Science of the Total Environment 651 (2019) 2712–2726

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Ground-based evaluation of MODIS snow cover product V6 across China: Implications for the selection of NDSI threshold

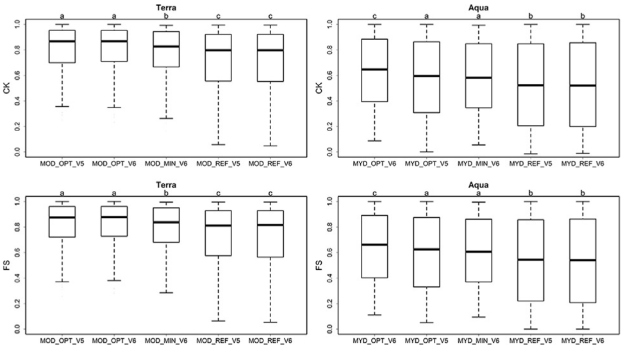
Hongbo Zhang [a], Fan Zhang [a,b,c,*], Guoqing Zhang [a,b], Tao Che [d], Wei Yan [e], Ming Ye [f], Ning Ma [a]

[a] *Key Laboratory of Tibetan Environmental Changes and Land Surface Processes, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, China*
[b] *CAS Center for Excellence in Tibetan Plateau Earth Sciences, Beijing, China*
[c] *University of Chinese Academy of Sciences, Beijing, China*
[d] *Northwest Institute of Eco-Environment and Resources, Lanzhou, China*
[e] *School of Geographic Sciences, Xinyang Normal University, Xinyang 464000, China*
[f] *Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL, USA*

HIGHLIGHTS

- The NDSI threshold of 0.1 is more reasonable than that of 0.4 for use in China.
- Terra V6 is comparative with Terra V5, but Aqua V6 is superior to Aqua V5.
- The revised temperature screen algorithm of V6 is found to be problematic in China.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 19 August 2018
Received in revised form 9 October 2018
Accepted 10 October 2018
Available online 13 October 2018

Editor: Ouyang Wei

Keywords:
MODIS snow cover product
Snow depth observation
NDSI threshold
China

ABSTRACT

The new MODIS daily NDSI snow cover product version 6 (V6) is released to replace V5 with significant revisions. This study evaluates, for the first time, the accuracy of product V6 across China based on daily snow-depth measurements during 2003–2013 from 279 and 252 stations for Terra and Aqua, respectively. Three schemes of selecting NDSI thresholds for Terra and Aqua were tested and compared including: (1) the locally optimal NDSI threshold, (2) the minimum valid NDSI of 0.1, and (3) the global reference NDSI threshold of 0.4. The mean Cohen's Kappa (CK) of the optimal, minimum and global reference thresholds for Terra (Aqua) are 0.80 (0.60), 0.77 (0.58), 0.72 (0.51), respectively, while snow depth ≥ 1 cm. The NDSI threshold of 0.1 is demonstrated to be more reasonable than the threshold of 0.4 for use in China. This is also supported by the accuracy comparison conducted for the clear-day snow-cover day calculation. Terra V6 and Terra V5 have comparable accuracies whereas Aqua V6 shows better accuracy than Aqua V5 does. The revised temperature screen algorithm employed in V6 is found to be problematic with large snow commission errors in high altitude stations. Regionally, product V6 presents low CKs of 0.61 and 0.35 for the optimal thresholds of Terra and Aqua in the Tibetan Plateau, which are attributed to its high elevation and relatively small snow depth. This study provides practical implications for use of MODIS snow cover production V6 in China.

* Corresponding author at: Key Laboratory of Tibetan Environmental Changes and Land Surface Processes, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, China.
E-mail address: zhangfan@itpcas.ac.cn (F. Zhang).

https://doi.org/10.1016/j.scitotenv.2018.10.128
0048-9697/© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Snow plays an important role in the global energy balance and atmospheric circulation due to its high albedo and low thermal conductivity (Hall and Riggs, 2007). It also serves as a temporal water storage in mountainous regions and may have a major impact on regional water resources (Barnett et al., 2005; Zhou et al., 2013). Accumulation and melt of snow are important components of the hydrologic cycle for mountainous areas, such as the Mediterranean mountainous regions (Fayad et al., 2017), the western United States (Dong et al., 2014; Franz et al., 2008; Musselman et al., 2017), and the Tibetan Plateau (TP) (Siderius et al., 2013; Xu et al., 2017; Yeo et al., 2017; Zhang et al., 2015; Zhang et al., 2013b). Accurately mapping the spatial distribution of snow is very important for investigating the regional response of hydrologic and environmental systems to climate change (Xiao et al., 2017; Zhang et al., 2004; Zhou et al., 2013) and for water resources management in mountainous basins (Thirel et al., 2013; Zhang et al., 2015). The studies of snow cover in China have covered a number of topics in regard to snow covered area (Dai and Che, 2014; Gao et al., 2012; Liang et al., 2008; Qin et al., 2006; Zhang et al., 2012), −days (Liu and Chen, 2011), snow depth (Che et al., 2008; Zhang and Ma, 2018), and snow phenology (Ke et al., 2016). Most of current studies in large-scale snow variability are based on remote sensing observations. Compared with the limited amount of in situ measurements of snow depth or snow water equivalent (SWE) that are usually provided by sparse meteorological stations, satellite-based snow cover estimates have the potential to produce spatiotemporal patterns of snow cover at a larger scale (Dozier and Painter, 2004; Immerzeel et al., 2009).

In the last several decades, a wide range of remote sensing products for estimating snow cover or SWE have been developed, such as the data derived from the Scanning Multichannel Microwave Radiometer (SMMR) (Chang et al., 1987), Special Sensor Microwave/Imager (SSM/I) (Grody and Basist, 1996), Shuttle Imaging Radar-C and X-Band Synthetic Aperture Radar (SIR-C/X-SAR) (Shi and Dozier, 1997), Advanced Very High Resolution Radiometer (AVHRR) (Simpson et al., 1998), Moderate Resolution Imaging Spectroradiometer (MODIS) (Hall et al., 2002), Advanced Microwave Scanning Radiometer-EOS (AMSR-E) (Kelly et al., 2003), Geostationary Observational Environmental Satellite (GOES) (Romanov and Tarpley, 2003), Chinese Fengyun 3 (FY3) (Che et al., 2016) and those developed by combing several of these products (Che et al., 2008; Huang et al., 2014; Ramsay, 1998; Yu et al., 2016). Among these products, the MODIS daily snow cover products have been extensively employed in a great number of studies regarding snow cover variation (Gao et al., 2012; Tang et al., 2013b; Zhang et al., 2012), snow phenology (Huang et al., 2017; Liu and Chen, 2011) and hydrologic application (Immerzeel et al., 2009; Karsten, 2011; Thirel et al., 2013; Zhang et al., 2015) owing to its relatively high temporal (1-day) and spatial (500-m) resolution as well as the convenience for use. MODIS snow cover data are also involved as a base for creating or improving various combined snow cover products (Gao et al., 2011; Huang et al., 2017; Yu et al., 2016; Zhou et al., 2013), which are spatially cloud-free for continuous snow monitoring. In this way, validation of MODIS snow cover product is thus very important for acquiring the information of accuracy for optimizing the usefulness of MODIS snow cover data (Hall and Riggs, 2007; Parajka and Blöschl, 2012).

Because of the snow spectral characteristics with high reflectance in visible bands and low reflectance in the near infrared (Hall et al., 2002), MODIS snow cover product mainly uses the Normalized Difference Snow Index (NDSI) to distinguish snow from other land cover types (Riggs et al., 2006). The previous versions (e.g., version 5, referred to as V5) of MODIS daily snow cover product provide both binary (i.e., snow or not snow) and fractional snow cover estimates (Riggs et al., 2006). The binary MODIS snow cover data are created using a NDSI threshold of 0.4 in the way that, for a pixel with NDSI ≥0.4, it is labeled as snow. The NDSI threshold is widely used for binary snow cover mapping (Huang et al., 2017; Tang et al., 2013b; Zhang et al., 2013a) and hydrologic applications (Andreadis and Lettenmaier, 2006; Immerzeel et al., 2009; Zhang et al., 2015). Numerous studies have evaluated and validated MODIS binary snow cover data around the world with the overall accuracy reported as 85–99% (Ault et al., 2006; Hall and Riggs, 2007; Klein and Barnett, 2003; Parajka and Blöschl, 2012; Wang et al., 2008).

Recently, a new version (version 6, referred to as V6) of MODIS snow cover product has been released with substantial revisions (Riggs et al., 2017). The new features include: (1) only NDSI snow cover is provided, and either binary or fractional snow cover data are no longer available; (2) a restored band 6 is in place of the previously used band 7 in the calculation of NDSI for Aqua; and (3) certain new data screening methods are adopted. Since product V5 was discontinued on 31 December 2016 (NSIDC, 2017), it is in a great need to evaluate the accuracy of daily NDSI snow cover data based on the latest MODIS product V6 to facilitate future snow cover monitoring. To our knowledge, only a few studies have evaluated product V6. For example, Dong et al. (2014) validated the MODIS daily NDSI snow cover data from product V6 in the United States using the SWE observations from 677 stations. In their study, the NDSI snow cover data were converted to binary snow cover with the NDSI threshold set to 0, which is clearly different from the global threshold of 0.4 used in product V5 (Riggs et al., 2006). A recent study employing the V6 product also follows a similar way with pixels of NDSI >0 identified as snow in the Upper Rio Grande Basin (Huang et al., 2018). The use of 0 as the NDSI threshold is based on the fact that positive NDSI values indicate that there is truly some snow present in the pixel, as indicated by Riggs et al. (2017). Because 0.1 is actually the minimum valid NDSI in product V6 (Riggs et al., 2016), using the NDSI threshold of 0 is equivalent to using the threshold of 0.1. Though the global reference of NDSI threshold being 0.4 is still recommended (Riggs et al., 2016; Riggs et al., 2017), taking different values of the NDSI threshold may improve the accuracy of snow cover data for current and future studies. Besides, Dong et al. (2014) also found that product V6 has a higher accuracy than that of product V5 because of the remarkable refinements of snow detection algorithm in the former. However, it remains suspicious whether product V6 significantly outperforms product V5 since different NDSI thresholds are used for the two products. Lastly, it should be noted that only the snow cover data from Terra were evaluated in previous studies. Considering the obvious revision of using the restored band 6 in snow detecting algorithm for Aqua, it is also necessary to explore the accuracy of Aqua snow cover data from product V6 relative to those from product V5.

