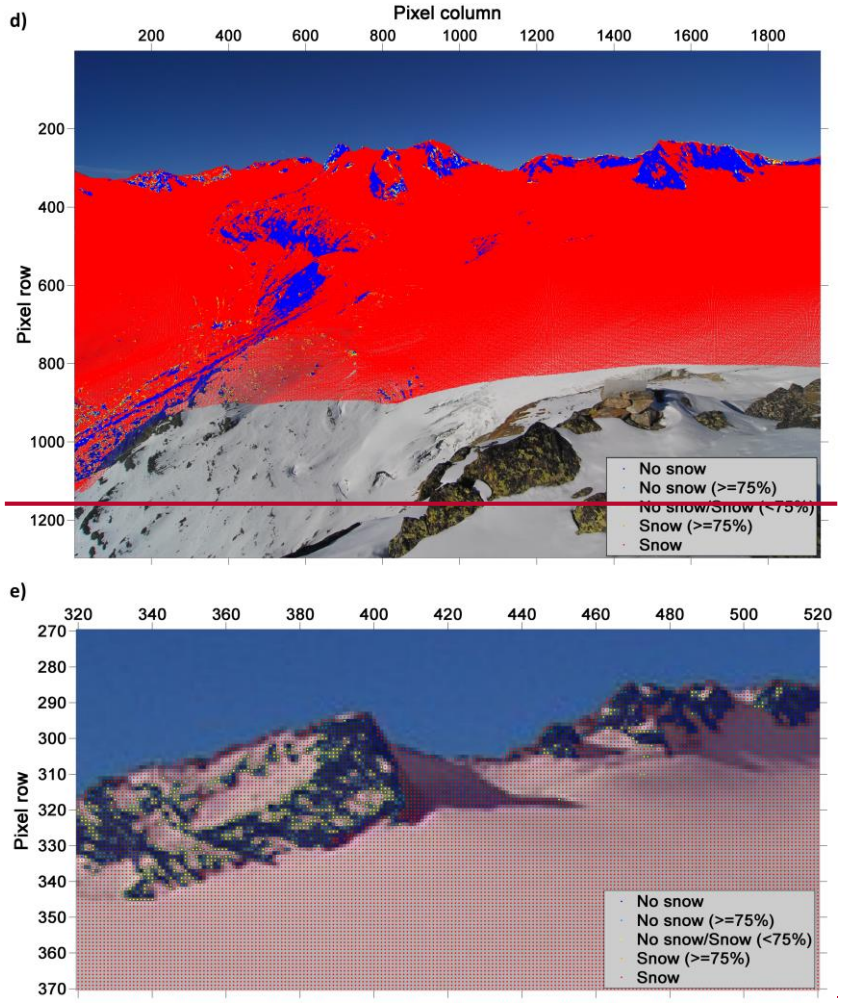


Figure 3: Schematic relationship between the camera location and orientation, and the two-dimensional photograph (blue) and three-dimensional real world coordinate system (black). The dashed line connects the locatons of three exemplary ground control points of the photograph with the real world to underline the concept.


a)
Pixel column
200
400
600
800
1000
1200
1400
1600
1800
Pixel row
Correct GCP positions
Rectified GCP positions after optimisation
b)
Longitude
6.36
6.37
6.38
6.39
6.4
6.41
6.42
Viewshed
Camera
Latitude
5.1945
5.194
5.1935
5.193
5.1925
5.192
5.1915
5.191
5.1905
5.19
x 10^6
x 10^5
c)
Pixel column
Pixel row


a)
Pixel column
200
400
600
800
1000
1200
1400
1600
1800
Pixel row
Correct GCP positions
Rectified GCP positions after optimisation
b)
Longitude
6.36
6.37
6.38
6.39
6.4
6.41
6.42
Viewshed
Camera
Latitude
5.1945
5.194
5.1935
5.193
5.1925
5.192
5.1915
5.191
5.1905
5.19
x 10^5
x 10^5
c)
Pixel column
Pixel row


d)
Pixel column
200
400
600
800
1000
1200
1400
1600
1800
Pixel row
200
400
600
800
1000
1200
No snow
No snow (>=75%)
No snow/Snow (<75%)
Snow (>=75%)
Snow
e)
320
340
360
380
400
420
440
460
480
500
520
Pixel row
270
280
290
300
310
320
330
340
350
360
370
No snow
No snow (>=75%)
No snow/Snow (<75%)
Snow (>=75%)
Snow

