Evaluating satellite retrieved fractional snow-covered area at a high-Arctic site using terrestrial photography

Kristoffer Aalstad, Sebastian Westermann, Laurent Bertino

Department of Geosciences, University of Oslo, Oslo, Norway
Nansen Environmental and Remote Sensing Center, Bergen, Norway

ARTICLE INFO

Edited by Menghua Wang
Keywords: Fractional snow-cover Optical Photography Unmixing NDSI Mixed pixel MODIS Landsat 8 Sentinel-2

A B S T R A C T

The seasonal snow-cover is one of the most rapidly varying natural surface features on Earth. It strongly modulates the terrestrial water, energy, and carbon balance. Fractional snow-covered area (fSCA) is an essential snow variable that can be retrieved from multispectral satellite imagery. In this study, we evaluate fSCA retrievals from multiple sensors that are currently in polar orbit: the operational land imager (OLI) on-board Landsat 8, the multispectral instrument (MSI) on-board the Sentinel-2 satellites, and the moderate resolution imaging spectroradiometer (MODIS) on-board Terra and Aqua. We consider several retrieval algorithms that fall into three classes: thresholding of the normalized difference snow index (NDSI), regression on the NDSI, and spectral unmixing. We conduct the evaluation at a high-Arctic site in Svalbard, Norway, by comparing satellite retrieved fSCA to coincident high-resolution snow-cover maps obtained from a terrestrial automatic camera system. For the lower resolution MODIS retrievals, the regression-based retrievals outperformed the unmixing-based retrievals for all metrics but the bias. For the higher resolution sensors (OLI and MSI), retrievals based on NDSI thresholding overestimated the fSCA due to the mixed pixel problem whereas spectral unmixing retrievals provided the most reliable estimates across the board. We therefore encourage the operationalization of spectral unmixing retrievals of fSCA from both OLI and MSI.

1. Introduction

Seasonally, snow covers between less than 10% to over 40% of the land area of the northern hemisphere, making it one of the most rapidly varying natural surface features on Earth (Hall, 1988). Given its unique combination of physical properties: high albedo, low thermal conductivity, and large water storing capacity, this seasonal snowpack strongly modulates the terrestrial surface energy and water balance in cold regions (Zhang, 2005; Dozier et al., 2016). These high latitudes and elevations are experiencing anthropogenically induced climatic warming that is amplified by positive climate feedbacks, including the snow-albedo feedback (Chapin et al., 2005; Pepin et al., 2015). A warmer climate can shift the phase of precipitation from snow to rain and lead to an earlier and slower snowmelt (Stewart et al., 2005; Musselman et al., 2017). Thereby, warming may reduce runoff efficiency and shift peak streamflow from summer and autumn, when demand is highest, to early spring and even winter (Barnett et al., 2005; Mankin et al., 2015). This is concerning given that more than 25% of the worlds population relies on mountain snowmelt for fresh water supply (Mankin et al., 2015). A shortening of the snow season due to climatic warming may also have a detrimental effect on Arctic and alpine vegetation, with implications for the albedo and carbon balance in these regions (Niittynen et al., 2018). At high latitudes, amplification accelerates the thawing of vast tracts of permafrost soils which mobilizes previously frozen carbon to the atmosphere and may further enhance global warming (Schuur et al., 2015; Chadburn et al., 2017).

Satellite remote sensing is an invaluable tool for monitoring the state of the seasonal snowpack (Bormann et al., 2018; Yilmaz et al., 2019). We are using polar-orbiting satellite sensors in this study as they provide measurements in polar regions where geostationary satellite sensors do not provide coverage. Most snow remote sensing techniques rely on the spectral signature of electromagnetic radiation induced by interactions with the snowpack. Optical sensors measuring reflected shortwave radiation in multiple bands in the visible-shortwave infrared (VSWIR) can be combined with the spectral signature of snow to retrieve snow-covered area, albedo, snow grain size, and impurity concentration (Nolin et al., 1993; Dozier et al., 2009; Painter et al., 2013) at scales from tens to hundreds of meters. Furthermore, high horizontal resolution (1–100 m) snow depth estimates can be retrieved by taking the difference of snow-covered and snow-free digital surface models.
(DSMs) obtained from stereo satellite imagery (Marti et al., 2016) or laser altimetry (Treiber and Kääb, 2017). As a caveat, cloud cover frequently prevents remote sensing of snowpack information with space-borne sensors operating in the VSWIR.

At this stage, primarily VSWIR sensors offer the spatiotemporal resolution and breadth of information needed to quantify essential snowpack properties at fine spatial scales in complex terrain (Dozier et al., 2016; Seidel et al., 2016; Bormann et al., 2018). VSWIR sensors can detect reflectance from the surface layer of the snowpack (upper 5–10 cm Dozier et al., 2009), such that individual images can be used to retrieve snow-covered area (binary or fractional), but not snow water equivalent (SWE) or snow depth. However, time series of satellite images have long been used to ingest the remotely sensed depletion of snow-covered area into snowmelt models to reconstruct SWE (Martinec and Rango, 1981; Girotto et al., 2014; Ritter et al., 2016; Dozier et al., 2016; Aalstad et al., 2018). Furthermore, satellite retrievals of snow-covered area have a wide range of uses, for example: automatic mapping of persistent ice and snow-cover (Winsvold et al., 2016; Selkowitz and Forster, 2016), investigating the snow-cover climatology (Gascoin et al., 2015; Bormann et al., 2018; Yilmaz et al., 2019), generating species distribution models (Niittynen et al., 2018), constraining permafrost models (Trofaier et al., 2017), and improving parametrizations used in Earth system models (Swenson and Lawrence, 2012).

The various applications of remotely sensed snow-cover products highlight the need to understand and characterize their uncertainty. Extensive evaluations of snow-covered area retrievals have been undertaken (Hall et al., 2002; Salomonson and Appel, 2004; Painter et al., 2009; Ritter et al., 2013; Cortès et al., 2014; Arsenault et al., 2014; Gascoin et al., 2015; Masson et al., 2018), but usually using other satellite retrievals as a reference. In this study, we retrieve reference fractional snow-covered area (ISCA) maps for a 1.77 km² area of interest in the high-Arctic, using imagery taken by an automatic camera system (ACS) deployed on a mountain top. These maps are retrieved at a high spatiotemporal resolution (0.5 m, daily during the snowmelt season) and are used to evaluate ISCA retrieved from three optical satellite sensors that are currently in polar orbit: the MODerate Resolution Imaging Spectroradiometer (MODIS) on-board the Terra and Aqua satellites, the Operational Land Imager (OLI) on-board Landsat 8, and the MultiSpectral Instrument (MSI) on-board Sentinel-2A and Sentinel-2B. In addition to evaluating existing products, such as MOD10A1 (Riggs et al., 2017) and MODSCAG (Painter et al., 2009), we also evaluate implementations of retrieval algorithms ranging from simple thresholding to fully constrained linear spectral unmixing.

2. Data and methods

2.1. Study area

The study area (see Fig. 1) is located in the Bayelva catchment on the Brøgger peninsula in north western Svalbard. The catchment is situated about 3 km west of the village of Ny–Ålesund (78°55′N, 11°50′E) which is the northernmost permanent civilian settlement in the world and serves as a major hub for polar research. The relatively warm West Spitsbergen (ocean) current ensures that this part of Svalbard has a maritime climate with mild winters and cool summers for this latitude (Boike et al., 2018). For the period 1981–2010, Ny-Ålesund received 427 mm annual average precipitation with winter, summer, and annual average air temperatures of −12.0, 3.5, and −5.2 °C (Farlend et al., 2011). At Bayelva, the climatological range of the catchment averaged peak SWE is from around 0.1 to 0.5 m w.e., the end of season snow density is typically 350 ± 50 kg m⁻³, and the snow-cover usually disappears between early June and mid-July (Aalstad et al., 2018; Boike et al., 2018). Given the high latitude, during solar noon on the summer solstice, when the sun is highest in the sky, the solar zenith angle reaches a minimum of 56° (i.e., a solar elevation angle as low as 34°). Between the 21st of October and the 20th of February, the sun is below the true horizon during polar night.

Our area of interest (AOI) is situated around the Bayelva climate station (BCS), where continuous measurements related to permafrost, snow, and hydrometeorology have been undertaken for the last two decades (e.g. Westermann et al., 2011, 2012; Aalstad et al., 2018; Boike et al., 2018). The area features elevations between 5 and 55 m a.s.l., and gently undulating topography with small hills (see Fig. 2), causing large differences in snow-cover due to wind drift. The surface cover alternates between bare soil (silty loam and silty clay), rocks (shale, sandstone, and coal), and low lying vegetation (mosses, lichens, and grasses). The AOI is located between two mountains, Zeppelinfjellet (556 m a.s.l.) to the south east and Scheteligfjellet (719 m a.s.l.) to the west on which the ACS is mounted. To the south the AOI is bordered by the two branches of the Brøggerbreen glacier that feeds the Bayelva river. The AOI highlighted in Fig. 1 spans most of the Bayelva catchment, covering an area of around 1.77 km² between Scheteligfjellet and Ny–Ålesund.

2.2. Terrestrial photography

Here, we describe the processing chain for the reference images obtained from terrestrial photography. The entire chain is shown schematically in Fig. 3. For all snowmelt seasons (May to August) from 2012 to 2017, an automatic time lapse camera system (ACS) was deployed near the summit of Scheteligfjellet (c.f. Fig. 1), overlooking the Bayelva site. A Canon EOS 1100D digital single-lens reflex camera was triggered by a Harbotronics time-lapse system, delivering high-quality daily images over the study area except for days with prolonged low-cloud cover or system malfunction. For each day, the photograph with the most favorable illumination conditions, i.e. close to solar noon and lack of cloud or other shadows, was selected. After filtering out low quality images, in total 305 terrestrial images from different days were available for further analysis.

A reference image and DEM were employed to georeference and orthorectify the images. We used an orthophoto and DEM derived from the airborne high resolution stereo camera (HRSC-AX) mission flown over the Brøgger peninsula around noon on the 17th of July 2008 (see Boike et al., 2018, and references therein) with a ground sampling distance (GSD) of 0.5 m. The DEM and associated true color image, cropped to our AOI, is displayed in Fig. 2, whereas the full HRSC true color image used for georeferencing is shown in Fig. 3.

For this study, we employ orthorectified images similar to the ones available through Westermann et al. (2015) which were already used to evaluate satellite retrievals of ISCA in a data assimilation experiment (Aalstad et al., 2018). The orthorectification procedure relies on a range of natural ground control points visible in both the camera images and the HRSC-AX orthophoto. We then fit the parameters of a simple camera model (Bouquet, 2015), while the position of the ACS was derived from Differential Global Positioning System (DGPS) measurements. For this study, we applied this standardized orthorectification process to the images from the snowmelt seasons in all the years 2012–2017. The AOI was chosen to exclude the outer edges of the orthorectified images where significant distortions occur.