The three main objectives of this study are to: 1) evaluate the accuracy of the new MODIS NDSI snow cover product V6 across China; 2) investigate whether 0.1 or 0.4 is more reasonable to be used as the NDSI threshold in China; and 3) reveal the effects of revisions in the snow detecting algorithm of product V6 compared with product V5. To achieve these goals, snow depth observations from 660 stations in China were filtered for validation. Six candidate evaluation metrics were then tested and compared to identify reliable metrics. Products V6 and V5 were validated and compared using three different schemes of NDSI thresholds. Lastly, uncertainties related with snow depth, NDSI thresholds, seasonal variation, cloud cover, and land cover were discussed in detail.

2. Data

2.1. Ground measurements

Daily snow depth observations during 2003–2013 were collected from a total of 660 meteorological stations covering the mainland of

China. The latitude, longitude, altitude and daily observations of six meteorological variables including mean air temperature, wind speed, precipitation, relative humidity, vapor pressure, and sunshine duration were also obtained from China Meteorological Administration (CMA) with the same time span of snow depth. The daily snow depth data were measured in an open ground near the meteorological station using a meter ruler, and the measurements were rounded to the closet integers on 8 a.m. (CMA, 2003). Measurements of snow depth <1 cm were recorded as thin or trace snow depth. Ke et al. (2016) indicates that thin snow depth can reduce data reliability for snow related studies in China. Therefore, records of thin snow depth of <1 cm are not considered in this study. The frequent cloud coverage may largely decrease the amounts of available data when using satellite snow cover products (Yu et al., 2016). To ensure a valid evaluation without cloud coverage in the meantime of maintaining enough stations, the stations with the number of true snow (≥1 cm) or non-snow (0 cm) cases of <20 are not considered following Metsämäki (2016). Finally, 279 and 252 stations are respectively selected for evaluating Terra and Aqua snow cover products due to the different cloud coverage conditions caused by different overpass time (Fig. 1). Among them, 246 stations are available for validation in both Terra and Aqua. Note that extent of the Tibetan Plateau (TP) is also plotted in Fig. 1 because the TP has obviously higher average elevation than other parts of China and its regional validation accuracies are further discussed later.

2.2. MODIS daily snow cover product

Two versions of MODIS daily snow cover products, i.e., V5 and V6, were evaluated in this study. The NDSI used for detecting snow cover is computed by Eq. (1), i.e.,

$$NDSI = (band4 - band6)/(band4 + band6) \tag{1}$$

In V5 product, both binary and fractional snow cover are provided. The MODIS binary snow cover is created using a global threshold of NDSI in the following way: the pixels with the NDSI values ≥0.4 are assigned as “snow”; otherwise they are considered “no snow” and are further set to one of the nine codes indicating missing data, no decision, night, snow-free land, lake, ocean, cloud, detector saturated, and fill (Riggs et al., 2006). MODIS sensors are on board two satellites including Terra (morning) and Aqua (afternoon). It should be noted that band 7 is used in place of band 6 as shown in Eq. (2) for Aqua snow cover product due to the dysfunction of the Aqua MODIS instrument band 6 (Salomonson and Appel, 2006), i.e.,

$$NDSI = (band4 - band7)/(band4 + band7) \tag{2}$$

That is, the NDSI threshold used for Aqua was different with that for Terra. To detect snow in dense vegetation, the Normalized Difference Vegetation Index (NDVI) is also used. For a pixel with the NDSI value between 0.1 and 0.4, it will also be identified as snow if it is located within a NDSI-NDVI space defined following Klein et al. (1998) and has the reflectance in band 2 and band 1 > 0.11 and 0.1, respectively. In addition to the global criteria using the NDSI threshold of 0.4, some screening algorithms were also applied. To separate snow from water, the pixels with the reflectance of band 2 < 0.11 would not pass the screen test. Those with the reflectance of band 4 < 0.1 will also fail in the screen tests to prevent detecting dark targets with high NDSI values as snow (Klein and Barnett, 2003). The last screening test is the temperature screen that the pixels with surface temperature >283 K are considered as snow-free (Riggs et al., 2006). The Terra MODIS fractional snow cover is calculated through a linear regression relationship between the MODIS NDSI and Landsat-observed snow cover fraction as described in Eq. (3) (Salomonson and Appel, 2004). Since band 6 is replaced with band 7, Eq. (4) is proposed for Aqua (Salomonson and Appel, 2006).

$$Fractional\ Snow\ Cover = -0.01 + 1.45 \times NDSI_{Terra} \tag{3}$$

$$Fractional\ Snow\ Cover = -0.64 + 1.91 \times NDSI_{Aqua} \tag{4}$$

In contrast, product V6 has made some significant changes. A major revision is that the binary and fractional snow cover estimates are no longer available and only the NDSI is provided. The valid NDSI codes are within 0–100 indicating that there exists some snow in the pixel (Riggs et al., 2016). The pixels with the other 8 codes may represent

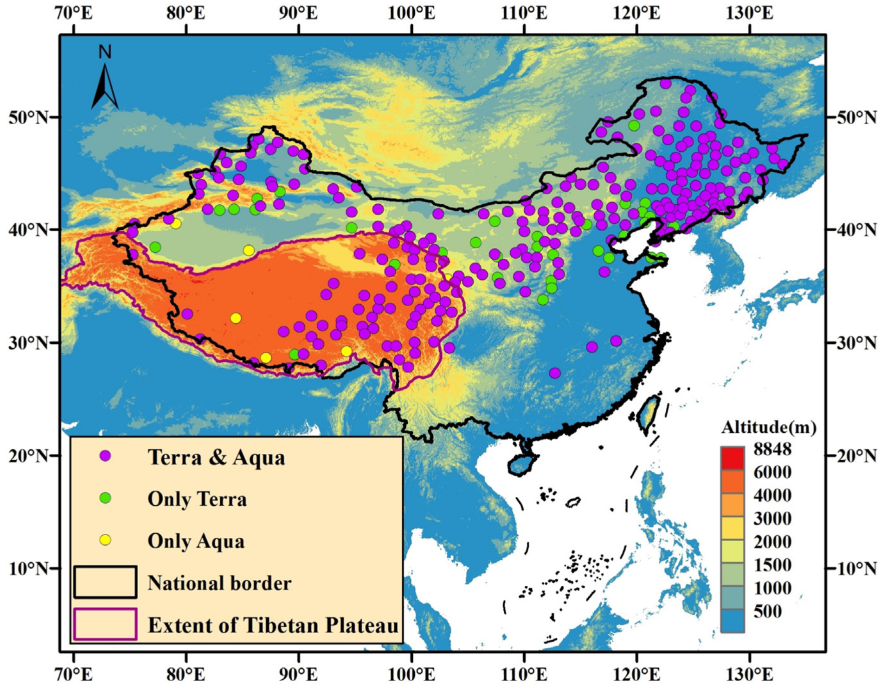


Fig. 1. Study area describing the locations of observation stations. The elevation data were derived from Shuttle Radar Topography Mission (SRTM). The extent of the Tibetan Plateau is also plotted.

missing data, no decision, night, inland water, ocean, cloud, detected saturated and fill (Riggs et al., 2016). Another significant revision is made for Aqua that the restored band 6 following the Quantitative Image Restoration (QIR) algorithm (Gladkova et al., 2012) is used for the NDSI calculation using Eq. (1) (Riggs et al., 2016). Compared with product V5, the screening tests mentioned above remain the same except that the temperature screen is revised to the combined temperature and elevation screen. In product V6, the temperature screen is only applied for "snow" pixels with altitude <1300 m and they will be reversed to non-snow if their estimated brightness temperatures from band 31 are ≥281 K. For locations with altitudes ≥1300 m, the "snow" pixels with surface temperature ≥281 K will not be reversed but be flagged as "warm snow". In addition, two new screening tests are employed including the low shortwave infrared (SWIR) reflectance screen and the low NDSI screen. The pixels with SWIR reflectance >0.45 are considered as non-snow (Riggs et al., 2016). Those with NDSI values <0.1 are taken as non-snow due to the large uncertainty of snow detection for low NDSI values, resulting that the actual minimum NDSI of product V6 is 0.1. These screening methods are all flagged in the layer called "NDSI_Snow_Cover_Algorithm_Flags_QA" of product V6 (Riggs et al., 2016).

The fractional snow cover data are used for retrieving the NDSI values using equations of (3) and (4) for Terra and Aqua of product V5, respectively, because the NDSI data are not provided in product V5. The NDSI data from both products are then converted to binary snow cover data using different schemes of NDSI threshold (see Section 3.3). The original binary snow cover data of product V5 are also used for comparison and are considered equal to using 0.4 as the global NDSI threshold.

3. Methods

A flowchart describing the evaluation process is shown in Fig. 2. The evaluation metrics were first selected by comparing six commonly used metrics. Then, locally optimal NDSI thresholds were calculated to explore the best accuracy of product V6 in comparison to snow depth observations from corresponding stations. Two different NDSI thresholds of 0.1 and 0.4 were compared for product V6 through the absolute validation and a simple application of calculating snow cover days (SCDs). The comparison of products V5 and V6 were also conducted.