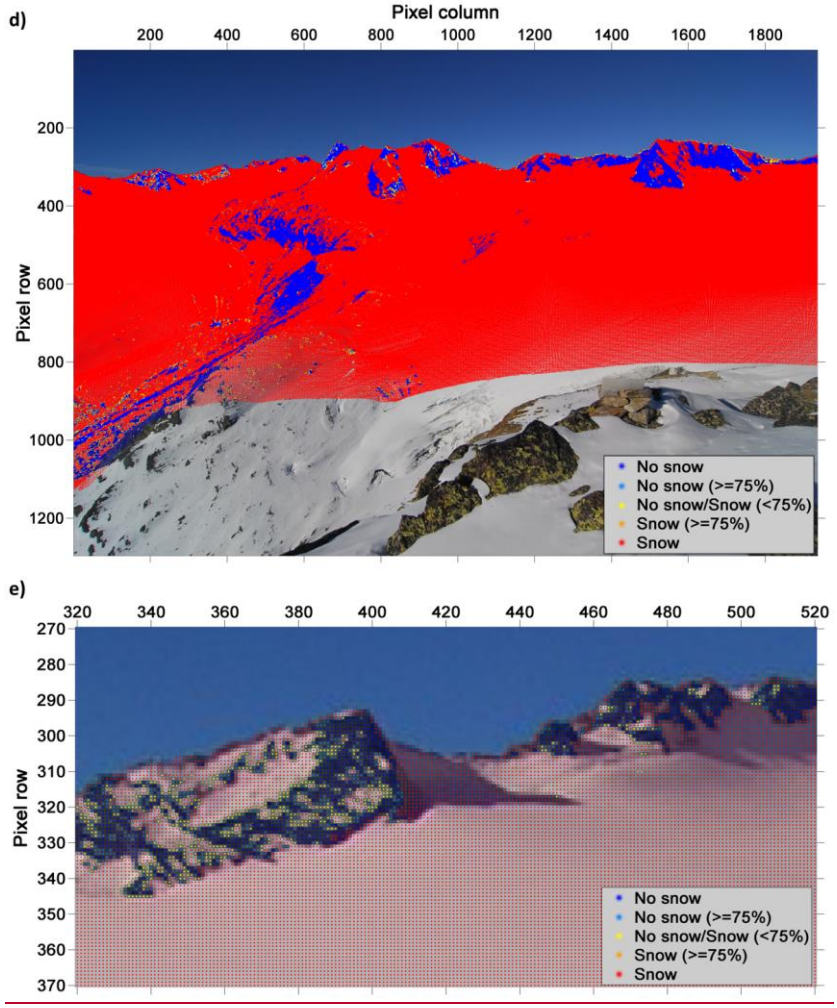


Figure 4: Internal processing steps within a single PRACTISE evaluation are shown for a photograph of VF on 17 November 2011. The figures chronologically show the routines for the photograph processing in PRACTISE which are **a)** the optimisation of the camera location and orientation using ground control points, **b)** the performed viewshed analysis from the resulting camera location and orientation, **c)** the projection and **d)** the classification of visible DEM pixels. More detail of the PCA based classification result in **d)** can be seen in an enlarged view in **e).**