We conducted an independent evaluation of the georeferencing, comparing the location of 19 landmark features (different to the ones used as ground control points in the orthorectification) in the reference (HRSC-AX) image to those in the orthorectified terrestrial photographs. For each snowmelt season, we randomly selected 5 images, such that a subset consisting of 6 × 5 = 30 images was used to diagnose the georeferencing error. Due to the snow-cover, it was not always possible to identify all the landmarks in an image. In such cases the obscured landmarks were ignored. Based on this exercise, we obtained a root mean squared error (RMSE) of 2.25 m and a bias of 1.85 m for the georeferencing in the orthorectified terrestrial photographs. An example of the georeferencing error present in the images is shown in Fig. 3 (c). By separating the georeferencing error into its directional
components, we found that the georeferencing biases in the northing and easting direction were quite small (0.16 m and 0.59 m, respectively) indicating no systematic shifts. Using the easting coordinate as a proxy for the distance from the camera, given that the camera is situated due west of the AOI, we found a negligible correlation (~0.11) with the georeferencing error.

Next, the orthorectified and georeferenced images were cropped to the AOI (c.f. Fig. 3), which is not exactly rectangular to exclude the runway at Ny-Ålesund airport. Due to the frequent plowing of snow on this runway it was excluded from the AOI as it could lead to undesirable situations where a satellite image was captured before and the terrestrial image after the plowing, or vice versa. Other than the runway, there are no significant anthropogenic disturbances within the AOI.

We manually classified the ~7 × 10^6 pixels in each cropped image in two categories, snow-covered and snow-free, using a simple image specific threshold in the blue band of each image. This method has been shown to perform reasonably well for much larger areas than our AOI (Fedorov et al., 2016). The threshold was chosen iteratively for each image until a good visual agreement was found between the snow-cover in the true color image and the resulting classification. In most cases, (see Fig. 3) the blue band histogram exhibited clear bimodality, and the threshold could easily be chosen as the local minimum between the snow peak and snow-free peak in the histogram. We gauged the sensitivity of the reference snow-cover maps to the choice of thresholds by varying the selected thresholds by ±5% for each of the aforementioned 30 randomly selected images. For an increase of 5%, we found a mean increase in fSCA of 0.01 and a maximum increase of 0.02. For a decrease of 5%, we also found a mean decrease in fSCA of 0.01 with a maximum decrease of 0.04. This suggests that the reference snow-cover maps are relatively insensitive to small variations in the selected thresholds. Three images contained sunglint over areas with surface water, which made it impossible to pick a valid threshold for these images. Consequently, these three sunglinted images were excluded from further analysis, resulting in a total of 302 classified images that we use as a ground truth for evaluating various satellite retrievals of fSCA.

As a final step, to gauge the uncertainty in our ground truth data, we translated the georeferencing error to a potential error in the reference fSCA maps using geostatistical analysis (see e.g. Wackernagel, 2003). For each reference snow-cover image, we calculated the fSCA for...
100 × 100 m moving windows using a 2D digital filter. We then computed the empirical variogram as a function of separation distance. Based on these variograms, assuming isotropic and stationary conditions, we could diagnose the estimation standard deviation (σ_E) for the reference fSCA as a function of separation distance for each image. Considering all the 302 images, for a separation distance equal

![Image](image_url)
Table 1
Overview of the fSCA retrievals that we evaluated. fSCA retrievals that are not operational products were performed on reflectances only as a part of this study.

| Abbreviation | Description | Sensor | Product | Reflectance |
|--------------|-------------|--------|---------|-------------|
| FT-MSI-TOA  | Fixed threshold | S2A/B MSI | No | TOA |
| AT-MSI-TOA  | Adaptive threshold | S2A/B MSI | No | TOA |
| FT-MSI-SFC  | Fixed threshold | S2A/B MSI | No | Surface |
| AT-MSI-SFC  | Adaptive threshold | S2A/B MSI | No | Surface |
| SLC-MSI     | Scene classification | S2A/B MSI | Yes | Surface |
| SU-MSI      | Spectral unmixing | S2A/B MSI | No | Surface |
| RG-MSI-TOA | Regression | S2A/B MSI | No | TOA |
| RG-MSI-SFC | Regression | S2A/B MSI | No | Surface |
| FT-OLI-TOA | Fixed threshold | LS8 OLI | No | TOA |
| AT-OLI-TOA | Adaptive threshold | LS8 OLI | No | TOA |
| FT-OLI-SFC | Fixed threshold | LS8 OLI | No | Surface |
| AT-OLI-SFC | Adaptive threshold | LS8 OLI | No | Surface |
| Fmask-OLI  | Scene classification | LS8 OLI | Yes | Surface |
| SU-OLI     | Spectral unmixing | LS8 OLI | No | Surface |
| RG-OLI-TOA | Regression | LS8 OLI | No | TOA |
| RG-OLI-SFC | Regression | LS8 OLI | No | Surface |
| MOD10A1-v5 | MOD10A1 version 5 | Terra MODIS | Yes | TOA |
| MOD10A1-v6 | MOD10A1 version 6 | Terra MODIS | Yes | TOA |
| MYD10A1-v5 | MYD10A1 version 5 | Aqua MODIS | Yes | TOA |
| MYD10A1-v6 | MYD10A1 version 6 | Aqua MODIS | Yes | TOA |
| MODSCAG    | MOD09GA v5 unmixing | Terra MODIS | Yes | Surface |
| SU-MOD09GA | MOD09GA v6 unmixing | Terra MODIS | No | Surface |
| SU-MYD09GA | MYD09GA v6 unmixing | Terra MODIS | No | Surface |

The georeferencing RMSE we obtained a worst case $\sigma_E \approx 0.01$, and for a separation distance equal to three times the georeferencing RMSE we obtained a worst case $\sigma_E \approx 0.03$.

2.3. Satellite retrievals

In this study, we employed fSCA retrievals from multiple satellite sensors using several different retrieval algorithms. The retrievals were both in the form of already processed fSCA products and our own retrievals obtained through algorithms of varying complexity. Here, we outline both the satellite platforms, sensors, and products as well as the retrievals algorithms. An overview of all the retrievals evaluated in this study is provided in Table 1.

2.3.1. MODIS

The MODIS instrument that is currently in orbit on the Terra (2000–present) and Aqua (2002–present) satellite platforms provides multispectral imagery in the VSWIR wavelength range at a daily revisit period and a GSD of approximately 500 m (for all the VSWIR bands). In this study, we used several MODIS products, namely: both collection/ version 5 and 6 of the NASA snow-cover products MOD10A1 (Terra) and MYD10A1 (Aqua) distributed by NSIDC (Hall et al., 2006a,b; Hall and Riggs, 2016a,b), the MODIS Snow Covered Area And Grain size (MODSCAG) product from Caltech/Jet Propulsion Laboratory (JPL) (Painter et al., 2009), and the NASA surface reflectances products MOD09GA (Terra) and MYD09GA (Aqua) distributed by LP DAAC (Vermote et al., 2000–present). The MOD10A1, MYD10A1, and MODSCAG products are used as is, while the MOD09GA and MYD09GA products are used as inputs for the unmixing described in Section 2.3.5. For all the MODIS products, we use the five pixels in tile h18v01 that fall completely within the AOI (see Fig. 1).

2.3.2. Landsat 8 OLI

The OLI sensor on-board Landsat-8 has been in orbit since February 2013 and delivers multispectral imagery in the VSWIR at a GSD of 30 m with a revisit period of 16 days at the equator. For this study, we used Landsat-8 OLI level 1 top of the atmosphere (TOA) reflectances and level 2 atmospherically corrected surface reflectances (Vermote et al., 2016; USGS, 2018), both distributed by USGS EROS through the EarthExplorer online user interface. Both the TOA and surface reflectances were used in NDSI thresholding and regression algorithms, while only the latter surface reflectances were used for spectral unmixing. Some scenes were not available in the correct UTM zone for our AOI (33X). In these cases, we repurposed the scenes to zone 33X and performed a nearest neighbor interpolation to a regular grid over our AOI. The level 2 product includes the pixel land-cover classification from the Fmask algorithm (Zhu and Woodcock, 2012; USGS, 2018) which we also evaluate in terms of snow-cover mapping performance. Fmask is based on a thresholding of the NDSI, in addition to many other tests to diagnose the pixel land cover and cloud status. We simplified the Fmask classification to be more relevant for snow-cover mapping as follows:

- Pixels classified as cloud shadows, terrain occluded, and/or saturated were reclassified as ‘missing’.
- Pixels classified as high confidence cloud (i.e. all cloud bits equal to 1) were reclassified to ‘cloudy’.
- Pixels that were not ‘missing’ or ‘cloudy’ but with snow bit equal to 1 were reclassified as ‘snow’ pixels.
- All remaining pixels were reclassified as ‘snow-free’ pixels.

For the Landsat 8 OLI imagery we manually selected cloud free scenes over our AOI. This resulted in a total of 26 scenes with accompanying reference snow-cover maps for the ablation seasons (May–July) from 2013 to 2017.

2.3.3. Sentinel-2 MSI

The MSI sensor is currently in orbit on-board Sentinel-2A (2015–present) and Sentinel-2B (2017–present) delivering VSWIR multispectral imagery at a GSD of 10–20 m with a revisit frequency of 5 days at the equator (Drusch et al., 2012). In our study, we used two products from Sentinel-2 MSI: level 1 (L1C) TOA reflectances and level 2 (L2A) atmospherically corrected surface reflectances distributed by the Copernicus Open Access Hub. Since not all L2A scenes had been processed for Svalbard yet, particularly in 2016, we performed conversions from L1C to L2A using the Sen2Cor python script (Müller-Wilm, 2018). No terrain correction was applied since the DEM required by Sen2Cor was not available for our AOI, but we do not expect this to adversely influence the results given that the AOI is relatively flat. As with Landsat 8 OLI, both the TOA and surface reflectances were used for NDSI thresholding and regression retrieval algorithms, while only the surface reflectances were used in spectral unmixing. Scenes that were not available in the correct UTM zone were reprojected and resampled in the same way as for the Landsat 8 OLI products. We also resampled the SWIR bands at 20 m GSD to be in line with the other bands at 10 m GSD using nearest neighbor interpolation. In addition to the L1C and L2A reflectances, we employed the scene land cover classification from the L2A product (“L2A_SceneClass”, henceforth SLCC-MSI) described in Richter et al. (2012) in our evaluation. The SLCC-MSI is also based on a thresholding of the NDSI, in addition to many other tests to diagnose the pixel land cover and cloud status. We simplified the SLCC-MSI classification to be more relevant for snow-cover mapping in the following way:

- Shadow, saturated, dark area, cloud shadows, and unclassified pixels were reclassified as ‘missing’ pixels.
- Cloud medium probability, cloud high probability, and thin cirrus pixels were all reclassified as ‘cloudy’ pixels.
- Snow pixels remained ‘snow’ pixels.
- Vegetation, non-vegetation, and water pixels were reclassified as ‘snow-free’ pixels.