3.1. Comparison and selection of evaluation metrics

Six kinds of evaluation metrics are commonly used in previous studies (Dong et al., 2014; Parajka and Blöschl, 2012; Rittger et al., 2013), including the recall (RC), precision (PC), false alarming rate (FAR), overall accuracy (OA), F-score (FS) and Cohen's kappa (CK). All the metrics can be derived on the basis of a confusion matrix as shown in Table 1. RC measures the proportion of correctly detected snow cases by MODIS in the actual snow cases (Rittger et al., 2013). RC is also called probability of detection (POD) in some studies (Dong et al., 2014; Zhou et al., 2013). PC measures the proportion of true snow cases in the detected snow cases by MODIS (Rittger et al., 2013). FAR measures the proportion of false snow cases which are actually not snow detected by MODIS in the actual non-snow cases (Dong et al., 2014). OA means the fraction of the correctly detected cases (snow-snow and land-land) in all cases (Parajka and Blöschl, 2012). As a harmonic mean of RC and PC, FS balances them and is expected to be more useful than both RC and PC (Dong et al., 2014). Kappa is an overall measurement

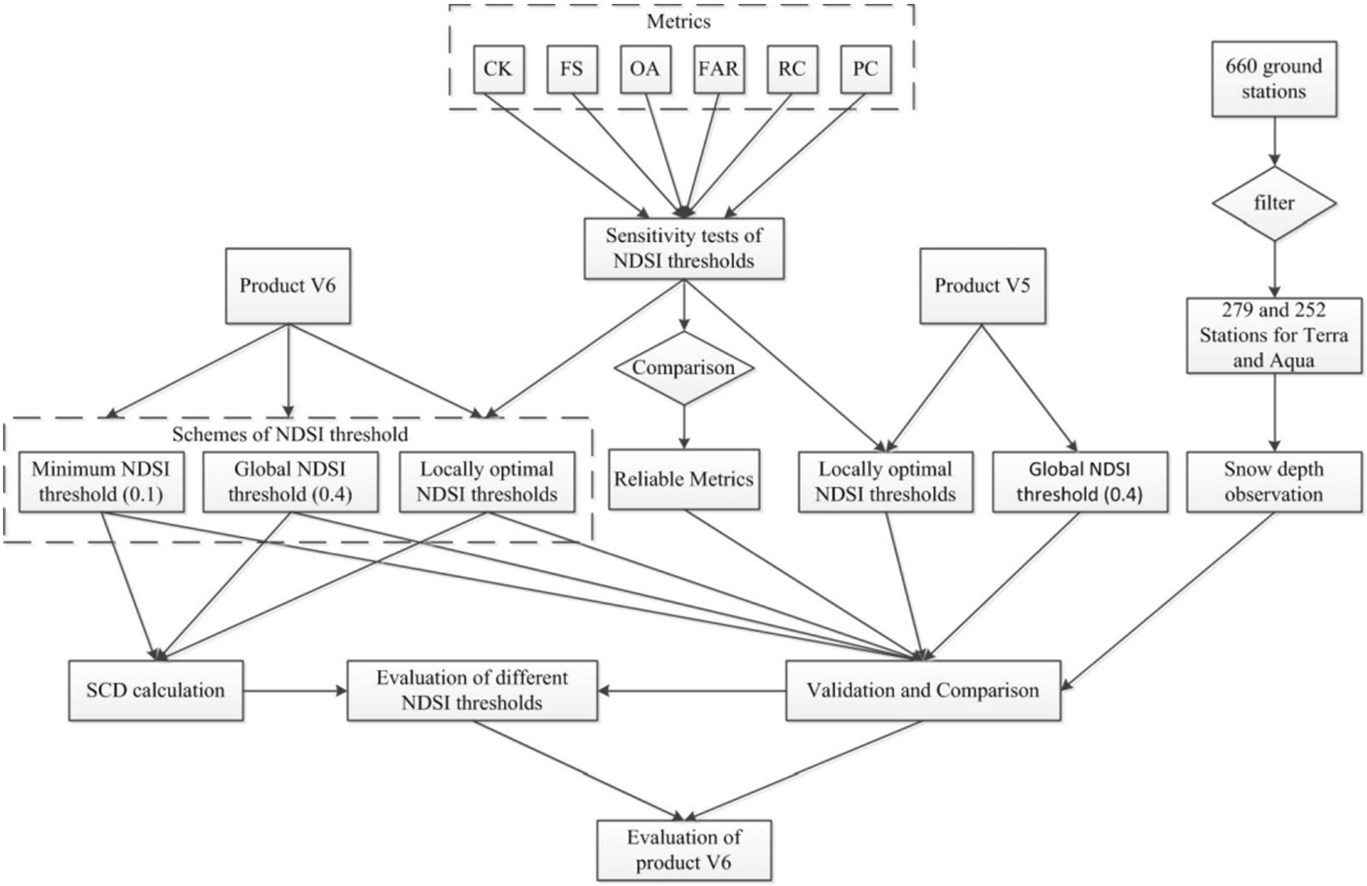

Metrics
CK
FS
OA
FAR
RC
PC
660 ground stations
filter
279 and 252 Stations for Terra and Aqua
Sensitivity tests of NDSI thresholds
Product V6
Product V5
Comparison
Schemes of NDSI threshold
Minimum NDSI threshold (0.1)
Global NDSI threshold (0.4)
Locally optimal NDSI thresholds
Reliable Metrics
Locally optimal NDSI thresholds
Global NDSI threshold (0.4)
Snow depth observation
SCD calculation
Evaluation of different NDSI thresholds
Validation and Comparison
Evaluation of product V6

Fig. 2. A flowchart describing evaluation methods. RC is Recall; PC is Precision; FAR is false alarming rate; OA is overall accuracy; FS is F-score; CK is Cohen's Kappa coefficient; SCD is snow covered day.

Table 1
Description of a confusion matrix for MODIS snow cover estimates versus ground observations and the definition of evaluation metrics.

		MODIS	
		Snow	Non-snow
Actual	Snow	*SS*	*SN*
	Non-snow	*NS*	*NN*
Metric		Definition	
RC		$\frac{SS}{SS+SN}$	
PC		$\frac{SS}{SS+NS}$	
FAR		$\frac{NS}{NN+NS}$	
OA		$\frac{SS+NN}{SS+SN+NS+NN}$	
FS		$\frac{2\times RC\times PC}{RC+PC}$	
CK		$\frac{OA-\ Pr(e)}{1-\ Pr(e)}$	
		Where, $Pr(e) = (\frac{SS+NS}{Total}\times\frac{SS+SN}{Total}) + (\frac{NN+NS}{Total}\times\frac{NN+SN}{Total})$ $Total = SS + SN + NS + NN$	

Note: SS, SN, NS and NN are all numbers, e.g., NS reps the number of cases that MODIS predicts snow covered while the snow depth observation indicates no snow. RC is Recall; PC is Precision; FAR is false alarming rate; OA is overall accuracy; FS is F-score; CK is Cohen's Kappa coefficient.

of agreement among different classifiers (Cohen, 1960), and is considered more meaningful than RC, PC and OA (Powers, 2011). The definitions of the six metrics are described in detail in Table 1.

To evaluate the efficiency of the six metrics, we also conducted a sensitivity test on the NDSI threshold. Specially, the NDSI threshold for creating binary snow cover was increased from 0 to 0.9 at a step of 0.01 resulting in 91 iterations. For each iteration, all the six metrics were calculated using all the available samples at all the stations. Six changing curves from the six metrics were obtained based on which the more suitable metrics can be determined.

3.2. Accuracy assessments based on locally optimal NDSI thresholds

The NDSI threshold may be locally accurate and the global NDSI threshold of 0.4 may not be always optimal (Riggs et al., 2017). To explore the best accuracy of MODIS binary snow cover estimates, the locally optimal NDSI thresholds are calculated as follows: for each station, 91 NDSI thresholds were tested increasing from 0 to 0.9 with an interval of 0.01 and the one having the highest performance metric was taken as the optimal NDSI threshold. There are totally 279 and 252 locally optimal NDSI thresholds for Terra and Aqua, respectively. The locally optimal NDSI thresholds are considered as the “true” NDSI thresholds, which can be further used to evaluate different schemes of global NDSI threshold.

The factors possibly affecting the spatial distribution of optimal NDSI thresholds are then analyzed by conducting a Pearson correlation analysis between ten meteorological and geographic variables and the NDSI threshold from all the stations. The ten variables include longitude, latitude, altitude, snow depth, precipitation, wind speed, air temperature, vapor pressure, relative humidity and sunshine duration.

The accuracies of product V6 including both Terra and Aqua were then assessed based on the locally optimal NDSI thresholds by calculating the reliable metrics selected from Section 3.1. The spatial distribution of validation accuracy was also analyzed using the correlation analysis based on the ten potential meteorological and elevation factors mentioned above.

3.3. Comparing different schemes of NDSI thresholds for product V6

Three schemes of NDSI threshold were used including: (1) the one using the locally optimal NDSI threshold, labeled as “MOD_OPT_V6” and “MYD_OPT_V6” for Terra and Aqua V6, respectively; (2) the one taking the 0.1 (the minimum valid NDSI in V6) as the NDSI threshold, labeled as “MOD_MIN_V6” and “MYD_MIN_V6” for Terra and Aqua V6, respectively; and (3) the one using the global reference NDSI threshold of 0.4, labeled as “MOD_REF_V6” and “MYD_REF_V6” for Terra and Aqua V6, respectively.

For each scheme, the reliable metrics were calculated resulting in 279 and 252 “metric observations” for Terra and Aqua, respectively, which can be further used for multiple comparisons. Statistical tests of multiple comparison were conducted to identify whether there were significant differences between the performance metrics of the three schemes, based on a paired unequal variances *t*-test with Bonferroni correction (Dunnett, 1955; Zhang et al., 2016).