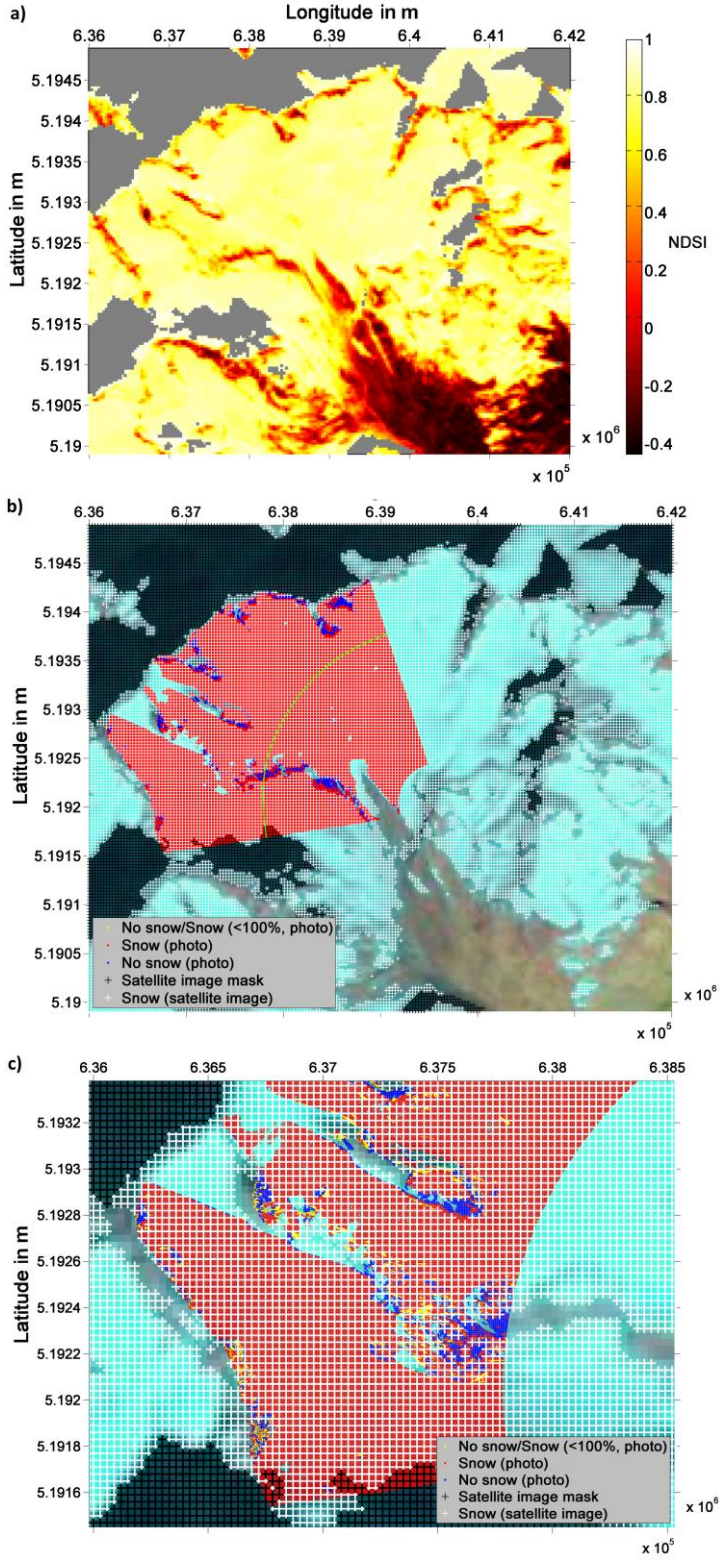


Figure 5: We outline here the internal processing steps within the remote sensing routines of PRACTISE. The Landsat ***NDSI*** ~~NDSI~~ map from 17 November 2011 is shown in **a)**. Clouds and shadows (grey areas) are excluded using fmask. The photograph and satellite snow cover map derived from the PRACTISE evaluation are superimposed on the Landsat Look image of 17 November 2011 in **b)**. Snow is depicted in red for the photograph snow map and white for the satellite snow map. The lower areas at VF (south-east of the green line in **b)**) were excluded from the complete analysis ~~as the combination of strong glacier retreat at VF and temporal difference between some analysis dates and the DEM recording dates resulted in a discrepancy of real elevations and DEM in the lower catchment areas that affected $NDSI_{thr}$ calibration results~~. The cutout in **c)** clarifies which photographed areas are part of the analysis and additionally underlines the high agreement between photograph and satellite snow cover map.