For the Sentinel-2 MSI imagery we manually selected cloud free scenes over our AOI. This resulted in a total of 24 Sentinel-2 scenes with accompanying reference snow-cover maps for the 2016 and 2017
2.3.4. Normalized difference snow index

The NDSI is the basis for the simpler snow-cover retrievals evaluated in this study. It exploits the fact that snow is highly reflective in the visible but a strong absorber in the shortwave infrared, which differentiates snow from most other natural surfaces (Dozier et al., 2009). The NDSI is defined as (Dozier, 1989)

\[
\text{NDSI} = \frac{r_{\text{Green}} - r_{\text{SWIR1}}}{r_{\text{Green}} + r_{\text{SWIR1}}},
\]

where \(r_{\text{Green}}\) is reflectance in the green band and \(r_{\text{SWIR1}}\) is the first shortwave infrared reflectance band (typically around 1.6 \(\mu\)m). In general, these may be either TOA or surface reflectances.

For version 6 of the MOD10A1 & MYD10A1 products the NDSI (based on TOA reflectances) is used when assigning a confidence on within-pixel snow presence through the ‘NDSI snow-cover’ field (Riggs et al., 2017). For pixels for which this NDSI snow-cover field is defined, the fSCA can be obtained through a simple linear relationship of the form

\[
\text{fSCA} = \beta_1 \cdot \text{NDSI} + \beta_0,
\]

where the \(\beta\) are regression coefficients. Eq. (2) is then subject to the physical constraints \(\text{fSCA} \in [0,1]\). In our study, we use the ‘universal FRA6T’ relationship from Salomonson and Appel (2004), where \(\beta_1 = 1.45\) and \(\beta_0 = -0.01\). This relationship was developed by correlating MODIS NDSI with fSCA obtained from coincident Landsat 7 Enhanced Thematic Mapper-Plus scenes (Salomonson and Appel, 2004). Note that employing separate relationships for Aqua MODIS is not necessary for version 6, since Aqua band 6 has been restored to scientific quality through the quantitative image restoration technique (Riggs et al., 2017).

For version 5 of the MOD10A1 & MYD10A1 products, fSCA is provided directly as a field based on separate linear relationships for Aqua and Terra (Salomonson and Appel, 2004). In addition to the MODIS products, we also apply the linear regression in Eq. (2) to Sentinel-2 MSI and Landsat 8 OLI derived NDSI to gauge the applicability of this relationship at higher resolution.

The NDSI can also be used directly to obtain binary snow cover maps. This is especially appealing for the higher resolution sensors on board Landsat 8 and the Sentinel-2 satellites, since the structure of individual snow patches is more readily resolved which moderates the problem of mixed pixels. To binarize pixels by classifying them as either bare or snow-covered, a threshold for the NDSI is required. Based on experience, several studies suggest a fixed threshold of 0.4 (e.g. Hall et al., 2002; Gascoin et al., 2015; Aalstad et al., 2018) which is often sufficiently accurate to discriminate snow and non-snow covered pixels. On the other hand, unsupervised adaptive detection algorithms can determine scene-specific thresholds that may yield superior results (Yin et al., 2013; Härer et al., 2018). In this study, we compare both techniques, using a fixed (0.4) and Otsu’s segmentation algorithm (Otsu, 1979) as an adaptive thresholding technique which performed best in the intercomparison of Yin et al. (2013). For an (adaptive or fixed) NDSI threshold \(\tau\), we define the binary snow-cover status of a pixel for the Sentinel-2 and Landsat 8 reflectance products through the following rules:

\[
\text{fSCA} = \begin{cases} 
1 & \text{if NDSI} > \tau \text{ and } r_{\text{Red}} > 0.12 \text{ and } r_{\text{SWIR1}} < 0.16, \\
0 & \text{otherwise},
\end{cases}
\]

where we again use either TOA or surface reflectances. The additional thresholds in Eq. (3), applied to the reflectances in the red (\(r_{\text{Red}}\)) and SWIR1 (\(r_{\text{SWIR1}}\), around 1.6 \(\mu\)m) bands are based on Dozier (1989) and Gascoin et al. (2015). The condition on the red band accounts for self, cast, or cloud shadows, while the threshold on the SWIR1 band helps to distinguish clouds (that may have a high NDSI) from snow. In agreement with Dozier (1989), the exact values of these individual band thresholds is not critical and we found little sensitivity to small (\(\pm 0.05\)) variations. Once Eq. (3) has been evaluated on a pixel level, the results are aggregated through spatial averaging to a desired coarser resolution to obtain the fSCA.

2.3.5. Spectral unmixing

fSCA retrieved using spectral unmixing (SU) of multispectral satellite imagery, reviewed in Keshava and Mustard (2002), was also employed in this study. The MODSCAG product detailed in Painter et al. (2009) applies unmixing to version 5 of the MOD09GA (Vermote and Wolfe, 2015a) product, and we base our unmixing procedure loosely on the MODSCAG algorithm. Other implementations of unmixing for snow-cover mapping include Vikhamar and Solberg (2003), Sirgury et al. (2009), and Cortés et al. (2014). In our implementation of spectral unmixing we solve the fully constrained least squares problem, i.e. we seek an optimal abundance \(a^s\), which satisfies

\[
a^s = \arg \min \| \mathbf{M} a^s - b^2 \|^2 \quad \text{s.t.} \quad a^s \geq 0 \quad \text{and} \quad 1^T a^s = 1,
\]

where \(\| \cdot \|\) is the \(L^2\)-norm, the \(N_b \times N_e\) (number of bands times number of endmembers) matrix \(\mathbf{M}\) contains the theoretical reflectances for each endmember (columns) averaged to the bands of the sensor in question (rows), the \(N_e \times 1\) vector \(a^s\) contains the fractional abundances of each endmember, the \(N_b \times 1\) vector \(\mathbf{b}\) contains the retrieved surface reflectances in each band, and \(1^T\) is a \(1 \times N_e\) vector of ones. The physical constraints on the fractional abundances of the endmembers are the abundance non-negativity constraint (ANC: \(a^s \geq 0\)) and the abundance sum-to-one constraint (ASC: \(1^T a^s = 1\)). The conversion from spectral reflectance, \(r(\lambda)\), to band reflectance, \(r^b\), is carried out through the following averaging operation:

\[
r^b = \int_0^\infty r(\lambda) \phi_0(\lambda) d\lambda,
\]

where \(\phi_0(\lambda)\) is the sensor specific spectral response function (normalized to a probability density) for band \(b\). In the SU experiments, we use the following six bands \(b \in \{\text{Blue, Green, Red, NIR, SWIR1, SWIR2}\}\) for all the sensors. As an example, the spectral response functions for the bands employed for MSI on-board Sentinel-2A are shown in Fig. 4 (a). In case of ambiguity, we use the band with the higher spatial (as opposed to radiometric) resolution, e.g. for MSI we employ near-infrared band 8 and not band 8a (see Müller-Wilm, 2018).

In practice, Eq. (4) can be solved using the fully constrained least squares unmixing approach described in Heinz and Chang (2001). This approach first augments both \(\mathbf{M}\) and \(\mathbf{b}\) with an extra row of ones to deal with the ASC and subsequently employs the non-negative least squares algorithm of Lawson and Hanson (1995) to enforce the ANC. In our study, theoretical spectra for the non-snow endmembers were obtained from the JPL spectral library (Baldrige et al., 2009). In particular, the non-snow endmembers employed were: vegetation (‘grass’), soil (‘pale brown dry silty clay loam’), and rock (‘gray sandstone’). These non-snow endmembers were chosen based on knowledge of the typical land cover of the study area (see Boike et al., 2018) and are visualized in Fig. 4 (b). We also added a shade endmember (zero reflectance in all bands) to account for shadows in the unmixing procedure. The snow endmember spectra were obtained from a look up table generated by running the SNICAR radiative transfer model (He et al., 2018) with varying solar zenith angles (bins of 5°) and 7 different effective snow grain radii in the range 100–5000 \(\mu\)m. For simplicity, light absorbing impurities in snow are ignored in this study. After fixing the solar zenith angle to the nearest value in the look up table, we looped over all possible grain radii and selected the one that minimized the root mean square residual error in Eq. (4). Some example snow endmember spectra are shown in Fig. 4 (b). As in Cortés et al. (2014), we ignore azimuthal effects since, despite the strong forward scattering properties of snow, these should not be significant for near nadir looking sensors on-board the Landsat 8 and Sentinel-2 satellites. For the off-nadir looking MODIS sensor this assumption may be more problematic. After obtaining the optimal \(a^s\) across all radii, the fSCA was obtained by
dividing the snow endmember abundance by the total unshaded abundance. The assumption in this normalization is that the relative abundances are the same in the shaded and non-shaded regions. All the non-snow endmembers and a subset of the snow endmembers modeled through SNICAR for different effective snow grain radii $r_e$ and solar zenith angles $\theta_0$.

2.3.6. Shadow masking

The fSCA retrieval algorithms do not work well if a considerable fraction of a satellite pixel is covered in shadow (see Fig. 5 (b)). For the higher resolution (OLI & MSI) sensors we diagnosed cast and self shadowing for each scene using a dynamic shadow mask. This was achieved in a two stage process. First the solar geometry (solar azimuth and elevation angle) was diagnosed for the AOI at the scene sensing time using the ephemeris routine in the PV_LIB toolbox (SNL, 2014). Our AOI is small enough that the solar geometry is invariant to a good approximation. The geometry was then used in conjunction with the TopoToolbox (Schwanghart and Scherler, 2014) and a 5 m GSD DEM of the entire Brøgger peninsula produced by the Norwegian Polar Institute (NPI, 2014) to calculate the self and cast shadows within the AOI. To account for uncertainties in the ephemeris and the employed DEM, we dilated the resulting shadow mask by one DEM pixel. The shadow mask is thus conservative in that it purposely slightly overestimates the shadowed area. For each satellite scene the binary shadow mask at the DEM resolution was aggregated to obtain fractional shadows at the resolution of the satellite image. For some satellite scenes, Scheteligfjellet casts a shadow over the AOI, so it was important to mask this out in an automatic fashion without having to discard the entire scene. An example of such a shadow and independently derived shadow mask is shown in Fig. 5. Cast shadows did not affect any of the reference fSCA maps derived from terrestrial imagery.

2.3.7. Aggregation

All the higher resolution retrievals from MSI and OLI are aggregated to coarser pixels at 100 m and 500 m resolution. The aggregation is performed in a conservative manner using a form of inverse distance weighting. For each coarse pixel, subpixels that either partially or fully fall within this pixel are weighted according to the fraction of their area that lies inside the coarse pixel. If more than 25% of a coarse pixel consists of missing data (e.g. if it has a shade fraction > 25%), the pixel is excluded from further analysis. A scale of 100 m was chosen in line with typical reanalysis experiments (e.g. Girotto et al., 2014; Cortés et al., 2016; Aalstad et al., 2018), since this scale has been shown to resolve the key modes of snow spatial variability. The scale of 500 m was chosen in order to compare the performance of the higher resolution retrievals with the MODIS retrievals. As an example, Fig. 6 visualizes how well some of the satellite retrieval algorithms for Sentinel-2 MSI can replicate the coincident reference snow cover prior to aggregation. For the regression and SU experiments that (unlike thresholding) provide fSCA without the need for spatial aggregation, we also performed the evaluation at the native pixel scale of the sensors.