3.4. Evaluation for calculation of clear-day SCDs

The SCDs is an important parameter for snow phenology analysis (Ke et al., 2016). MODIS snow cover data have been used for calculating SCDs in China (Liu and Chen, 2011). The clear-day SCD calculation is thus considered as a good application example to further compare different schemes of NDSI threshold. Their performances on clear-day SCD calculation were then evaluated. The SCD is defined as a day with snow depth of ≥1 cm following Ke et al. (2016). Due to the large proportion of missing data caused by cloud coverage, 11-year (2003—2013) accumulated clear-day SCDs were calculated at each station. The SCDs calculated using observed snow depth data were taken as the truth. The mean absolute error (MAE) was chosen as the evaluation metric. The multiple comparisons of the three schemes were also conducted for both Terra and Aqua following the procedure in Section 3.3.

3.5. Comparison of products V6 and V5

Substantial revisions have been made in product V6 compared with product V5. To examine whether product V6 is better than product V5, two schemes of NDSI threshold were tested for product V5 including: (1) the one using the global reference NDSI threshold of 0.4, labeled as “MOD_REF_V5” and “MYD_REF_V5” for Terra and Aqua, respectively; and (2) the one using the locally optimal NDSI threshold, labeled as “MOD_OPT_V5” and “MYD_OPT_V5” for V5 Terra and Aqua, respectively. They were then compared with MOD_REF_V6, MYD_REF_V6, MOD_OPT_V6 and MYD_OPT_V6. The same multiple comparisons were conducted as described in Sections 3.3 and 3.4. It should be noted that the threshold of 0.1 was not used for the comparison because the locally optimal NDSI threshold already represents the best accuracy that the snow cover products can achieve and the use of the threshold of 0.4 was to make our results comparable with previous studies which use the same threshold. The three new or revised screening algorithms including the combined temperature and height screen, the low NDSI screen and the high SWIR reflectance screen were all evaluated by calculating the snow commission or omission errors.

4. Results and discussions

4.1. The reliable metrics

Comparison results of the six metrics are shown in Fig. 3. The results of Terra (Aqua) are based on 519,953 (434703) daily snow depth observations from 279 (252) stations, of which 43,097 (29283) observations have snow depth ≥1 cm. Since the low NDSI screen is applied in product V6, there is a horizontal line for all the six metrics at the NDSI thresholds of 0–0.1, indicating that the minimum valid NDSI of V6 is actually 0.1. Both OA and FAR seem stable with the variation of NDSI threshold and their largest differences (i.e., the maximum minus the minimum) are merely 0.06 and 0.01, respectively. This is because OA and FAR are greatly affected by the large proportion of actual non-snow cases making them less effective. For example, for a station with 90% of total days snow-free, even all the days are predicted snow-free, the OA is 90% but the evaluated method actually presents no predictive capacity for detecting snow. FAR has the similar problem. It is clear that for the NDSI

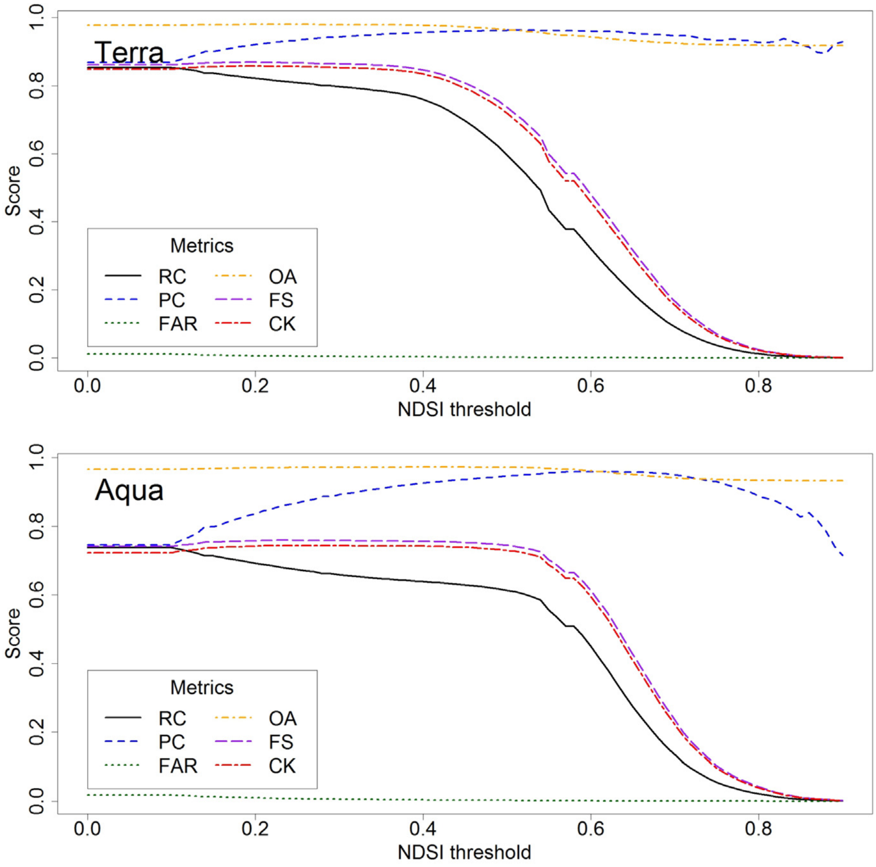


Fig. 3. Comparison of the responses of six kinds of metrics to the variation of NDSI threshold for Terra (upper) and Aqua (lower).

thresholds of 0.1–0.6, RC is always decreasing while PC is always increasing. Actually, RC and PC are closely related with the two well-known kinds of errors, i.e., snow commission (falsely classifying as snow whereas actually snow-free) and omission (falsely classifying as snow-free whereas actually snow covered) errors (Arsenault et al., 2014), respectively. RC plus the commission error equals to 1; PC plus the omission error equals to 1 (Zhou et al., 2013). The lower the NDSI threshold, the more the detected snow cases whereas in the meantime the lower the precision of estimates made by MODIS. RC and PC actually describe the two different aspects of validation accuracy, respectively and it is unreasonable to rely on either of them (Rittger et al., 2013). Compared with RC and PC, FS and CK are expected to balance them and selected as the relatively reliable metrics for comprehensively describing the performance of binary snow cover estimates. In addition, CK is chosen as the decisive metric (e.g. for determining the optimal NDSI threshold) and FS is taken as an auxiliary reference in this study.

4.2. Optimal NDSI thresholds and accuracy of product V6

The locally optimal NDSI threshold are obviously spatial heterogeneous for both Terra and Aqua with the spatially mean ± standard error of 0.17 ± 0.09 and 0.19 ± 0.12, respectively (Fig. 4a and b). This indicates that the global NDSI threshold of 0.4 employed in V5 product may not be appropriate in China because the locally optimal NDSI thresholds are mostly lower than 0.4. This finding is also consistent with several local studies in China with the optimal NDSI thresholds of 0.33 in the Middle Qilian Mountains (Hao et al., 2008) and 0.37 in Qinghai Province (Wang et al., 2012), which both indicate that the global NDSI threshold of 0.4 is on the high side for use in China. It is hard to predict the distribution of the optimal NDSI threshold because no substantial correlation was found between it and the ten climatic and elevation factors mentioned in Section 3.2 with the highest absolute correlation coefficient <0.3.

The CK values of <0, 0–0.2, 0.21–0.4, 0.41–0.6, 0.61–0.8 and 0.81–1 are evaluated as “poor”, “slight”, “fair”, “moderate”, “substantial” and “almost perfect” performances, respectively (Landis and Koch, 1977). The CK for Terra products based on locally optimal NDSI threshold (i.e. MOD_OPT_V6) is generally high with a spatially mean ± standard error of CK as 0.80 ± 0.18 (Fig. 4c). Terra product V6 is therefore demonstrated to have a nearly “almost perfect” performance in China. Similar results are observed for the FS with a spatially mean ± standard error value of 0.81 ± 0.18 (Fig. 4e). For Aqua (i.e. MYD_OPT_V6), the accuracies are obviously lower with the spatially mean CK and FS as 0.60 ± 0.26 (Fig. 4d) and 0.62 ± 0.26 (Fig. 4f), respectively. Though remarkable improvements have been made for Aqua V6 snow cover data (Riggs et al., 2016; Riggs et al., 2017), its accuracy is yet much lower than that of Terra in China.

The altitude, snow depth and latitude seem to have relatively strong influence on the accuracies of product V6 with correlation coefficients all higher than 0.4 for both Terra and Aqua (Fig. 5). High altitude areas generally featured with complex terrains that can add more complexity and heterogeneity of the pixel, which makes it difficult for MODIS snow cover detection. For Terra, the averaged CK for stations with the altitude ranges of 0–2000, 2000–4000 and >4000 m above sea level (a.s.l.) are 0.87, 0.62 and 0.59 respectively. High altitude (>2000 m a.s.l.) stations, accounting for ~25% of total stations, present much lower accuracy. Most of them are located in the Tibetan Plateau, which is referred to as the “Third Pole” due to its high altitudes (Qiu, 2008; Yao et al., 2012). Here, the averaged OA and PC of stations with elevation >2000 m a.s.l. are 97% and 83%, respectively, seemingly indicating a good agreement. However, the averaged RC of them is only 53% resulting a clearly lower averaged FS and CK of 0.63 and 0.61, respectively. This further demonstrates the deficiencies of OA, RC and PC, and the reliability of FS and CK. For Aqua, similar results can be concluded but with much lower accuracies that the averaged CK for high