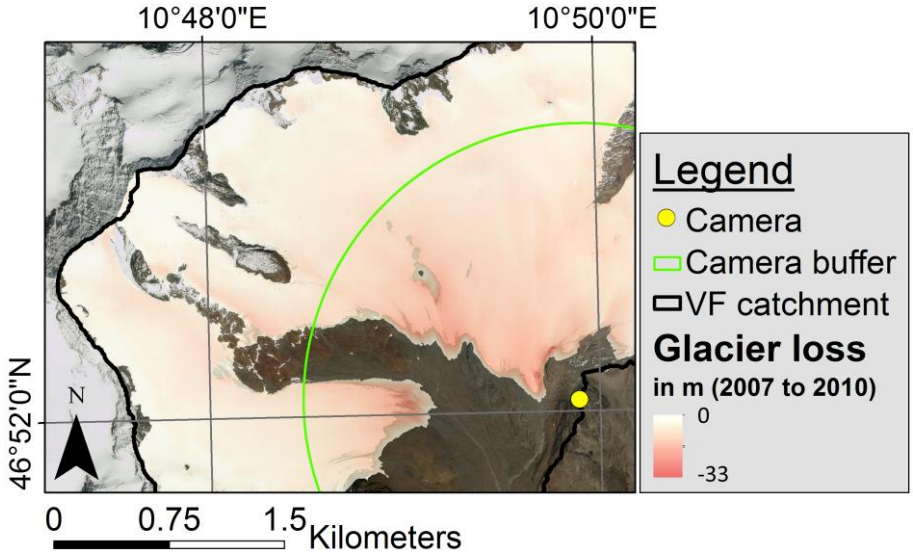


Figure 6: Glacier retreat from 2007 to 2010 causes a loss in elevation of up to -33m at VF. The green line depicts the buffer distance around the camera which was excluded from the analysis due to significant glacier loss which in turn lead to geometric inaccuracies in the photograph rectification and incorrect ***NDSI*** ~~NDSI~~ threshold calibration results.