2.4. Evaluation metrics

We perform the evaluation by pairing coincident (i.e. same day and spatial area) pixels from the satellite fSCA retrievals (the estimate) and the reference fSCA retrievals (the ‘truth’). The number of sample pairs is
different depending on the type of satellite retrieval. For example, there are many more MODIS scenes than there are Sentinel-2 scenes, so the number of MODIS samples at 500 m resolution is considerably larger. At this resolution, the number of sample pairs for Sentinel-2 MSI (144) is nonetheless more than enough for a robust evaluation. The estimate-truth pairs are used to compute various evaluation metrics, most of which are typical of earlier fSCA evaluation studies (e.g. Rittger et al., 2013; Gascoin et al., 2015; Masson et al., 2018). The evaluation metrics can either be fractional, being derived directly from the fSCA, or binary, derived from binary snow cover (snow-covered or bare) obtained through thresholding the fSCA.

For the fractional metrics, the fSCA error of each sample pair is computed by subtracting the true fSCA from the estimated fSCA. Based on this error we compute the bias (mean error) as a measure of systematic differences and the root mean squared error (RMSE) as a measure of spread. In addition, we also compute the correlation coefficient to measure the strength of the linear correlation (R) between the truth and the estimate. In this study, the truth and the estimate are strongly and positively correlated. We therefore report $R^2$ instead to better distinguish between high correlations. Moreover, we perform a standard scatter plot analysis where best fit lines are computed using simple linear regression through ordinary least squares estimation.

For the binary metrics, we first convert fSCA to a binary snow-covered or bare pixel status, for which a threshold in the range [0,1] must be defined. Previous studies have used various thresholds, for example 0.15, 0.5 (Rittger et al., 2013) or 0 (Masson et al., 2018). Since fSCA can be interpreted as the probability of a randomly selected point in a pixel being snow-covered, it is difficult to justify a specific threshold. As the binary evaluation is nonetheless highly sensitive to this choice, we employ three thresholds representing various degrees of

![Fig. 5. The effect of Scheteligfjellet’s cast shadow on a Landsat 8 OLI level 2 scene sensed at 18:23Z on the 15.06.2014; (a) False color RGB = (SWIR1,NIR,R) image (note the shadow to the left in the image) with the yellow polygon delineating the AOI; (b) The fSCA determined from SU (white = 100%, black = 0%); (c) The independently modeled shade fraction (white = 0%, black = 100%) from the DEM based shadow mask. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)](image)

![Fig. 6. Sentinel-2 snow-cover mapping example: (a) Reference orthophoto taken around noon on the 21.06.2017; (b) Reference snow-cover map (white = snow, black = no snow); (c) S2A MSI L2A FC RGB = (SWIR1,NIR,R) image sensed 12:47:00Z on the same day; (d) Snow-cover map derived using a fixed NDSI threshold on TOA reflectances (FT-MSI-TOA) from the same S2A scene; (e) fSCA derived using spectral unmixing (SU-MSI) on the same S2A scene; (f) Snow-cover map derived from the L2A scene classification (SLCC-MSI) applied to the same S2A scene.](image)
fSCA in this study, specifically: 0.1, 0.5, and 0.9.

Given a specific threshold, which we apply to both the true and estimated fSCA, we may diagnose the number of true positives (TP, both are snow-covered), true negative (TN, both are bare), as well as false positives (FP, estimate is snow-covered while truth is bare) and false negatives (FN, estimate is bare while truth is snow-covered). Based on these counts we can compute various binary metrics. First we diagnose the positive predictive value (also called precision) which is the ratio of true positives to the total number of estimated positives (TP + FP). Then we compute the true positive rate (TPR, also known as recall and hit rate) which is the ratio of the number of true positives to the number of actual positives in the true data (TP + FN). We also diagnose the false positive rate (FPR) which is the ratio of the number of false positives to the actual number of negatives (TN + FP).

By varying the estimate binarization threshold from 0 to 1 and keeping the true binarization threshold fixed, TPR and FPR can be plotted against one another to construct receiver operating characteristics (ROC) curves as described in Fawcett (2006). Briefly, the ROC curve starts in the top right corner of ROC space (TPR vs FPR) for an estimator binarization threshold of −∞ where FPR = TPR = 1 and ends in the bottom left corner for a binarization threshold of ∞ where FPR = TPR = 0. In the ROC curves we emphasize the points where the two binarization thresholds are equal since for a perfect classifier these should fall in the top left corner of ROC space. We employ the area under the ROC curve (denoted $A$) to summarize the ROC performance in a single number. For an ideal classifier $A = 1$. Conversely, $A = 0.5$ indicates that the classifier is no better than a random guess. The extreme case $A = 0$ indicates that the classifier has its labels completely reversed (i.e. positive is negative and vice versa). In addition to the area under the curve, in line with Rittger et al. (2013) and Masson et al. (2018), we also calculate the F-score which is the harmonic mean of the positive predictive value and the true positive rate, where a score of $F = 1$ is perfect and $F = 0$ is the worst possible score. The F-score penalizes both errors of commission and omission. For both $A$ and $F$ we employ the three previously mentioned binarization thresholds. These two metrics are fundamentally different: $F$ is a mean measure of how well an algorithm detects positives (true negatives are not considered), whereas $A$ is the probability that the algorithm ranks a positive higher than a negative.

3. Results

3.1. Annual snow-cover depletion

We present results for the remotely sensed snow-cover depletion, aggregated to the scale of the AOI, for the six ablation seasons considered in this study (2012–2017). Fig. 7 shows time series of the fSCA depletion as sensed by the ACS and a subset of the satellite retrieval algorithms. The years 2012, 2013, 2015, and 2017 can all be characterized as normal snow years with the snowpack completely disappearing towards the end of June/early July (see black dots in Fig. 7), around a month long melt season, and a peak snow depth at the BCS around 1 m (Boike et al., 2018). In 2014 the snowpack was unusually high, persisting until mid July (Aalstad et al., 2018) with a peak depth at the BCS of 1.6 m. However, in 2014, the ACS malfunctioned, so that ground truth imagery was not available in most of the melt season (c.f. Fig. 7). In 2016, the snowpack had completely disappeared already in mid June, with a peak depth of 0.8 m at the BCS and a melt period of only half a month.

Using the ACS as a reference, Fig. 7 shows that the plotted MODIS products (MOD10A1 & MYD10A1 version 6) tend to slightly overestimate the fSCA during the snowmelt. While there are a few gaps in the snowmelt season due to clouds, these products are nonetheless still able to identify the end of the snowmelt season fairly well. The SU retrievals from Sentinel-2 MSI and Landsat 8 OLI do not suffer from the same positive bias as the MODIS products during the snowmelt period.

The unmixing retrievals are also remarkably accurate throughout the entire ablation season. Still, given the lower revisit frequency, there are slightly fewer retrievals than for MODIS, at least prior to 2017 and the launch of Sentinel-2B. Due to orbital convergence, the revisit frequency of the higher resolution sensors (OLI and MSI) is considerably higher at high latitudes than at the equator. In fact, Sentinel-2 MSI provides a near daily (as opposed to 10 day) revisit frequency in the high-Arctic. Using all the satellite sensors together (see 2017 in Fig. 7), it is possible to get quite a robust reconstruction of the true fSCA depletion curve at the scale of our AOI.

3.2. Fractional evaluation

Scatter plots of all the available sample pairs for the various ISCA satellite retrievals at 100 m and 500 m resolution are shown in Figs. 8 and 9, respectively. The fractional metrics that summarize the performance visualized in these scatter plots, namely bias, RMSE, and $R^2$, are listed in Tables 2, 3, and 4 for the 100 m, 500 m, and pixel-scale retrievals, respectively. We begin by describing the evaluation of the higher resolution retrievals.

Fig. 8 shows that all thresholding based algorithms at 100 m resolution have a tendency to overestimate the fSCA for both the OLI and MSI sensors. The overestimation is visible through the clustering of scatter points above the 1:1 line (blue) especially when the reference fSCA is higher than 0.4. This positive bias results in linear best fit lines with positive intercepts that lie above but run nearly parallel to the 1:1 line. For all thresholding retrievals, the biases are on the order of 0.03–0.05 for MSI and 0.05–0.07 for OLI. Of the thresholding experiments, the built in scene classifications, i.e. Fmask-OLI and SLCC-MSI, perform worst in terms of the fractional evaluation metrics, showing the highest combined bias and RMSE and the lowest correlation coefficient. The performance of SLCC-MSI is particularly poor with a 34% reduction in the number of available samples compared to the other MSI experiments. For Fmask-OLI the reduction in the number of samples, compared to the other OLI retrievals, is only 5%. For both scene land cover classifications, the reduction in the number of samples is due to cloud commission errors.

For both OLI and MSI, there is no noticeable performance gain when using an adaptive (AT) instead of a fixed (FT) threshold on the NDSI. In fact, for all the experiments at 100 m resolution, the adaptive threshold performs slightly worse than the corresponding fixed threshold, exhibiting a higher bias and RMSE. Based on the scatter plots, the most visible discrepancy between the adaptive and fixed thresholds are at low reference fSCA, with the adaptive threshold leading to a larger overestimation as the true fSCA approaches zero. Using surface instead of TOA reflectances for MSI (see e.g. FT-MSI-TOA vs FT-MSI-SFC in Table 2), i.e. applying an atmospheric correction, only results in a marginal improvement in performance, while there is no difference for OLI (Table 2, Fig. 8). Using the regression relationship in Eq. (2), quite precise retrievals were obtained with considerably lower RMSE than the thresholding based retrievals for OLI and MSI, but a positive (albeit lower) positive bias remained. For these regression-based retrievals there was a marginal improvement in RMSE and correlation when using surface rather than TOA reflectances. For both OLI and MSI, the SU technique shows the best performance. For the unmixing, the positive bias is almost eliminated (0.02 for OLI and 0.00 for MSI) and the linear best fit line closely tracks the 1:1 line. In addition, the RMSE is around half of the worst performing thresholding experiments. While the unmixing experiments still show a slight tendency towards overestimating fSCA at intermediate (0.25–0.75) fSCA, it is smaller than for the thresholding retrievals and not visible at low and high fSCA values.

At 500 m resolution, fSCA retrievals from both the moderate (MODIS) and higher resolution (OLI and MSI) sensors are available (Fig. 9). For all higher resolution sensors, the RMSE is reduced by around 0.02 (see Table 3) due to averaging of random errors in the spatial aggregation procedure. The reduction in random error also
results in a higher linear correlation for these retrievals at 500 m than at 100 m resolution. Note that the bias is not reduced accordingly as bias is a systematic error that does not cancel out upon spatial aggregation. As for the 100 m resolution retrievals, the overestimation of fSCA in the thresholding experiments is clearly visible, particularly at intermediate fSCA. This effect is still present, but less pronounced for the OLI and MSI SU retrievals. The low number of samples for the SLCC-MSI stands out in Fig. 9 and complicates interpretation.