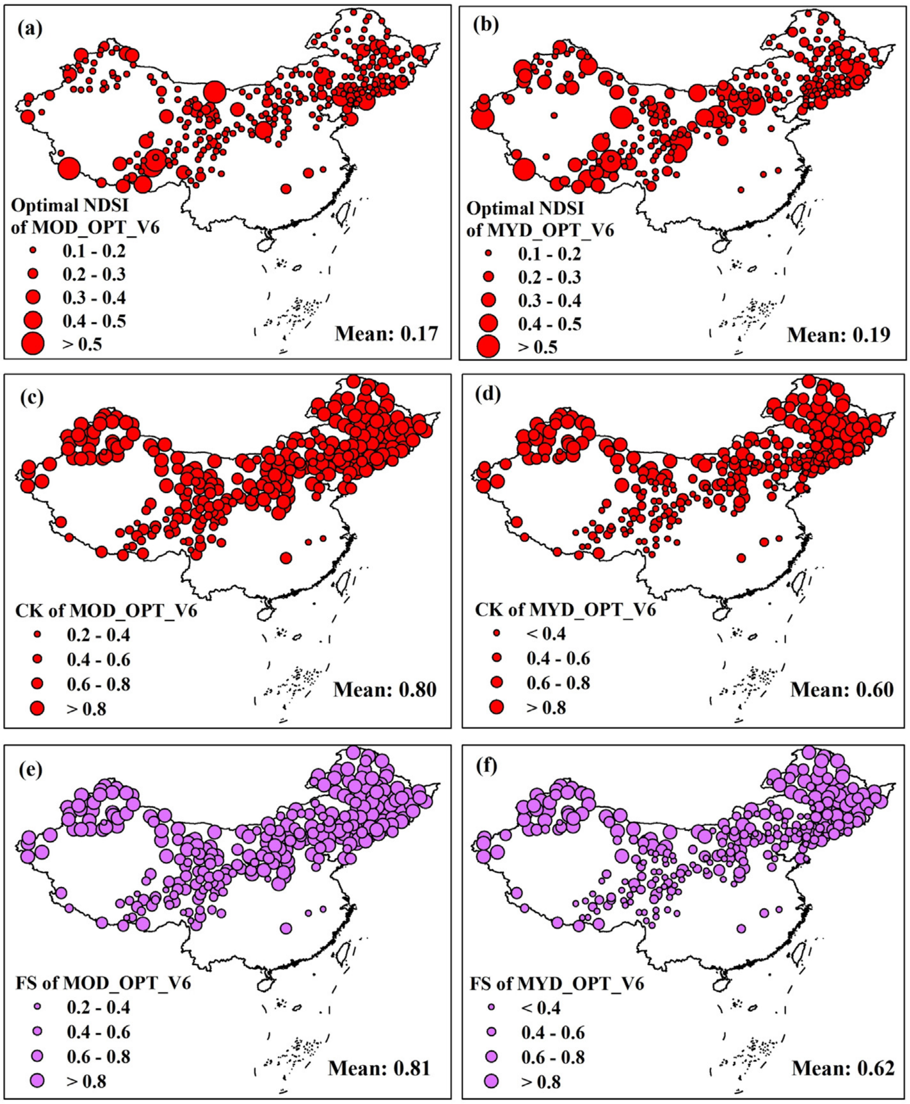


Fig. 4. Spatial distribution of locally optimal NDSI thresholds for Terra (a) and Aqua (b), and accuracies including the CK of Terra product V6 based on locally optimal NDSI thresholds (i.e. MOD_OPT_V6) (c), the FS of MOD_OPT_V6 (e), the CK of MYD_OPT_V6 (d), and the FS of MYD_OPT_V6 (f).

altitude (>2000 m) stations is only 0.35. More discussions focusing on the results of the Tibetan Plateau are presented in Section 4.5.4. Though latitude has a strong correlation with the CK, its effects are considered largely related with altitude. Most of high altitude stations are distributed in the Tibetan Plateau with lower latitudes than the remaining stations in northern China (Fig. 1). Altitude thus has very high negative correlation coefficients (~0.8) with latitude in this study. It is not surprising that snow depth shows strong impacts on the accuracy, considering the widely reported increasing validation accuracy with snow depth, e.g. in Xinjiang (Wang et al., 2008) and Qinghai (Wang et al., 2012) provinces and in the Tibetan Plateau (Yang et al., 2015) in China.

4.3. More efficient NDSI threshold for use in China

It is not surprising that using locally optimal NDSI thresholds (MOD_OPT_V6) has superiority in accuracy (mean CK of 0.8 and 0.61 for Terra and Aqua) over the other two schemes merely using a spatially constant NDSI threshold (Fig. 6). However, it is interesting that using 0.1 as the NDSI threshold (MOD_MIN_V6, mean CK of 0.77 and 0.58 for Terra and Aqua) shows better accuracy than the previously used 0.4 (MOD_REF_V6, mean CK of 0.72 and 0.51 for Terra and Aqua) in China. The paired t-test result shows that their accuracy differences are statistically significant ($p < 0.001$). The results based on FS also support this finding (see the bottom row of Fig. 6). In space, MOD_MIN_V6 shows higher accuracy (i.e., higher CK and FS) than MOD_REF_V6 in ~74% of total stations. For Aqua, MYD_MIN_V6 also shows a better accuracy than MYD_REF_V6 in ~71% of total stations. Thus, a value of 0.1 is considered as a more efficient NDSI threshold than 0.4 for use in China. This finding is important for practical application, because a locally optimal NDSI threshold is usually unknown for a specific area and is hard to predict as discussed in Section 4.2.

The comparison of 11-year clear-day SCD accuracies further validate our hypothesis: MOD_OPT_V6 shows the highest accuracy for calculating SCD followed by MOD_MIN_V6 and MOD_REF_V6, with the MAE values of 20, 28 and 38 days, respectively; their accuracy differences based on stations are statistically significant (Fig. 7). For Aqua, the MAE values increase to 24, 32 and 41 days for MYD_OPT_V6, MYD_MIN_V6 and MYD_REF_V6 respectively. MOD_MIN_V6 and MYD_MIN_V6 present higher accuracies than MOD_REF_V6 and MYD_REF_V6 in 82% and 83% of total stations, respectively.

We further analyzed in which cases MOD_MIN_V6 would be more likely to give more accurate snow cover estimates than MOD_REF_V6. For stations lower than 2000 m, MOD_MIN_V6 has the CK greater than MOD_REF_V6 by 0.04 and for those higher than 2000 m a.s.l., it

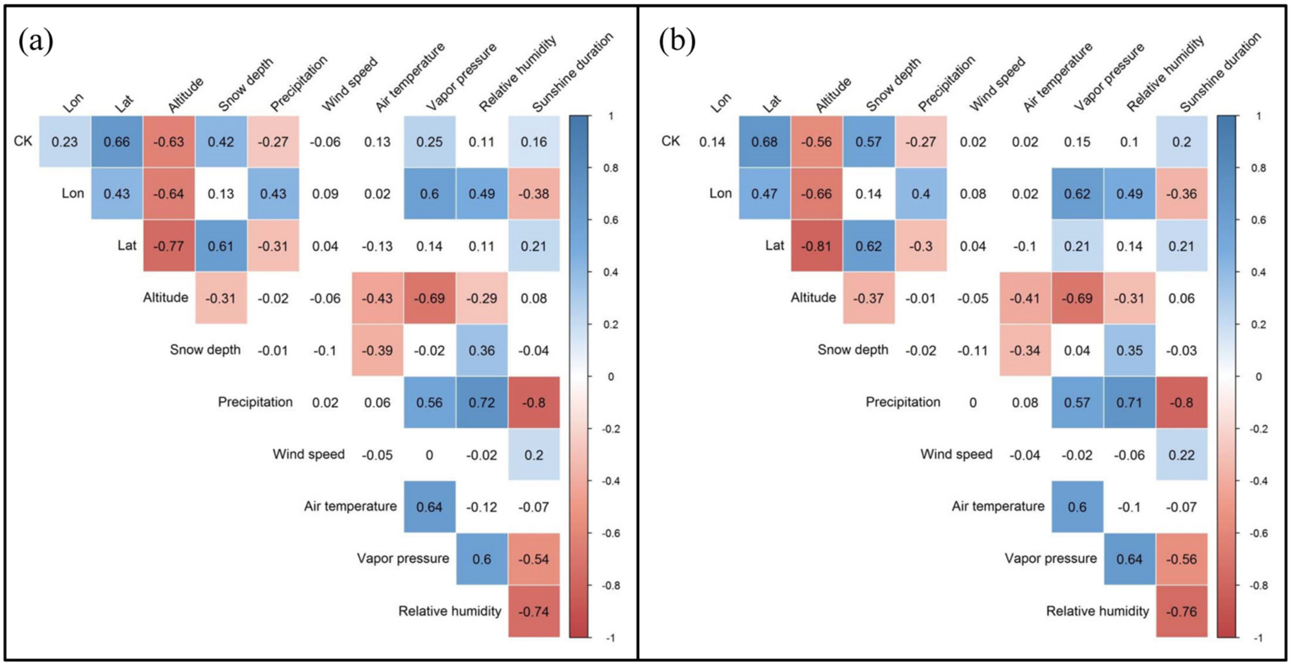


Fig. 5. Correlation matrix of the validation accuracies and ten potential factors for Terra (a) and Aqua (b). Significantly (at 0.01 significance level) positive correlation values are shown in blue; significantly (at 0.01 significance level) negative correlation values are shown in red; insignificant correlation values are filled as blank.

increases to 0.07. Thus, the NDSI threshold of 0.1 may be more appropriate for use in high altitude areas. In addition, snow depth shows some effects that for stations with multi-year averaged daily snow depth <1 cm, those with MOD_MIN_V6 superior to MOD_REF_V6 account for ~71% of the total and it increases to ~80% for stations with multi-year averaged daily snow depth >3 cm.