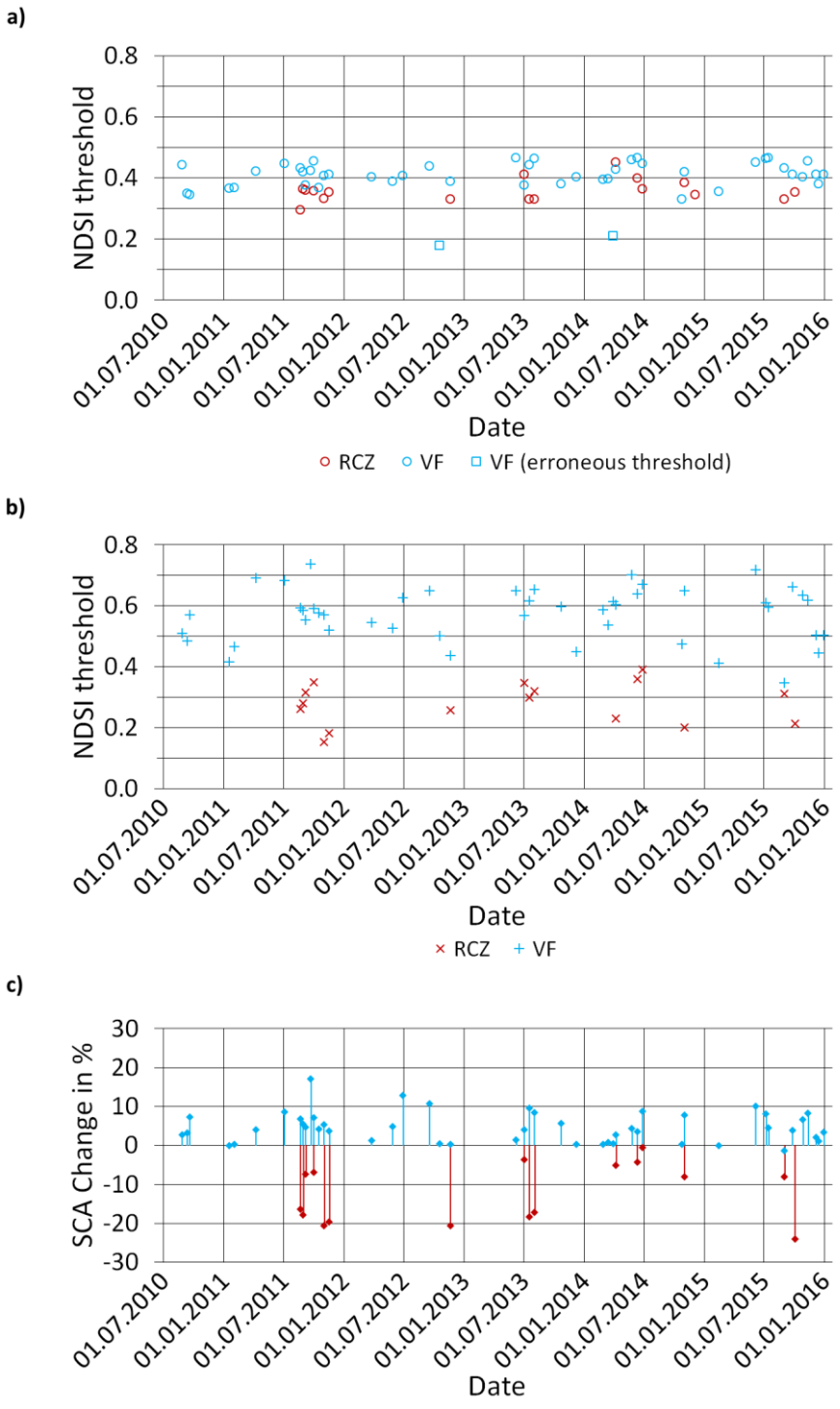


Figure 7: The Figure displays in **a)** the complete time series of adjusted ***NDSI*** ~~NDSI~~ thresholds using the Otsu segmentation method (circles, erroneous thresholds as squares) at RCZ (red) and VF (blue) and depicts in **b)** the camera calibrated ***NDSI*** ~~NDSI~~ thresholds at these two sites utili~~s~~z~~z~~ing ground-based photographs as in situ measurements (blue pluses for VF and red crosses for RCZ). Relative SCA changes at RCZ and VF resulting from the application of the standard instead of the camera calibrated reference ***NDSI*** ~~NDSI~~ threshold are shown in **c)**.