For the ~ 500 m resolution MODIS retrievals, products based on the regression in Eq. (2) (version 5 and 6 of MOD10A1 and MYD10A1), as well as unmixing retrievals from the MODSCAG product and experiments with location specific endmembers are available (SU-MYD09GA and SU-MOD09GA). On the one hand, there is typically considerably more scatter than for the higher resolution sensors, which leads to a generally higher RMSE for the MODIS retrievals. On the other hand, the biases are typically around the same magnitude for the MODIS
retrievals (0–0.09) and the higher resolution sensors (0.01–0.08). We emphasize that the 500 m evaluation is at the scale of individual pixels for MODIS, but combines approximately 275 aggregated (30 m GSD) OLI pixels and 2500 (10 m GSD) MSI pixels for the higher resolution sensors.

The regression relationship (2) of Salomonson and Appel (2004) with the ‘universal FRA6T’ coefficients was applied to the NDSI obtained from the ‘NDSI snow-cover’ field inversion 6 of the MOD10A1 & MYD10A1 products. This brought the version 6 product in line with version 5, with an negligible difference in performance. For the bias and RMSE the largest change had a value of 0.01 going from version 5 to version 6, whereas the square correlation coefficient remained the same. Note that there are considerably more (23% for Aqua and 20% for Terra) retrievals available in the newer version 6 given that version 5 was discontinued after 2016.

Surprisingly, the empirically based MOD10A1 & MYD10A1 for the majority of the fractional evaluation metrics (see Table 3) outperform the unmixing based MODIS retrievals. Of the latter, MODSCAG performs worst with the algorithm struggling to recover the true fSCA in the low (< 0.25) and high (> 0.75) snow-cover cases, and with considerable scatter for intermediate fSCA. To further investigate this unfavorable performance, we performed two custom experiments: SU-MOD09GA and SU-MYD09GA. Unlike the MODSCAG products for the AOI, these custom unmixing experiments are based on version 6 (as opposed to version 5) of the MODIS surface reflectance products (MOD09GA, MYD09GA). Moreover, we considered regionalized (i.e., AOI specific) endmembers for these custom MODIS unmixing experiments, whereas the endmembers used in MODSCAG are unknown to us. The custom MODIS unmixing experiments performed slightly better than MODSCAG with a marked reduction in RMSE and increase in

---

**Table 2**

Summary of evaluation metrics, i.e., number of samples (N), Bias, RMSE, square correlation coefficient ($R^2$), area under the ROC curve ($A$), and F-score (F) for the satellite retrievals at 100 m resolution. The subscript for the binary metrics ($A$ and F) indicates the fSCA binarization threshold employed. For an overview of the various satellite retrievals see Table 1.

| Name          | N   | Bias | RMSE | $R^2$ | $A_{0.1}$ | $A_{0.5}$ | $A_{0.9}$ | $F_{0.1}$ | $F_{0.5}$ | $F_{0.9}$ |
|---------------|-----|------|------|------|-----------|-----------|-----------|-----------|-----------|-----------|
| FT-MSI-TOA   | 3960| 0.03 | 0.10 | 0.96 | 0.99      | 1.00      | 0.94      | 0.98      | 0.98      | 0.92      |
| AT-MSI-TOA   | 3960| 0.05 | 0.12 | 0.94 | 0.99      | 1.00      | 0.96      | 0.95      | 0.97      | 0.93      |
| FT-MSI-SFC   | 3960| 0.03 | 0.08 | 0.97 | 1.00      | 1.00      | 0.96      | 0.99      | 0.98      | 0.93      |
| AT-MSI-SFC   | 3960| 0.04 | 0.10 | 0.96 | 1.00      | 1.00      | 0.95      | 0.97      | 0.97      | 0.92      |
| SLCC-MSI     | 2628| 0.05 | 0.15 | 0.85 | 1.00      | 0.92      | 0.81      | 1.00      | 0.95      | 0.90      |
| SU-MSI       | 3960| –0.00| 0.07 | 0.98 | 1.00      | 1.00      | 0.98      | 0.98      | 0.99      | 0.90      |
| RG-MSI-TOA   | 3960| 0.02 | 0.08 | 0.97 | 0.99      | 1.00      | 0.98      | 0.97      | 0.98      | 0.95      |
| RG-MSI-SFC   | 3960| 0.02 | 0.07 | 0.98 | 1.00      | 1.00      | 0.98      | 0.98      | 0.99      | 0.95      |
| FT-OLI-TOA   | 4205| 0.05 | 0.13 | 0.92 | 0.99      | 0.99      | 0.87      | 0.99      | 0.97      | 0.90      |
| FT-OLI-SFC   | 4205| 0.06 | 0.14 | 0.90 | 0.99      | 0.99      | 0.88      | 0.98      | 0.97      | 0.90      |
| AT-OLI-TOA   | 4205| 0.05 | 0.13 | 0.92 | 0.99      | 0.99      | 0.87      | 0.98      | 0.97      | 0.91      |
| FT-OLI-SFC   | 4205| 0.06 | 0.14 | 0.90 | 0.99      | 0.99      | 0.86      | 0.98      | 0.97      | 0.90      |
| Fmask-OLI    | 3980| 0.07 | 0.17 | 0.83 | 0.99      | 0.95      | 0.80      | 0.98      | 0.96      | 0.88      |
| SG-OLI       | 4205| 0.02 | 0.09 | 0.96 | 1.00      | 1.00      | 0.96      | 0.98      | 0.98      | 0.91      |
| RG-OLI-TOA   | 4205| 0.04 | 0.11 | 0.93 | 0.99      | 1.00      | 0.91      | 0.98      | 0.98      | 0.90      |
| RG-OLI-SFC   | 4205| 0.04 | 0.10 | 0.94 | 0.99      | 1.00      | 0.92      | 0.99      | 0.98      | 0.92      |

![Fig. 9. Same as Fig. 8, but at 500 m spatial resolution. For an overview of the various satellite retrievals (indicated by the title of each panel) see Table 1. Results from several retrievals are not shown for visibility (see Table 3 metrics).](image-url)
correlation, which suggests that the use of regionalized endmembers in particular leads to better spectral unmixing results. Notably, the custom experiments performed better in high fSCA situations, but still had problems in the low fSCA case. Despite the slight improvements relative to MODSCAG, the performance of these custom unmixing experiments was still worse than the simpler MOD10A1 & MYD10A1 products. At the same time, it is worth emphasizing that the SU-based MODIS retrievals feature an equal or lower bias than the regression-based MOD10A1 & MYD10A1 products.

3.3. Binary evaluation

Receiver operating characteristics (ROC) curves for the binarized fSCA satellite retrievals at 100 m and 500 m resolution are shown in Figs. 10 and 11, using different binarization thresholds (fSCA = 0.1, 0.5 and 0.9, see Section 2.4). The binary metrics that summarize the performance of the binarized retrievals, namely the area under the ROC curve (A) and the F-score (F) are listed in Tables 2, 3, and 4.

For all retrievals at 100 m resolution (Fig. 10), the area under the curve is always lowest for a reference binarization threshold of fSCA = 0.9 (blue curve). This indicates that the algorithms perform worse for high snow-cover situations compared to intermediate (fSCA = 0.5) and low (fSCA = 0.1) snow-cover situations. For the lower binarization thresholds, all algorithms perform well (A ≥ 0.99), with the exception of the native scene classifications. Considering all binarization thresholds, the regression and unmixing retrievals from MSI and OLI perform best with respect to A, which is clearly visible in the ROC curves especially for the higher threshold. The native scene classifications (Fmask-OLI and SLCC-MSI) again perform worst. For the F-score, all algorithms perform favorably for the lower thresholds with F > 0.95, indicating a very good performance both in terms of the positive predictive value (precision: ratio of true positives to predicted positives) and the true positive rate (recall: ratio of true positives to the number of actual positives). This implies that both the number of errors of omission (false negatives) and omission (false negatives) are low. As for A, the F-score is worst (lowest) for the higher binarization threshold. It is worth highlighting that the true positive rate (recall) is near unity (i.e. a perfect score) in the ROC curves for most of the retrievals when the retrieval binarization threshold equals to the reference binarization threshold (cases marked by circles in Fig. 10). This suggests that low positive predictive values (precision) reduce the F-score from an ideal value of 1, indicating more errors of omission than omission.

For the retrievals at 500 m resolution, we see a similar picture. Once more, the performance for the lower reference binarization thresholds is noticeably (see Fig. 11) better than for the highest reference binarization threshold. For the lower thresholds, A exceeds 0.98 for all MSI and OLI retrievals, with the exception of the native scene classifications. Furthermore, the F-score shows a similar pattern as for the high resolution retrievals, with markedly worse scores for the highest binarization threshold. Once more, this is primarily due to a low positive
predictive value (see the circular markers in Fig. 11), indicating that the algorithms produce more false positives than false negatives. This finding is in line with the fractional evaluation metrics for which a positive bias occurred for most of the products. As previously noted, the bias does not disappear upon spatial aggregation, showing the same patterns in the binary evaluation at 100 m and 500 m. For the MODIS retrievals, we note that the high binarization threshold yields the lowest area under the ROC curves (Fig. 11). As for the fractional evaluation, MODSCAG performs considerably worse than both generations of the regression-based MODIS snow products with an area under the ROC curve well below unity for all the reference binarization thresholds. The performance is improved for the customized SU-MOD09GA and SU-MYD09GA experiments with both a higher $A$ and $F$-score closer to those of the regression-based retrievals.

4. Discussion

4.1. Error analysis

In this study, we used high (0.5 m) resolution reference snow-cover maps obtained from accurately georeferenced (bias and RMSE on the order of 2 m) orthorectified photographs. The high spatial resolution of...
these reference binary snow-cover maps means that they are less susceptible to the mixed pixel problem in that the majority of 0.5 m pixels are either fully snow-covered or fully snow-free. The estimated error in our reference fSCA retrievals due to the georeferencing error is on the order $\sigma_E = 0.01$ at 100 m (see Section 2.2), which is considerably smaller than the RMSE of the best performing satellite retrievals (e.g. SU-MSI with RMSE = 0.07, see Table 2). Even in the infrequent cases of a $3\sigma$ georeferencing error, inaccurate georeferencing translates to an error on the reference fSCA of $\sigma_E = 0.03$, which is less than half the RMSE (based on the reference fSCA) of the best performing retrieval. Due to the effects of spatial averaging, the estimation standard deviation for the reference fSCA at 500 m will be even lower. Thus, our estimates of the satellite retrieval error can closely represent the true error of the retrievals for our site. Although we had a relatively small (1.77 km$^2$) AOI, 302 high resolution reference images were available for evaluation, so that we consider our reference dataset to be sufficiently extensive for the evaluation of higher resolution imagery from OLI and MSI. For the moderate (500 m) resolution MODIS retrievals, our study area only contains 5 pixels which is rather low. Nonetheless, the fact that we had hundreds of reference images translated into a total of on the order 400 fSCA retrievals at 500 m resolution that we could evaluate for each of the MODIS products. Combined with the fact that our study area is relatively flat and representative of large swatches of Arctic tundra, the validation of MODIS fSCA retrievals at our study site becomes a valuable exercise. It should be noted, however, that we do not expect our evaluation to necessarily reflect the general performance of the global MODIS snow-cover products given the limited size of the evaluation area and the lack of global representativeness.