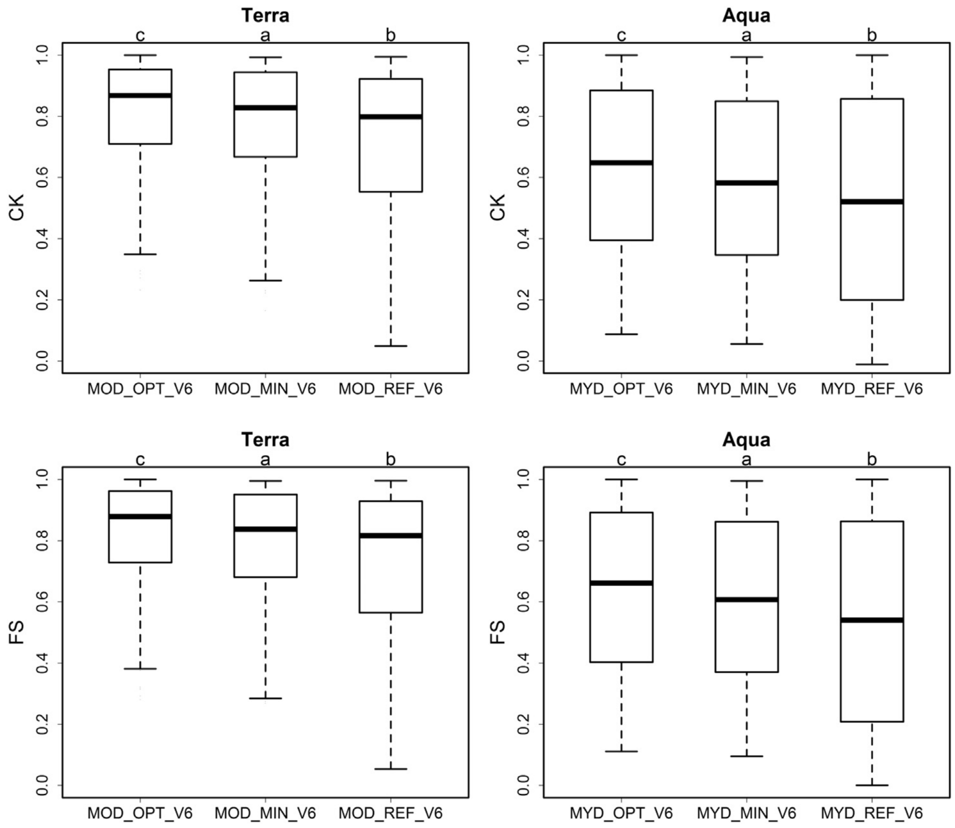


Fig. 6. Comparison of station-based accuracies (CK in upper part and FS in lower part) from the three different NDSI threshold schemes for Terra (left) and Aqua (right) based on paired *t*-tests. Letters at the top indicate the significance of the differences: the schemes with a same letter at the top indicate insignificant difference. The box and whiskers show the distributions of station-based metrics.

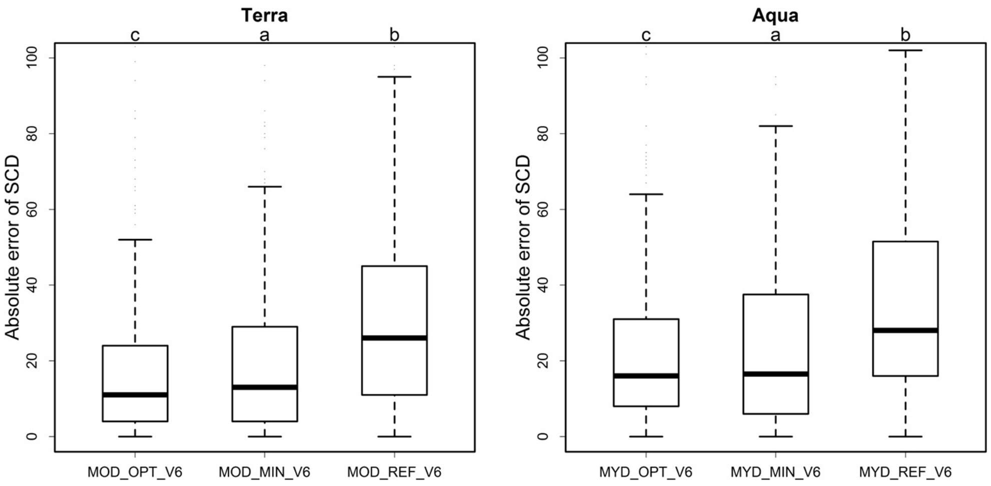


Fig. 7. Comparison of station-based accuracies for calculating SCD from the three different NDSI threshold schemes for Terra (left) and Aqua (right) based on paired *t*-tests. Letters at the top indicate the significance of the differences: the schemes with a same letter at the top indicate insignificant difference. The box and whiskers show the distributions of station-based absolute errors of SCD.

4.4. Is MODIS snow cover product V6 better than product V5 in China?

Fig. 8 presents the comparison results of products V5 and V6 using different NDSI threshold schemes. It can be seen that there is no statistically significant difference between schemes of MOD_OPT_V5 and MOD_OPT_V6 suggesting that the two Terra snow cover products of V5 and V6 almost have identical accuracies when using locally optimal NDSI thresholds. This finding seems conflicting with that from Dong et al. (2014) who indicated that Terra V6 has clearly higher accuracies than Terra V5. However, it should be noted that in their study, the binary snow cover estimates calculated from Terra V6 used 0 as the NDSI threshold (like MOD_MIN_V6 used here) and those from Terra V5 used 0.4 (i.e. MOD_REF_V5 used here) instead. Their findings have also been supported in this study that MOD_MIN_V6 has statistically significantly better accuracies than MOD_REF_V5 as shown in Fig. 8. However, the differences in accuracy of the two versions are not due to the improvements involved in snow detecting algorithm from V5 to V6, but the different values of the NDSI threshold employed to create the binary snow cover.

Another interesting finding is that the CK and FS of MOD_OPT_V5 and MOD_REF_V5 are even slightly larger than those of MOD_OPT_V6 and MOD_REF_V6, respectively (Fig. 8). This is attributed to the revisions of the temperature screening, after a careful examination of the flags for data screening algorithms applied in product V6. In product V5, all "snow" pixels with surface temperature ≥10 K were reversed to be non-snow (Riggs et al., 2006) whereas the "snow" pixels with

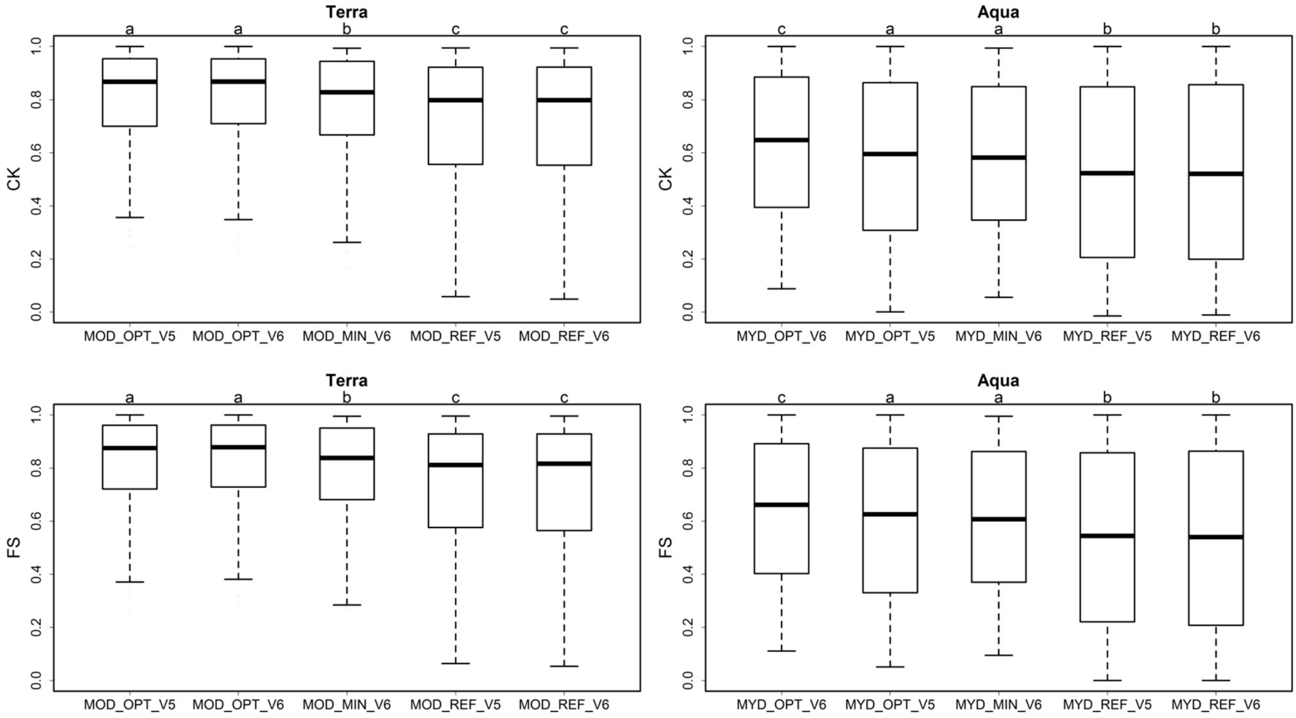


Fig. 8. Comparison of station-based accuracies (CK in upper part and FS in lower part) from the five different NDSI threshold schemes of products V5 and V6 for Terra (left) and Aqua (right) based on paired *t*-tests. Letters at the top indicate the significance of the differences: the schemes with a same letter at the top indicate insignificant difference. The x axis is in order of the average CK or FS. The box and whiskers show the distributions of station-based metrics.

Table 2
Summary of validation results for snow depth ≥ 1 cm.