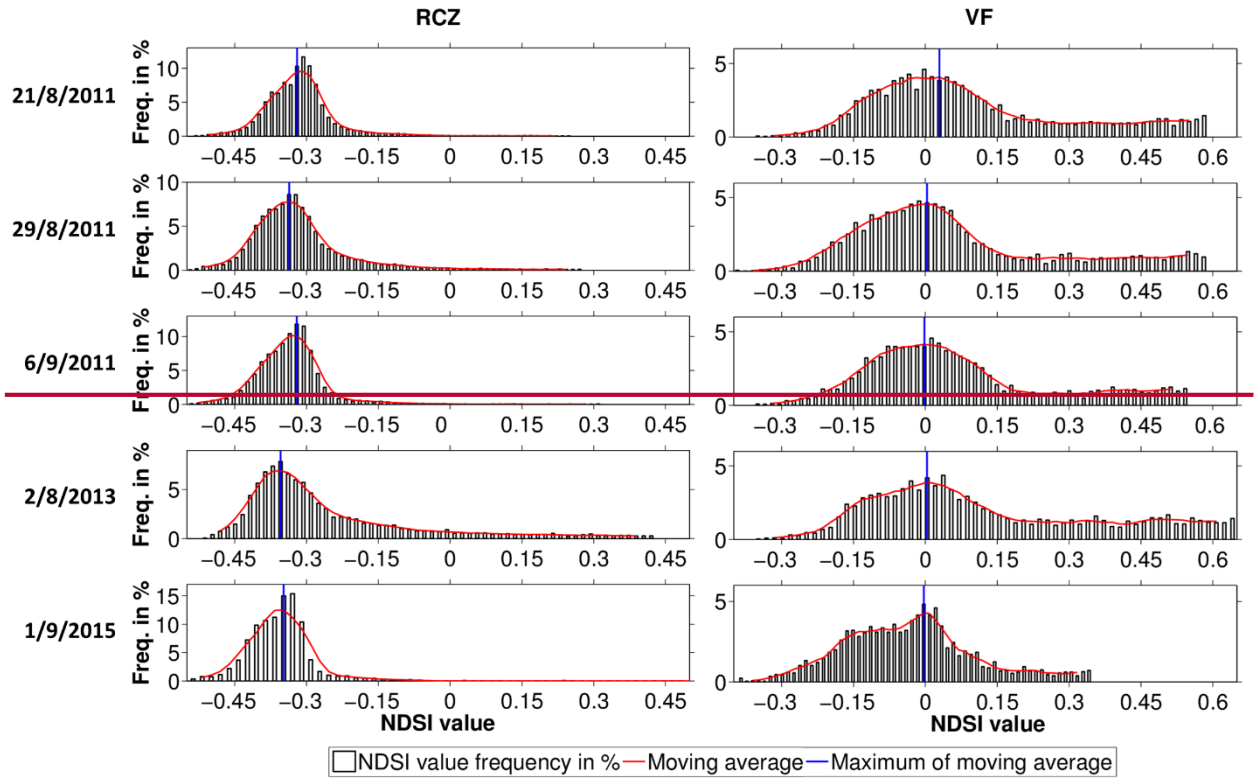


Figure 8: Representative NDSI reflectance values for the rock surfaces in RCZ and VF catchment are determined using frequency histograms of the snow-free bare rock NDSI values for five summer dates. These are then smoothed applying a moving average of 5 histogram classes. The maxima of the smoothed histograms are stable for each catchment and the investigated dates and result in mean NDSI values for rock surfaces at RCZ of -0.34 and at VF of 0.01.

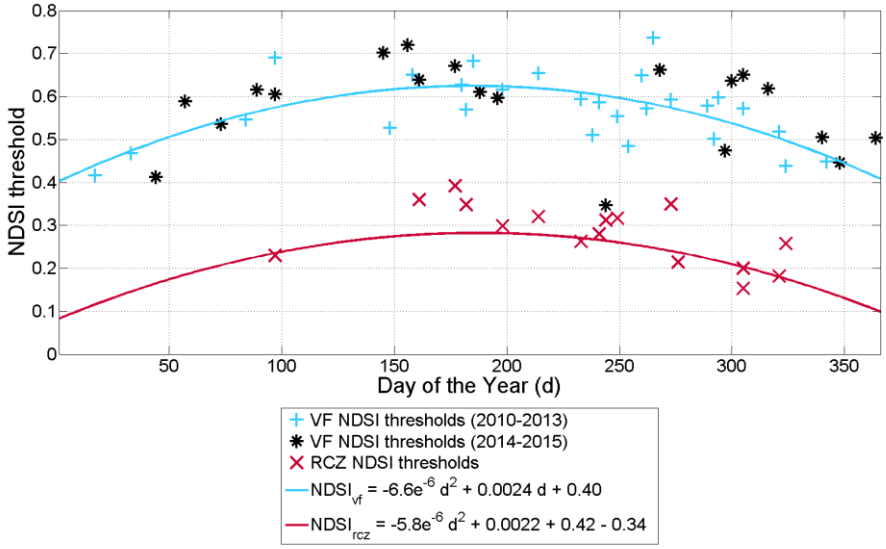


Figure ~~9~~8: Estimates of $NDSI_{thr}$~~NDSI~~ threshold values at VF are predicted ~~for each day of the year~~ by a quadratic polynomial model ($NDSI_{vf}$~~$NDSI_{vf}$~~, blue line) which was fitted for each day of the year to the calibrated $NDSI_{thr}$~~NDSI~~ thresholds between 2010 and 2013 ($NDSI_{thr}$~~$NDSI_{thr}$~~, blue pluses). ~~The coefficient of determination (r^2) of this model is 0.45 and the root mean square error (RMSE) is 0.06.~~ The black stars represent the $NDSI_{thr}$~~$NDSI_{thr}$~~ from 2014 to 2015 at VF used for evaluation of $NDSI_{vf}$~~$NDSI_{vf}$~~. Additionally, a $NDSI_{thr}$~~$NDSI_{thr}$~~ prediction model for RCZ ($NDSI_{rcz}$~~$NDSI_{rcz}$~~, red line) is defined by a quadratic polynomial model fitted to the complete time series of calibrated $NDSI_{thr}$~~$NDSI_{thr}$~~ at VF (blue pluses and black stars, r^2=0.36, RMSE=0.07) and an additional term of -0.34 ~~to account for the NDSI reflectance difference between the different rock surfaces at RCZ and VF~~. $NDSI_{rcz}$~~$NDSI_{rcz}$~~ is evaluated against the calibrated $NDSI_{thr}$~~$NDSI_{thr}$~~ of RCZ (red crosses).