Earlier work (e.g. Cortés et al., 2014; Masson et al., 2018) has emphasized the need for an evenly sampled fSCA so that situations with no-snow or full snow-cover are not over-represented. These edge cases should be the easiest for any algorithm to capture, with intermediate partial snow-cover being more challenging. In our case, we only employ reference fSCA maps from the entire ablation season (see Fig. 7), which ensures that the entire range from 0 to 1 is represented. However, the reference data set does not feature an even distribution of fSCA values, as many almost fully snow-covered and bare scenes were included. Our study shows that it is important to test the algorithms in these simple cases as well. As can be seen in Figs. 10 and 11, many of the algorithms did not perform well in situations with high snow-cover ($fSCA > 0.75$). When computing fractional metrics, such as RMSE, we did not see the need (c.f. Masson et al., 2018) to remove situations with zero snow-cover since the MODIS retrievals in particular had some large excursions from the truth in situations with zero fSCA. Such excursions are clear signs of cloud contamination and errors of omission in the MODIS cloud mask. Cloud masking is a challenging task that we did not
investigate explicitly (see Zhu and Woodcock, 2012); instead we manually selected the highest resolution imagery to ensure that it was cloud free. Cloud commission or omission errors are a large source of uncertainty in any automatic snow-cover mapping algorithm as discussed in Gascoin et al. (2015) and Stillinger et al. (2019).

For the binary evaluation, we considered more than just a single fSCA binarization threshold. With the ROC analysis, all the algorithms performed least favorably for high snow-cover, i.e. for a reference binarization threshold of 0.9. This would not have been clear if only a single low threshold (e.g. 0) would have been employed (e.g. Masson et al., 2018). The ROC analysis showed that low F-scores were due to low precision (errors of commission) rather than recall (errors of omission). This is connected to the positive bias in the thresholding algorithms due to the mixed-pixel problem. One benefit of a ROC analysis over other binary evaluations is that it is quite insensitive to a skewed data distribution, i.e. to the distribution of positives and negatives in the reference data. Moreover, ROC evaluation is natural when one considers fSCA defined as the probability of occurrence of snow-cover within a pixel since the ROC analysis is specifically designed to deal with probabilistic classification (Fawcett, 2006).

The evaluation showed that the errors for most of the retrievals were dependent on the fSCA itself (i.e. heteroscedastic). For most of the retrievals, this heteroscedasticity manifests itself (see Fig. 12) through errors that are higher for intermediate fSCA and lower for high and low fSCA. Our results suggest that the error variance in satellite retrieved fSCA can be parametrized through the following expression (after Sakov et al., 2012)

\[ \sigma^2_{\text{err}} = \sigma^2 + \left( \frac{\sigma^2 - \sigma^2_0}{\alpha} \right)(x - |x - \text{fSCA}|) \]  

(6)

where \( \sigma^2_0 \) is the minimum error variance, \( \sigma^2 \) is the maximum error variance, and we propose \( \alpha = 0.5 \) for symmetry. Thereby, the error variance is at its minimum for full or zero fSCA and at its maximum for intermediate fSCA. For spectral unmixing at 100 m resolution, appropriate values for the error standard deviations are \( \sigma_0 = 0.05 \) and \( \sigma_1 = 0.1 \). Expression (6) can be a useful addition for fSCA data assimilation studies which so far have typically employed a fixed observation error variance (e.g. Girotto et al., 2014; Cortés et al., 2016; Aalst et al., 2018).

4.2. Thresholding algorithms

The native scene land cover classification algorithms Fmask-OLI and SLCC-MSI are designed to conservatively map changes in land use and land cover by masking out undesired artifacts, such as clouds or snow (see Zhu and Woodcock, 2012). Perhaps not surprisingly, they perform worst of all the thresholding algorithms with respect to snow cover mapping. Both the Fmask (Zhu and Woodcock, 2012) and SLCC-MSI (Richter et al., 2012) algorithms perform an expansion of the snow mask through a dilation operation, which likely explains the high bias. While all the selected scenes were cloud free, SLCC flagged a considerable amount of snow-covered pixels as cloudy, which led to an omission of 34% of the valid data. The cloud commission errors were not as severe in Fmask, with only 5% data loss. The fact that OLI (in contrast to MSI) has a thermal band may explain some of the differences in cloud masking, although temperature based cloud masking can be challenging over snow (e.g. Westermann et al., 2012; Østby et al., 2014). For larger scale snow-mapping, there is potential to improve the cloud cover masking in the SLCC-MSI algorithm, for example following Gascoin et al. (2019) and Stillinger et al. (2019).

For the remaining thresholding retrievals, no appreciable difference was found between using top of the atmosphere or surface reflectances. This most likely can be explained with the relative insensitivity to atmospheric scattering and absorption of the green and SWIR band, which are employed for the NDSI (Dozier, 1989; Riggs et al., 2017). Use of the NDSI, as with other band ratio methods, lessens the influence of atmospheric effects (Salomonson and Appel, 2004). Accordingly, Yin et al. (2013) found little performance gain when employing atmospheric correction for NDSI snow-cover mapping, and the MODIS NDSI-based snow-cover product does not employ any atmospheric correction (Riggs et al., 2017). As accurate atmospheric correction (see e.g. Vermote et al., 2016) can be difficult and often introduces additional uncertainty in the retrieval process, it is encouraging that NDSI-based fSCA detection performs well when employing TOA reflectances directly. Note that for MODIS Masson et al. (2018) demonstrated an improvement in NDSI snow-cover mapping when using atmospheric and topographically corrected reflectances in mountainous terrain. Our results may be an indication that the improvement in their results is largely due to the topographic correction.

Previous studies (e.g. Yin et al., 2013; Härer et al., 2018) suggest a performance gain when employing an adaptive (scene-dependent) NDSI threshold instead of a fixed threshold of 0.4. We tested the best performing algorithm in Yin et al. (2013), namely the binarization of Otsu (1979), only constraining the threshold to lie in the range [0.1,0.7] (which is reasonable for snow-covered pixels, Riggs et al., 2017). Our evaluation did not show an improvement in performance when using this algorithm (see Fig. 8), and the performance was even slightly worse for low fSCA for which there was a noticeable increase in the positive bias. The algorithm tended to pick a threshold near the lower limit of 0.1 for low fSCA scenes, which resulted in an overestimation of the snow-covered area, including several errors of commission over the Bayelva-river floodplain. While it is worth optimizing the location and scene-specific choice of the NDSI threshold, our results indicate that using an automatic thresholding algorithm may not be ideal in all situations. This is in line with findings of Härer et al. (2018) who only found small differences when implementing the segmentation method of Otsu (1979). This segmentation method does not work well when the foreground (snow) and background (bare) clusters are poorly defined, especially in the case that one cluster covers a much smaller portion of the scene than the other. This happens when the fSCA for the entire AOI is either very low or very high. In agreement with Härer et al. (2018), we do not recommend applying automatic NDSI threshold selection techniques without ground truth data to support their validity.

All thresholding algorithms exhibited a considerable positive fSCA bias (see Tables 2, 3), which compromises the use of aggregated fSCA maps produced by thresholding of higher resolution NDSI maps as a reference for moderate resolution retrievals from sensors, such as MODIS (e.g. Salomonson and Appel, 2004; Gascoin et al., 2015; Masson et al., 2018). In Rittger et al. (2013), the potential limitations of using a thresholding based retrieval as reference were pointed out. At least for our AOI, the errors in the aggregated threshold-based retrievals are on the same order of magnitude as the errors of some of the coarser-scale MODIS retrievals. Therefore, if the threshold-based retrievals were used as a reference, many of the MODIS retrievals would appear to perform much worse than in reality. In fact, the retrieval algorithms used by all MODIS products partly overcome the mixed pixel problem that affects all thresholding-based retrievals. For the validation exercise, we strongly recommend exclusively making use of ISCA retrievals that have accounted for the mixed pixel problem, such as SU or regression, for the reference data set, as in the study of Rittger et al. (2013). An alternative is to employ thresholding retrievals from very high resolution satellite imagery (Cortés et al., 2014), as this strongly moderates the mixed pixel problem.

4.3. Mixed pixel problem

In natural landscapes, spatial heterogeneity often leads to mixed pixel problems: a single satellite pixel is typically composed of more than one endmember (Painter et al., 2003). This is often the case in the middle of winter, not just the ablation season, due to rocks and vegetation that protrude through the snow. As such, classifying a pixel into a
single land-cover is inherently problematic, although it is implicitly understood that the land-cover assigned to a pixel represents the most abundant endmember in that pixel. The mixed pixel problem becomes less severe with higher spatial resolution, but does not fully disappear until the scale of a single point.

As a result, any retrieval algorithm that uses thresholds to binarize a pixel to a single endmember is inherently flawed, but one assumes that the error introduced by thresholding is not systematic and cancels out in the process of aggregation. Our results suggest that this is not necessarily the case for snow-cover, with thresholding algorithms upon aggregation systematically overestimating fSCA. This is even the case at the higher resolution of 100 m, for which all thresholding based retrievals exhibit a positive bias and lower scores for the binary evaluation metrics, especially for high binarization thresholds. The positive bias in these retrievals is explained by the aggregation of pixels that are only partly snow-covered, but are classified as snow-covered by the binary thresholding. If as many low-snow-cover pixels were mapped as snow-free as high snow-cover pixels were mapped as snow-covered, thresholding would give an on average unbiased response, but this is clearly not the case in our study. The overestimation can be identified in the ROC space (see Fig. 10) where the false positive rate is generally high (⩾0.1) for the larger binarization thresholds (blue and gray circles). This effect is also seen in Fig. 6 in which most of the Bayelva floodplain is mapped as fully snow-covered by the thresholding algorithms, while it is only partially snow-covered in reality.

Spectral unmixing algorithms, pioneered for fSCA retrieval by Nolin et al. (1993) and operationalized by Painter et al. (2009), explicitly take the mixed pixel problem into account (see Eq. 4)). In Fig. 6, we note that much of the Bayelva river floodplain is correctly retrieved as partially snow-covered in the unmixing experiment with Sentinel-2 MSI. For the higher resolution sensors, the unmixing algorithms clearly outperform all thresholding algorithms. For SU, applying an atmospheric correction is vital, given that the employed endmember spectra are surface reflectances, not TOA reflectances (see discussion in Schaepman-Strub et al., 2006). As discussed, this atmospheric correction procedure may introduce additional uncertainty, but the unmixing-based experiments in our study nonetheless performed best for the higher resolution sensors (OLI and MSI), with almost zero bias and relatively low RMSE values. The favorable performance can in large parts be attributed to the physically-based nature of the SU retrieval algorithm and the fact that the mixed-pixel problem is explicitly accounted for (Keshava and Mustard, 2002).