Scheme		CK	FS	OA	FAR	RC	PC
Terra	MOD_OPT_V6	0.80	0.81	0.98	0.00	0.75	0.92
	MOD_OPT_V5	0.80	0.81	0.98	0.00	0.75	0.93
	MOD_MIN_V6	0.77	0.78	0.98	0.01	0.77	0.84
	MOD_REF_V5	0.73	0.74	0.98	0.00	0.65	0.95
	MOD_REF_V6	0.72	0.73	0.98	0.00	0.64	0.95
Aqua	MYD_OPT_V6	0.61	0.63	0.97	0.01	0.56	0.77
	MYD_OPT_V5	0.58	0.59	0.97	0.01	0.52	0.80
	MYD_MIN_V6	0.58	0.60	0.96	0.02	0.59	0.67
	MYD_REF_V5	0.52	0.53	0.97	0.01	0.46	0.78
	MYD_REF_V6	0.51	0.52	0.97	0.00	0.45	0.82

Note: all the metrics are spatially averaged values based on stations.

elevation >1300 m are considered as “warm snow” even with surface temperature ≥10 K in product V6 (Riggs et al., 2016). Our results show that the total number of “warm snow” cases for Terra is 1045, of which only 118 cases are actually snow-covered and 927 cases are actually snow-free. This leads to the snow commission error rate as high as ~89%. There are 10 high altitude stations with the number of “warm snow” cases ≥20 as plotted in the Supplementary Material (Fig. S1a). Their altitudes are all higher than 1300 m varying from 2065 m to 4672 m. For Aqua, the problem is more serious in that the general snow commission error rate is ~91% (11,191 out of 11,377). There are 30 high altitude stations with the number of “warm snow” cases ≥20 as plotted in the Supplementary Material (Fig. S1b) with altitudes varying from 1331 to 4672 m. In contrast, the snow omission error rate caused by the temperature screening applied for low altitude (<1300 m) stations is only 0.6% and 1.6% for Terra and Aqua, respectively. This indicates that the “warm snow” pixels flagged in product V6 may be problematic for high altitude stations used in our study and most of them should be reversed to non-snow like what are done for stations lower than 1300 m. However, warm snow has been widely identified in mountainous areas especially at spring and product V5 is considered to fail in detecting them due to the temperature screening applied for all elevations (Riggs et al., 2016).

The two new screening methods employed in product V6 were proven to be efficient in China. For Terra V6, the snow omission error rates of the low SWIR reflectance screen and the low NDSI screen are only 13% and 14%, respectively. They presented even better performances in Aqua V6 with lower snow omission error rates of 10% and 7%, respectively. The high efficiency of the low NDSI screening also indicates that most of cases within the NDSI ranges of 0–0.1 are truly non-snow, further supporting that MOD_MIN_V6 (or MYD_MIN_V6) selecting 0.1 as the NDSI threshold is reasonable.

Lastly, Aqua V6 has better accuracy than Aqua V5 as shown in Fig. 8 that MYD_OPT_V6 has higher averaged CK (0.61) and FS (0.63) than those of MYD_OPT_V5 (0.58 and 0.59) and their differences based on stations are statistically significant. The accuracy improvement for Aqua may be due to the restored band 6 using the QIR method as mentioned above. It may be surprising that MYD_REF_V6 does not show better accuracy than MYD_REF_V5, largely because that MYD_REF_V5 uses band 7 in place of band 6 for calculating the NDSI and its NDSI threshold is actually not 0.4 but ~0.34 (based on our data analysis).

The validation results of all the ten schemes from both Terra and Aqua shown in Fig. 8 are summarized in Table 2. In addition to CK and FS, the other four kinds of metrics including OA, FAR, RC and PC are also provided for reference and comparison with other validation studies.

4.5. Discussion

4.5.1. Strong effects of snow depth

Measured snow depth greater than a specific threshold is generally taken as the truth in validation and the threshold varies with studies. The 1 cm is used as the snow depth threshold here similar to Parajka and Blöschl (2006), Ault et al. (2006), Wang et al. (2008) and Huang et al. (2011). Some other studies chose 0 cm (Dong et al., 2014; Yang et al., 2015), 0.25 cm (Gao et al., 2011), 2 cm (Metsämäki, 2016) and 2.54 cm (i.e. 1 in. (Arsenault et al., 2014; Maurer et al., 2003). The snow depth of 0–1 cm is generally considered as trace introducing large uncertainties possibly due to more susceptibility to the time difference between satellite and ground observations, more patchy vegetation, higher possibility of erroneously classifying thin snow as clouds, etc. (Ault et al., 2006; Hall and Riggs, 2007; Ke et al., 2016). Therefore, we also tested larger snow depth thresholds (Fig. 9). Generally, better accuracies (higher CKs) are obtained with increasing thresholds of observed snow depth for both Terra and Aqua. When the snow depth threshold increases from 1 cm to 3 cm, the spatially averaged CK reaches 0.90 (0.79) (Table 3) with an increase from 0.80 (0.61) (Table 2) for MOD_OPT_V6 (MYD_OPT_V6). This indicates that validation accuracy is particularly sensitive to the selected snow depth threshold, which should be considered when comparing results from different validation studies. The results are consistent with several previous studies reporting the same accuracy improvements with increasing snow depth (Klein and Barnett, 2003; Wang et al., 2008; Yang et al., 2015). The strong effect of snow depth threshold may be due to two reasons: (1) the snow/cloud confusion error is most likely occurring when snow is shallow (Riggs et al., 2006); and (2) small snow depth may not be able to represent the “truth” in some situations such as that the snow measured at morning melts at the overpass time of satellite and that snow does not cover a large part (>50%) of the pixel. It is hard to determine a suitable threshold of observed snow depth for validating MODIS snow cover products because the snow depth does not equal to or is not closely associated with snow cover fraction and their relationship may be complex. The correlation between MODIS NDSI and observed snow depth was analyzed and a relatively low correlation coefficient (<0.4) was found. Considering that most of previous studies took 1 cm as the threshold and that it may represent a wider range of snow conditions, previous results of this study were based on snow depth ≥1 cm.

The threshold of snow depth can also affect the locally optimal NDSI thresholds in that they are getting higher with the increasing snow depth (Fig. 9b and d). It should be noted that the locally optimal NDSI thresholds were calculated individually for each threshold of snow depth. Locally optimal NDSI thresholds of snow depth ≥6 cm are higher than those of ≥1 cm by 0.08 and 0.11 on average for Terra and Aqua, respectively. Though they get larger, most of them are still much lower than 0.4.

4.5.2. Suggested NDSI threshold of 0.1

This study suggests 0.1 as the NDSI threshold for use in China, in place of the 0.4 used in product V5. The NDSI threshold of 0.1 generally shows better accuracies than using 0.4 as discussed in Section 4.3. However, it should be noted that such results are based on observations with snow depth ≥1 cm. As shown in Fig. 9, when the snow depth increases to 3 cm, using 0.4 as the NDSI threshold starts to present higher accuracy than using 0.1. The accuracy differences between MOD_MIN_V6 and MOD_REF_V6 are generally small (the maximum difference of CK < 0.02) for snow depth thresholds of 3–6 cm. To help identify whether significant differences exist, a paired unequal variances *t*-test was conducted for each snow depth threshold and the result indicates that the accuracy differences between MOD_MIN_V6 and MOD_REF_V6 for snow depth thresholds of 3–6 cm are actually insignificant (Fig. S2). For Aqua, situations are slightly different in that MYD_REF_V6 shows significantly better accuracies than MOD_MIN_V6 for snow depth thresholds of 4–6 cm (Fig. S2), indicating that the NDSI threshold of 0.4 is more suitable for deep snow than the threshold of 0.1 when processing Aqua product V6. This is also supported by the spatially averaged optimal NDSI thresholds of approximately 0.3 at snow depth thresholds of 4–6 cm for Aqua (Fig. 9d) indicating that a higher (than

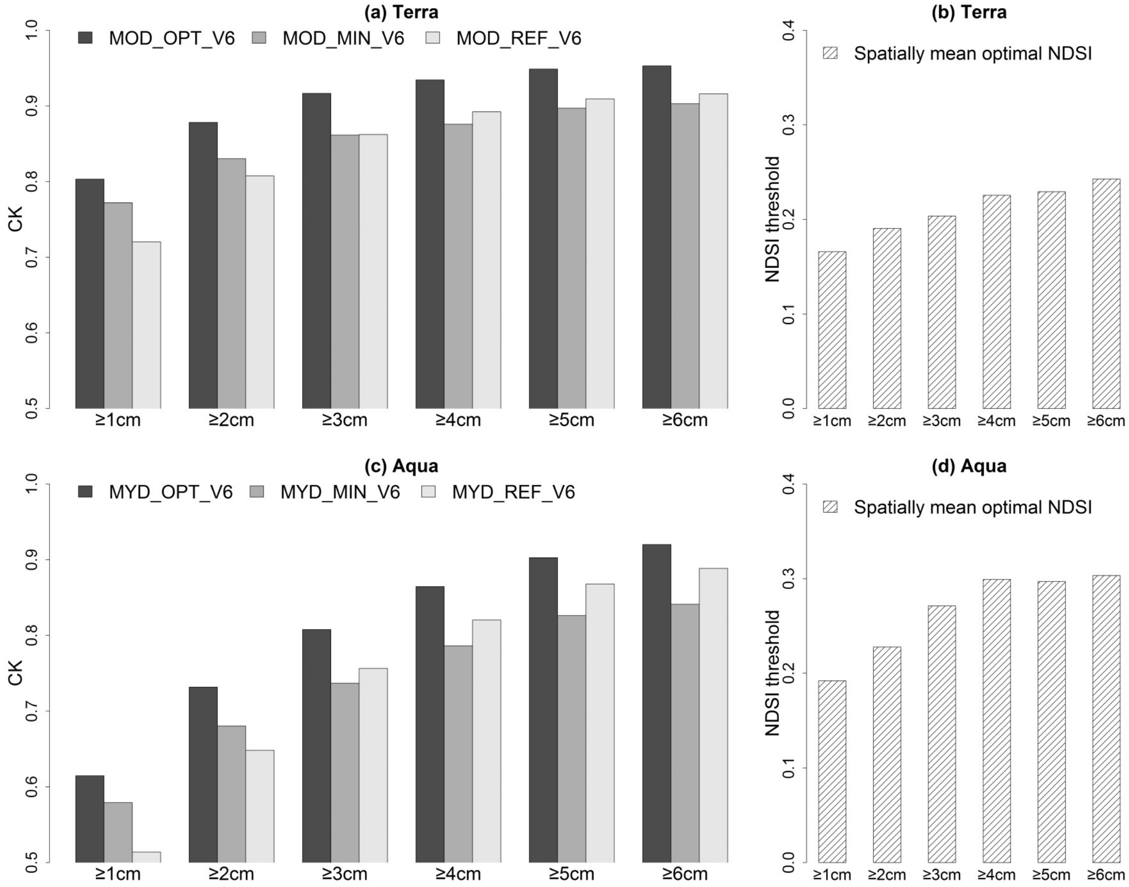


Fig. 9. Comparison of validation accuracies (CK) using different thresholds of snow depth for Terra (a) and Aqua (c), and the corresponding spatially mean optimal NDSI for Terra (b) and Aqua (d).