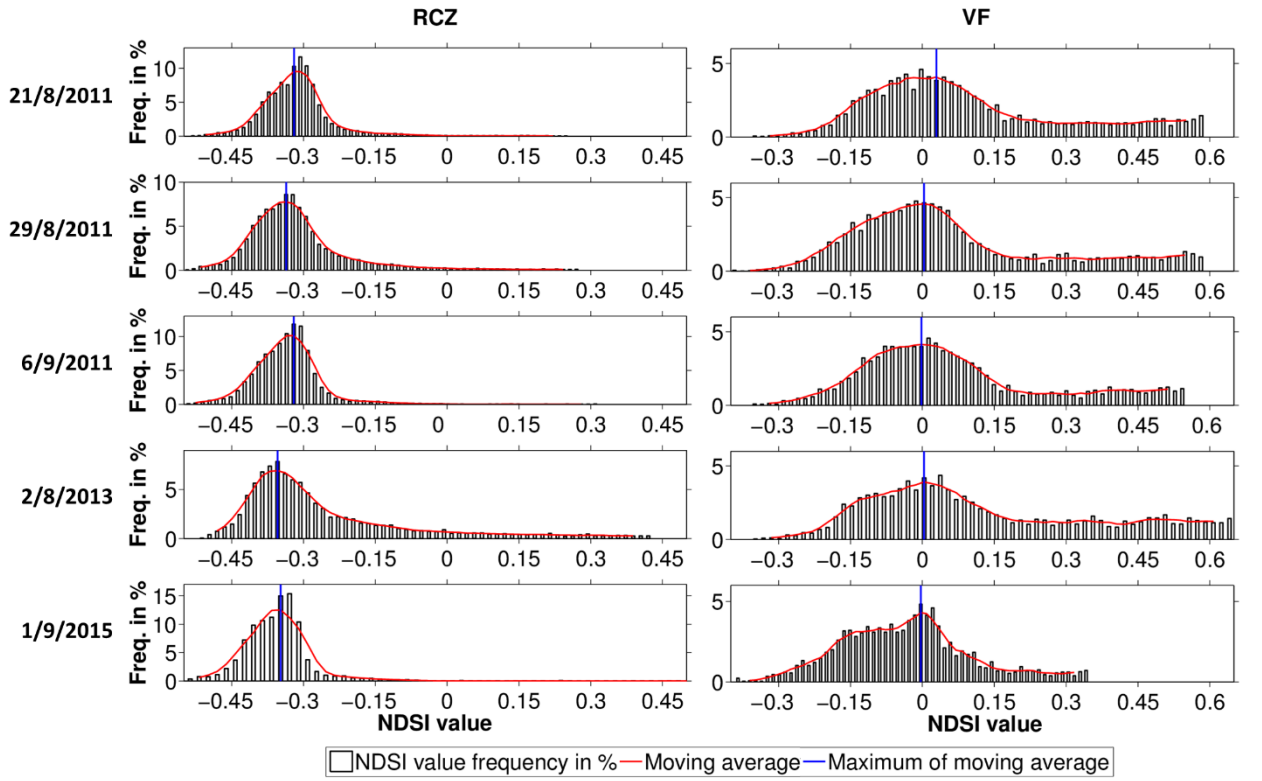


Figure 9: Representative ***NDSI*** values for the rock surfaces in RCZ and VF catchment are determined using frequency histograms of the snow-free bare rock ***NDSI*** values for five summer dates. A smoothed moving average of 5 histogram classes is shown with red. The maxima of the smoothed histograms are depicted in blue for each catchment and the investigated dates.