In contrast to the favorable performance for the higher resolution sensors and unlike previous studies (e.g. Rittger et al., 2013), the unmixing-based MODSCAG algorithm (Painter et al., 2009) did not perform well in our study (see Fig. 9). We initially suspected that the (unknown) choice of endmembers in MODSCAG could explain the poor performance, and by picking AOI specific endmembers we could slightly improve the performance in the custom experiments SU-MOD09GA and SU-MYD09GA (see Table 3). These custom experiments also benefited from the updated (version 6) MODIS surface reflectance products, as opposed to the version 5 products that were still used in the MODSCAG retrievals available for our AOI. Still, even the custom experiments were outperformed by the simpler MOD10A1 & MYD10A1 products. We highlight that, to the best of our knowledge, MODSCAG has not previously been evaluated in the Arctic. It is therefore possible that some of the assumptions in MODSCAG, such as ignoring directional effects (Painter et al., 2009), are more problematic in the Arctic where solar zenith angles are much higher during the ablation season than at lower latitudes (minimum \( \theta_s = 56° \) in the study area). This can be especially problematic for MODIS which, unlike OLI or MSI, is not a nadir looking sensor, since high solar zenith angles make forward scattering from snow in the view direction more likely (Painter et al., 2009). The combination of off-nadir view angle and high solar zenith angle may also make situations with sunglint possible (which is not considered in the unmixing). Nonetheless, we found negligible correlation between the retrieval errors and sensor zenith angle as well as solar zenith angle for MODSCAG and the custom SU experiments, respectively. This indicates that directional effects are not likely to be the source of error in line with the findings of Painter et al. (2009).

Dozier et al. (2008) show that the effective pixel dimension can increase by a factor of 10 when the view angle of MODIS is greatly off nadir, which could also be a significant source of error in unmixing with MODIS (Painter et al., 2009). In our study, this is probably not the case as this error source should then also manifest itself in the regression-based MODIS products MOD10A1 & MYD10A1 which it does not seem to do. Based on the formulae provided by Dozier et al. (2008), the pixel stretching factor (change in area) is 1.5 for a sensor view zenith angle of 30°. For the MODSCAG retrievals analyzed herein we found that the view zenith angle is lower than 30° in the overwhelming majority of cases. Due to this geometric stretching and the triangular response function of the whiskbroom MODIS sensor (see Nishihama et al., 1997) admixing of signals emanating from neighboring pixels is a potential issue. To check if this was the case, we diagnosed the correlation in the fSCA time series between the different MODIS pixels. For all of the MODIS retrievals, we found high (> 0.9) between pixel fSCA correlations, indicating that our AOI is homogeneous in terms of fSCA at the MODIS pixel-scale. Moreover, strongly different surfaces, in particular the ocean, are in general one pixel or more away from the MODIS pixels considered in our AOI (see Fig. 1). Thus admixing of neighboring pixels does not appear to be a major issue for our AOI and does not change the conclusions provided.

Both versions of the MOD10A1 & MYD10A1 products performed reasonably well for both the binary and fractional evaluation metrics (see Table 3), performing similarly to the thresholding experiments. At first glance, this is surprising given that the thresholding experiments at 500 m are based on the aggregation of a large number of smaller pixels, which considerably reduces random error. At the same time, the thresholding retrievals are biased due to the mixed pixel problem, and this bias does not disappear upon aggregation. The MOD10A1 & MYD10A1 products are based on the empirical regressions in Eq. (2) with coefficients presented in Salomonson and Appel (2004). As noted by Riggs et al. (2017), the NDSI in (fractionally) snow-covered landscapes typically ranges from 0 to 1, which the linear regression can account for. As such, the NDSI regression implicitly considers the mixed pixel problem. This implicit accounting also explains why the regression-based retrievals for MSI and OLI markedly outperformed the thresholding-based retrievals, although a slight bias remained. Being based on the NDSI, which is relatively robust to atmospheric effects (Salomonson and Appel, 2004), these regression-based retrievals performed similarly across all evaluation metrics regardless of whether or not an atmospheric correction was applied. Both versions 5 and 6 of MOD10A1 & MYD10A1 performed reasonably well, with for the most part superior evaluation metric scores compared to the SU-based MODIS fSCA retrievals. The biases in these regression-based retrievals were, however, comparable or even larger than those in the SU-based retrievals. Thus, at a certain level of aggregation (upsampling), due to the cancellation of random errors, the SU-based MODIS retrievals will match or even outperform the regression-based MODIS retrievals. It is possible that a better performance can be obtained by tuning the coefficients of the regressions for this particular site (Riggs et al., 2017). However, as noted by Masson et al. (2018), employing the ‘universal’ coefficients is advantageous for validation purposes, since the results can be generalized beyond the study area.

Based on our results, we make the general recommendation of employing ISCA retrieval algorithms that take into account the mixed pixel problem. It is only in this way that systematic error can be eliminated from the retrieval process. Such algorithms could go beyond the linear mixed pixel accounting approaches considered here (i.e., SU and NDSI regression), and include nonlinear regression approaches from the machine learning community such as artificial neural networks (e.g. Czyzowska-Wisniewski et al., 2015). To extend the
applicability of our results for the higher level MODIS products, which were based on a low number of pixels, we envisage increasing the size of the validation area and making use of higher resolution satellite retrievals (from e.g. Sentinel-2 MSI or Landsat 8 OLI) as a reference. These reference satellite retrievals would have to account for the mixed pixel problem to avoid a biased validation procedure. Such an extension is, however, beyond the scope of the present study which focuses on the use of a dense time series of very high resolution retrievals from a terrestrial automatic camera system that are only available for a small limited area.

5. Conclusion

In this study, we evaluated satellite retrievals of fractional snow-covered area (fSCA) using several hundred high resolution snow-cover maps retrieved from orthorectified terrestrial images. These images were taken over the course of six snowmelt seasons (2012–2017) by an automatic camera system installed on a mountain peak near Ny-Ålesund in the high-Arctic archipelago of Svalbard. We considered ISCAs retrievals from three different optical satellite sensors, namely: the operational land imager (OLI) on-board Landsat-8, the multispectral instrument (MSI) on-board the Sentinel-2 satellites, and the moderate resolution imaging spectroradiometer (MODIS) on-board Terra and Aqua. The algorithms employed to retrieve ISCAs from multispectral satellite imagery ranged from simple thresholding of the normalized difference snow index (NDSI) to fully constrained spectral unmixing (SU). The satellite retrievals were extensively evaluated with respect to the difference snow index (NDSI) to fully constrained spectral unmixing (SU). The algorithms employed to retrieve ISCAs from multispectral satellite imagery ranged from simple thresholding of the normalized difference snow index (NDSI) to fully constrained spectral unmixing (SU). The satellite retrievals were extensively evaluated with respect to the difference snow index (NDSI) to fully constrained spectral unmixing (SU).

The Satellite retrievals were extensively evaluated with respect to the difference snow index (NDSI) to fully constrained spectral unmixing (SU). The algorithms employed to retrieve ISCAs from multispectral satellite imagery ranged from simple thresholding of the normalized difference snow index (NDSI) to fully constrained spectral unmixing (SU). The satellite retrievals were extensively evaluated with respect to the difference snow index (NDSI) to fully constrained spectral unmixing (SU).

The NDSI-regression based ISCAs retrievals implicitly account for the mixed pixel problem and can thus provide satisfactory results.

The Fmask-OLI and SLCC-MSI native scene classifications overestimate both cloud and snow-cover.

The unsupervised adaptive NDSI threshold selection does not necessarily outperform a fixed threshold of 0.4.

The atmospheric correction has little impact on ISCAs retrieved using NDSI thresholding and regression.

The conservative uncertainty estimates established in this study are useful for reanalyses that ingest satellite-retrieved ISCAs in models using data assimilation. Given that the SU retrievals for the higher resolution OLI and MSI sensors performed considerably better than all the other retrievals, we strongly recommend efforts towards operationalizing merged SU products from these sensors. Spectral unmixing is only slightly more computationally expensive than the other retrievals, it appears that the benefit of moderating the mixed pixel problem by providing improved nearly unbiased results through unmixing outweighs these costs. Although the 500 m operational MOD10A1 and MYD10A1 products performed reasonably well, a higher resolution merged operational unmixing product would yield marked reductions in error without significantly compromising the revisit time, especially in the Arctic.

CRediT authorship contribution statement

Kristoffer Aalstad: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. Sebastian Westermann: Conceptualization, Methodology, Software, Investigation, Resources, Data curation, Writing - review & editing, Supervision, Project administration, Funding acquisition. Laurent Bertino: Methodology, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Our work was funded by SatPerm (239918; Research Council of Norway), in collaboration with EmblA (56801; Nordforsk) and the European Space Agency Permafrost CCI project (http://cci.esa.int/Permafrost). This study is a contribution to the strategic research area LATICE at the University of Oslo. We gratefully acknowledge all the data providers for making their data open and freely available. Landsat 8 OLI scenes were downloaded from USGS Earth Explorer (http://earthexplorer.usgs.gov). Sentinel-2A/2B MSI scenes were downloaded from the Copernicus Open Access Hub (https://sci.hub.copernicus.eu). MODIS products were downloaded from NSIDC (MOD10A1/MYD10A1), LP DAAC (MOD09GA/MYD09GA), and JPL (MODSCAG). The spectral response functions of the different sensors were obtained from EUMETSAT and ESA. The HRSC-AX imagery and DEM was obtained from the data set associated with the study of Boike et al. (2018). The DEM for the entire Brøgger peninsula was obtained from the Norwegian Polar Institute (NPI, 2014). The vast majority of the work conducted in this study was carried out in the MATLAB computing environment using native functions and toolboxes. In addition, solar geometry was diagnosed using the ephemeres routine in the PV_LIB toolbox developed at Sandia National Labs. Based on the ephemeres, shadows were obtained using the TopoToolbox (Schwanghart and Scherler, 2014) thanks to a suggestion from Bas Altena. We would like to thank Ge et al. (2018) for making the SNICAR MATLAB routines available. Thanks also to George A. Biggs and Kat Bormann for their helpful comments about the MODIS products. We gratefully acknowledge field support from Kjersti Gisnås, Jaroslav Obu, and Håvard Kristiansen. We also want to thank the editor, Menghua Wang, and three anonymous reviewers, all of whom contributed to significantly improving this paper.

References

Aalstad, K., Westermann, S., Schuler, T.V., Boike, J., Bertino, L., 2018. Ensemble-based assimilation of fractional snow-covered area satellite retrievals to estimate the snow distribution at Arctic sites. Cryosphere 12, 247–270. https://doi.org/10.5194/cp-12-247-2018.

Arsenault, K.R., Houwer, P.R., De Lannoy, G.J.M., 2014. Evaluation of the MODIS snow cover product fraction product. Hydrolog. Process. https://doi.org/10.1002/hyp.9636.

Baldridge, A.M., Hook, S.J., Grove, C.I., Rivera, G., 2009. The ASTER spectral library version 2.0. Remote Sens. Environ. 113, 711–715. https://doi.org/10.1016/j.rse.2008.11.007.

Barnett, T.P., Adam, J.C., Lettenmaier, D.P., 2005. Potential impacts of a warming climate on water availability in snow-dominated regions. Nature 438. https://doi.org/10.1038/nature04141.

Boike, J., Jussak, I., Lange, S., Chadburn, S., Burke, E., Overduin, P.P., Roth, K., Ippisch, O., Bornemann, N., Stern, L., Gouttevin, I., Hauber, E., Westermann, S., 2018. A 20-year record (1998–2017) of permafrost, active layer and meteorological conditions at high Arctic permafrost research site (Bayoqta, Spitsbergen). Earth Syst. Sci. Data 10, 355–396. https://doi.org/10.5194/essd-10-355-2018.