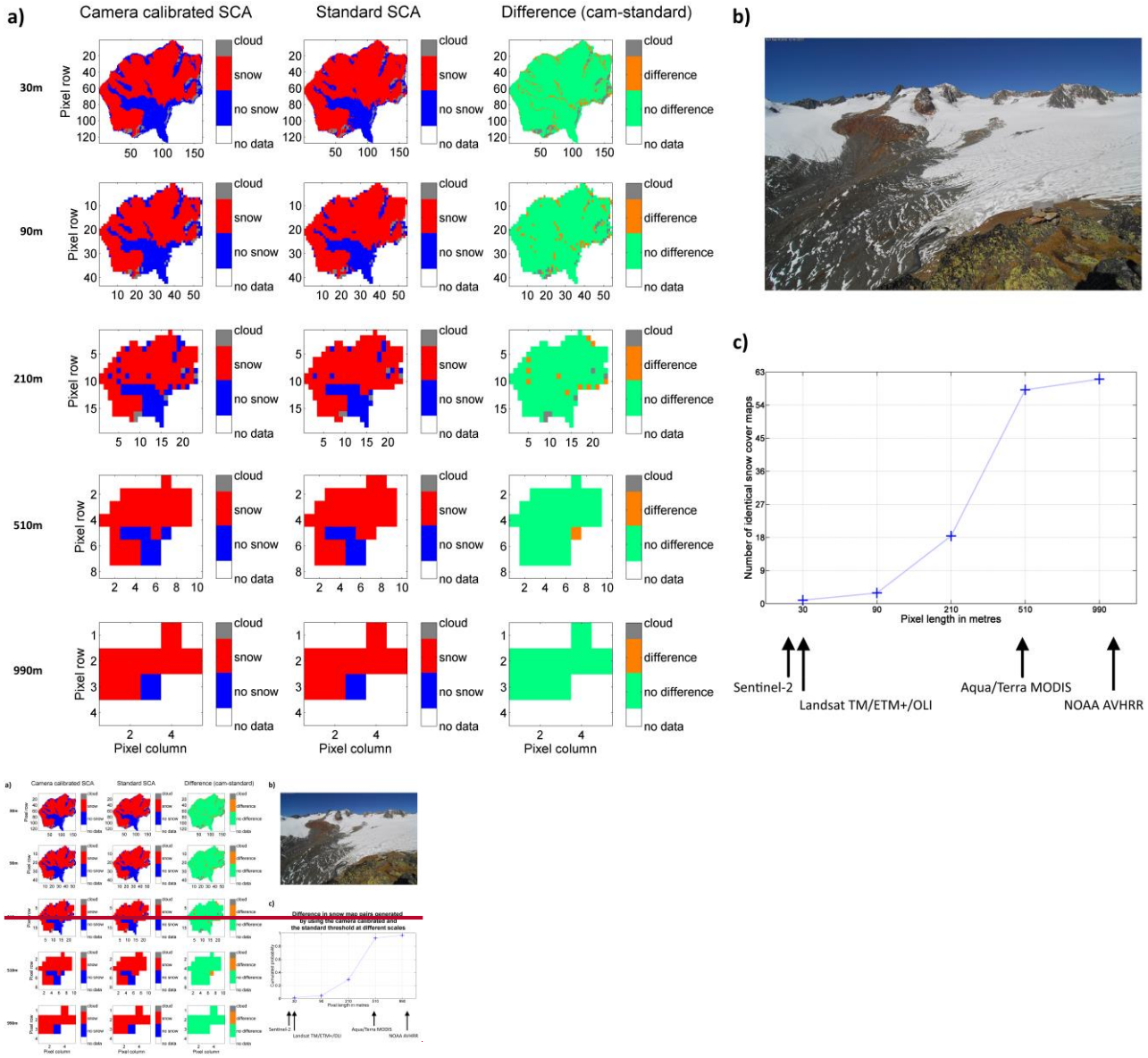


Figure 10: At VF, we exemplarily show in **a)** the effect of scaling to ***NDSI*** ~~NDSI~~ based snow cover products for a Landsat 7 scene at 16 September 2012. The ~~first~~ columns from left to right are ~~outlines~~ the camera calibrated SCA, ~~in the second column~~ the standard threshold SCA ~~is depicted~~, and ~~in the third column~~ their differences at VF ~~are presented~~. The different rows show different scaling factors, ~~starting from the top with the original resolution and a factor of~~being 1 (30 m) ~~to~~ 3 (90 m), 7 (210 m), 17 (510m) and ~~at the bottom a factor of~~ 33 (990 m) from the top to the bottom. The concurrent photograph in **b)** depicts the snow situation at VF in our example. The analysis of all investigation dates in **c)** shows at which pixel size ~~that~~ how many of the camera calibrated and standard threshold snow cover maps become ~~more and more~~ identical ~~with lower resolutions~~. The spatial resolutions of the Sentinel- 2, Landsat, MODIS and NOAA AVHRR satellites are outlined for orientation~~positive effect of the camera calibration for Landsat and presumably Sentinel-2 data thus diminishes for pixel sizes of 500 m and higher and hence for snow cover products derived from the MODIS or the AVHRR sensor~~.