Borumann, K., Brown, R.D., Derksen, C., Painter, T.H., 2018. Estimating snow-cover trends from space. Nat. Clim. Chang. https://doi.org/10.1038/s41558-018-0318-3.

Bouguet, J.Y., 2015. Camera Calibration Toolbox for Matlab. Available at: http://www. vision.caltech.edu/bouguetj/calib_doc/ (last access: December 2018).

Chadburn, S.E., Burke, E.J., Cox, P.M., Friedlingstein, P., Hugelius, G., Westermann, S., 2017. An observation-based constraint on permafrost loss as a function of global warming. Nat. Clim. Chang. 7, 340–344. https://doi.org/10.1038/nclimate3262.

Chapin, F.S., Sturm, M., Serreze, M.C., McFadden, J.P., Key, J.R., Lloyd, A.H., McGuire, A.D., Rupp, T.S., Lynch, A.H., Schimel, J.P., Beringer, J., Chapman, W.L., Epstein, H.E., Euskirchen, E.S., Hinman, L.D., Jia, G., Ping, C.-L., Tape, K.D., Thompson, C.D.C., Walker, D.A., Welker, J.M., 2005. Role of land-surface changes in Arctic
summer warming. Science 310, 657–660. https://doi.org/10.1126/science.1127368

Cortés, G., Girotto, M., Margulis, S.A., 2014. Analysis of sub-pixel snow and ice extent over the extratropical Andes using spectral unmixing and historical Landsat imagery. Remote Sens. Environ. 141, 64–78. https://doi.org/10.1016/j.rse.2013.05.023.

Cortés, G., Girotto, M., Painter, T.H., 2016. Snow cover estimation over the extratropical Andes using a data assimilation framework integrating MERRA data and Landsat imagery. Water Resour. Res. 52. https://doi.org/10.1002/2015WR1018376.

Czyzowska-Wisniewski, E.H., van Leeuwen, J.D., Hirschboeck, K.H., Marsh, S.E., Wisniewski, V.T., 2014. Fractional snow cover estimation in complex alpine-forested environments using an artificial neural network. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2019.09.026.

Dozier, J., 1989. Spectral signature of Alpine snow cover from the Landsat Thematic Mapper. Remote Sens. Environ. 28, 9–22. https://doi.org/10.1016/0034-4257(89)90101-6.

Dozier, J., Bair, E.H., Davis, R.E., 2016. Estimating the spatial distribution of snow water equivalent in the world's mountains. WIREs Water 3, 461–474. https://doi.org/10.1002/wat2.12140.

Dozier, J., Green, R.O., Nolin, A.W., Painter, T.H., 2009. Interpretation of snow properties from imaging spectrometry. Remote Sens. Environ. 113, S25–S37. https://doi.org/10.1016/j.rse.2009.04.005.

Dozier, J., Painter, T.H., Rittger, K., Frew, J.E., 2008. Time-space continuity of daily maps of fractional snow cover and albedo from MODIS. Adv. Water Resour. https://doi.org/10.1016/j.advwatres.2008.08.011.

Drusch, M., Del Bello, U., Carlier, S., Collin, O., Fernandez, Y., Gascon, F., Hoersch, B., Isola, C., Labernini, P., Martimort, P., Meyret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P., 2012. Sentinel-2: ESA’s optical high-resolution mission for GMES operational services. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2011.11.014.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 26, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Fedorov, R., Caramoti, A., Fraternali, P., Tagliasacchi, M., 2016. Estimating snow cover fraction using availability-based software. IEEE Trans. Multimedia 18, 1187–1200. https://doi.org/10.1109/TMM.2016.255356.

Foreland, J.E., Benestad, R., Hansen-Bauer, I., Haugen, J.E., Skaugen, T.E., 2011. Temperature and precipitation development at Svalbard 1900–2100. Adv. Meteorol. 2011, 1–14. https://doi.org/10.1155/2011/697960.

Gascon, S., Grisonnet, M., Bouchet, M., Salgues, G., Hagolle, O., 2019. Theias snow cover monitoring. Remote Sens. Environ. 225, 221–276. https://doi.org/10.1016/j.rse.2019.09.013.

Gasco, G., Hagoort, C., Haen, M., Jarlan, L., Szenczi, C., Märti, R., Sánchez, R., 2015. A snow cover climatology for the pyrenees from mods snow products. Hydrol. Earth Syst. Sci. 19, 2317–2331. https://doi.org/10.5194/hess-19-2317-2015.

Girotto, M., Margulis, S.A., Durand, M., 2014. Probabilistic SWE reanalysis as a generalization of deterministic SWE reconstruction techniques. Hydrol. Process. 28. https://doi.org/10.1002/hyp.9887.

Hall, D.K., Salomonson, V.V., Appel, I., 2004. Development of the Aqua MODIS NDSI fractional snow cover algorithm. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2003.10.016.

Hall, D.K., Painter, T.H., Rittger, K., Nolin, A.W., 2013. Sentinel-2 L2A MSI - Level 2A Products Algorithm Theoretical Basis Document. Issue 2.0. https://sentinel.esa.int/web/sentinel/missions/sentinel-2-lvl-2a.

Hall, D.K., Salomonson, V.V., Riggs, G.A., 2006a. MODIS/Aqua Snow Cover Daily L3 Global 500m Grid, Version 6, h18v01, Boulder, CO, USA, NASA NSIDC DAAC. https://doi.org/10.5067/63NQASRDPDB0. Accessed: 06-2018.

Hall, D.K., Salomonson, V.V., Rittger, K., 2006b. MODIS/Terra Snow Cover Daily L3 Global 500m Grid, Version 6, h18v01, Boulder, CO, USA, NASA NSIDC DAAC. https://doi.org/10.5067/moda/MOD09A16. Accessed: 06-2018.

Hall, D.K., Dirgelamo, N.E., Bayr, K.J., 2002. MODIS snow-cover products. Remote Sens. Environ. 83, 181–194. https://doi.org/10.1016/S0034-4257(02)00009-0.

Hall, D.K., Salomonson, V.V., Riggs, G.A., 2006a. MODIS/Aqua Snow Cover Daily L3 Global 500m Grid, Version 6, h18v01, Boulder, CO, USA, NASA NSIDC DAAC. https://doi.org/10.5067/63NQASRDPDB0. Accessed: 06-2018.

Hall, D.K., Salomonson, V.V., Riggs, G.A., 2006b. MODIS/Terra Snow Cover Daily L3 Global 500m Grid, Version 6, h18v01, Boulder, CO, USA, NASA NSIDC DAAC. https://doi.org/10.5067/63NQASRDPDB0. Accessed: 06-2018.

Häuser, S., Bernhardt, M., Siebers, M., Schulz, K., 2018. On the need for a time- and location-dependent estimation of the NDSI threshold value for reducing existing uncertainties in snow cover maps at different scales. Cryosphere 12, 1629–1642. https://doi.org/10.5194/cr-12-1629-2018.

He, C., Shi, M.G., Chen, L., Pohl, J., Liu, X.K., Kang, S., Ming, J., Qian, Y., 2018. Black carbon-induced snow albedo reduction over the Tibetan Plateau: uncertainties from snow grain shape and aerosol-snow mixing state based on an updated SNICAR model. Atmos. Chem. Phys. 18. https://doi.org/10.5194/acp-18-11507-2018.

Heinz, D.C., Chang, G., 2014. Fully constrained least squares linear spectral mixture analysis for material quantification in hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. https://doi.org/10.1109/TGERS.2013.2298318.

K. Aalstad, et al. Remote Sens. ENVIRON. 239 (2020) 116168

Lawson, C.L., Hanson, R.J., 1995. Linear least squares with linear inequality constraints. In: Solving Least Squares Problems. SIAM, pp. 160–165. https://doi.org/10.1137/1.9780898719562.ch4.

Mankin, J.S., Vivoli, D., Singh, D., Hoekstra, A.Y., Diffenbaugh, N.S., 2015. The potential for snow to supply human water demand in the present and future. Environ. Res. Lett. 10. https://doi.org/10.1088/1748-9326/10/11/114016.

Martín-R., Gascon, S., Berthier, E., de Pinedo, M., Hower, T., Laffly, D., 2016. Mapping snow depth in open alpine terrain from stereo satellite imagery. Cryosphere 10, 1361–1380. https://doi.org/10.5194/tc-10-1361-2016.

Martínez, J., Rango, A., 1981. Areal distribution of snow water equivalent evaluated by snow cover monitoring. Water Resour. Res. 17, 1480–1488. https://doi.org/10.1029/RW017i005p01480.
Stillinger, T., Roberts, D.A., Collar, N.M., Dozier, J., 2019. Cloud masking for Landsat 8 and MODIS Terra over snow-covered terrain: error analysis and spectral similarity between snow and cloud. Water Resour. Res. https://doi.org/10.1029/2019WR024932.

Swenson, S.C., Lawrence, D.M., 2012. A new fractional snow-covered area parameterization for the Community Land Model and its effect on the surface energy balance. J. Geophys. Res.: Atmos. https://doi.org/10.1029/2012JD018178.

Treichler, D., Kääb, A., 2017. Snow depth from ICESat laser altimetry - a test study in southern Norway. Remote Sens. Environ. 191, 389–401. https://doi.org/10.1016/j.rse.2017.01.022.

Trofaier, A.M., Westermann, S., Bartsch, A., 2017. Progress in space-borne studies of permafrost for climate science: towards a multi-ECV approach. Remote Sens. Environ. 203, 55-70. https://doi.org/10.1016/j.rse.2017.05.021.

USGS, 2018. Product Guide Landsat 8 Surface Reflectance Code (LASRC) Product. Available at: https://www.usgs.gov/land-resources/nli/landsat/files (last access: December 2018).

Vermote, E., Justice, C., Claverie, M., Franch, B., 2016. Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2016.04.008.

Westermann, S., Boike, J., Guglielmin, M., Gisnås, K., Etzelmüller, B., 2015. Snow Melt Monitoring Near Ny-Ålesund, Svalbard, Using Automatic Camera Systems. Department of Geosciences, University of Oslo, PANGAEA. https://doi.org/10.1594/PANGAEA.846617. Last access: 5 May 2017.

Westermann, S., Langer, M., Boike, J., 2011. Spatial and temporal variations of summer surface temperatures of high arctic tundra on Svalbard - implications for MODIS LST based permafrost monitoring. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2010.11.018.

Yılmaz, Y., Aalstad, K., Sen, O.L., 2019. Multiple remotely sensed lines of evidence for a depleting seasonal snowpack in the near East. Remote Sens. https://doi.org/10.3390/rs11050483.

Zhang, T., 2005. Influence of the seasonal snow cover on the ground thermal regime: an overview. Rev. Geophys. 43. https://doi.org/10.1029/2004RG000157.

Zhu, Z., Woodcock, C.E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. 118, 83–94. https://doi.org/10.1016/j.rse.2011.10.028.