0.1) NDSI threshold should be used. Thus, there is a tradeoff between MOD_REF_V6 (MYD_REF_V6) and MOD_MIN_V6 (MYD_MIN_V6) that the latter scheme can capture more shallow snow cases whereas the former one can reduce uncertainty related with shallow or thin snow such as patchy snow and snow/cloud confusion.

For most of practical situations, the locally optimal NDSI threshold is unknown and a globally/regionally suggested one is needed. It must be admitted that the proposed 0.1 is not a globally optimal NDSI threshold for use in China because the optimal one seems to be higher than 0.1 as shown in Fig. 9b and d. However, the present study gives some insights for reexamining the reasonability and efficiency of the previously used 0.4. Based on the validation accuracies for snow depth thresholds of 1 and 2 cm, and their performances on SCD calculation, we conclude that using 0.1 is more suitable than using 0.4 as the NDSI threshold in China.

Table 3
Summary of validation results for snow depth ≥ 3 cm.

Scheme		CK	FS	OA	FAR	RC	PC
Terra	MOD_OPT_V6	0.90	0.91	0.99	0.00	0.91	0.91
	MOD_OPT_V5	0.91	0.91	0.99	0.00	0.91	0.92
	MOD_MIN_V6	0.86	0.87	0.98	0.01	0.92	0.85
	MOD_REF_V5	0.87	0.87	0.99	0.00	0.83	0.95
	MOD_REF_V6	0.86	0.87	0.99	0.00	0.83	0.96
Aqua	MYD_OPT_V6	0.79	0.79	0.98	0.01	0.81	0.80
	MYD_OPT_V5	0.79	0.80	0.99	0.01	0.78	0.83
	MYD_MIN_V6	0.74	0.75	0.97	0.02	0.82	0.72
	MYD_REF_V5	0.76	0.76	0.98	0.01	0.73	0.85
	MYD_REF_V6	0.76	0.76	0.98	0.01	0.72	0.88

Note: all the metrics are spatially averaged values based on stations.

4.5.3. Seasonal variation, land cover, cloud coverage and other uncertainties

Monthly accuracies are shown in Fig. 10. Because the number of station for each month are greatly decreased due to the filtering process mentioned above, only five months, November–March, were considered to have sufficient number of stations to provide spatially representative validation results. The validation accuracies of November and December present clearly higher accuracies than those of January, February and March for all the three NDSI threshold schemes, and for both Terra and Aqua. It is clear that MOD_MIN_V6 (MYD_MIN_V6) is always superior to MOD_REF_V6 (MYD_REF_V6) for all the five months.

Land cover may be another important factor according to previous studies (Hall and Riggs, 2007). They indicate that MODIS has more difficulty in areas with dense vegetation (Riggs et al., 2006) resulting lower accuracy (Hall et al., 2001; Yang et al., 2015). MODIS snow cover data are considered to have the lowest accuracies in forested areas and very high accuracies in cropland or agriculture areas (Hall et al., 2001; Hall and Riggs, 2007). According to MODIS land cover product (MCD12Q1), there are mainly three land cover types for stations used in this study including the “cereal crops” (accounting for 14%), grassland (38%) and “urban and built-up” areas (36%), with the averaged CKs of 0.92, 0.72 and 0.85, respectively. No other land cover type has the station number >10. Due to the limited amount of stations, the important land cover type of forests lack enough observations to conduct a reliable evaluation here and it need to be furthered investigated in future study.

The averaged cloud cover fractions of stations used during the study period are as high as 52% and 55% for Terra and Aqua, respectively. The 11-year ground observations of snow depth were thus obtained to

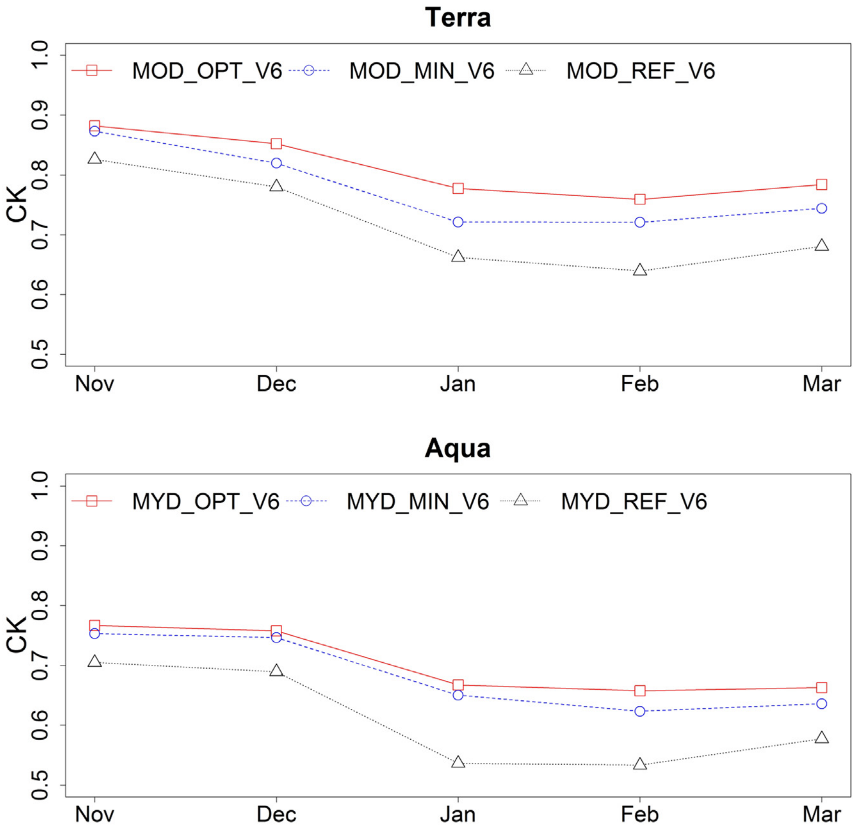


Fig. 10. Monthly validation accuracies (CK) of product V6 for Terra (upper) and Aqua (lower).

provide sufficient data for a reliable validation. The snow/cloud confusion is a known problem in MODIS snow cover products because that some thin snow may be flagged as clouds, especially at snow cover edges and that sub-pixel contaminated clouds may be identified as snow, especially at cloud coverage margins, due to similar spectral characteristics (Riggs et al., 2016; Riggs et al., 2006; Tang et al., 2013a). To alleviate this problem, observations with thin snow depth (<1 cm) were removed in this study. In addition, the correlation analysis shows that the cloud fraction has a relatively weak correlation coefficient (<0.25) with validation accuracies (CKs).

The validation method using ground observed snow depth intrinsically have some deficiencies such as the time difference between satellite overpass and ground measuring time, as well as the spatial representativeness (Hall and Riggs, 2007). It is generally impossible to achieve the ideal conditions that the snow depth data from ground stations are measured at the exact same time of satellite overpass and that its area covers at least 50% of the pixel as indicated by Hall and Riggs (2007). Future study may make use of other satellite data with higher resolution such as Landsat images (Crawford, 2015; Huang et al., 2011) which can provide more detailed snow observations to combine with snow depth observations to present more comprehensive evaluation and validation of product V6 for use in China.

The filtering method employed on stations may impose some uncertainty. Considering the quickly decreasing number of available stations with the increasing filtering threshold as shown in the Supplementary Material (Fig. S3b and S3d), a filtering threshold of 20 was used in present study following Metsämäki (2016). Fig. S3a and S3c show that the validation accuracy of both Terra and Aqua snow cover data increases with the filtering threshold. Generally, its effect on Terra snow cover data is small with the biggest accuracy (CK) change between the filtering thresholds of 10 and 100 about 0.06, whereas that on Aqua snow cover data is relatively large with the highest accuracy (CK) change ~0.16.

4.5.4. More challenges for the Tibetan Plateau

The dramatic elevation difference between the stations within the TP and those outside the TP results in a sharp contrast between their validation accuracies due to the strongly negative correlation between altitude and validation accuracy as described in Section 4.2. The evaluation metrics of CK and FS were calculated for each station and all the stations were divided into two groups including one containing the stations within the TP and the other containing those outside the TP. The average metric value was then calculated for each group. According to MOD_OPT_V6 (mean CK of 0.61) and MYD_OPT_V6 (mean CK of 0.35), the averaged CK (FS) of stations within the TP are obviously lower than that of stations outside the TP by 0.25 (0.24) and 0.36 (0.34) for Terra and Aqua, respectively (Fig. 11). This seems somewhat different from a high OA (>93%) of product V5 reported in the TP (Yang et al., 2015). However, the OA is considered biased and less effective as discussed in Sections 3.1 and 4.1 as well as other validation studies (Metsämäki, 2016; Rittger et al., 2013). The averaged OAs of MOD_REF_V6 and MYD_REF_V6 using stations within the TP are also as high as 0.96 and 0.95, respectively, whereas the significantly lower CK, FS and RC values once again demonstrate that OA is not a reliable evaluation metric (Fig. 11).

In addition to the clearly higher altitude, another cause of the remarkably lower accuracy for the TP may be related with the snow depth. The multi-year average daily snow depth of all stations within the TP is 0.3 (0.3) cm whereas that of stations outside the TP is 1.8 (2.0) cm for validating Terra (Aqua) V6. This indicates that small snow depth is more frequently occurring in the TP compared with the other regions because of overall low precipitation in winter (Wang et al., 2017; Zhang and Ma, 2018). Small snow depth may not allow snow to

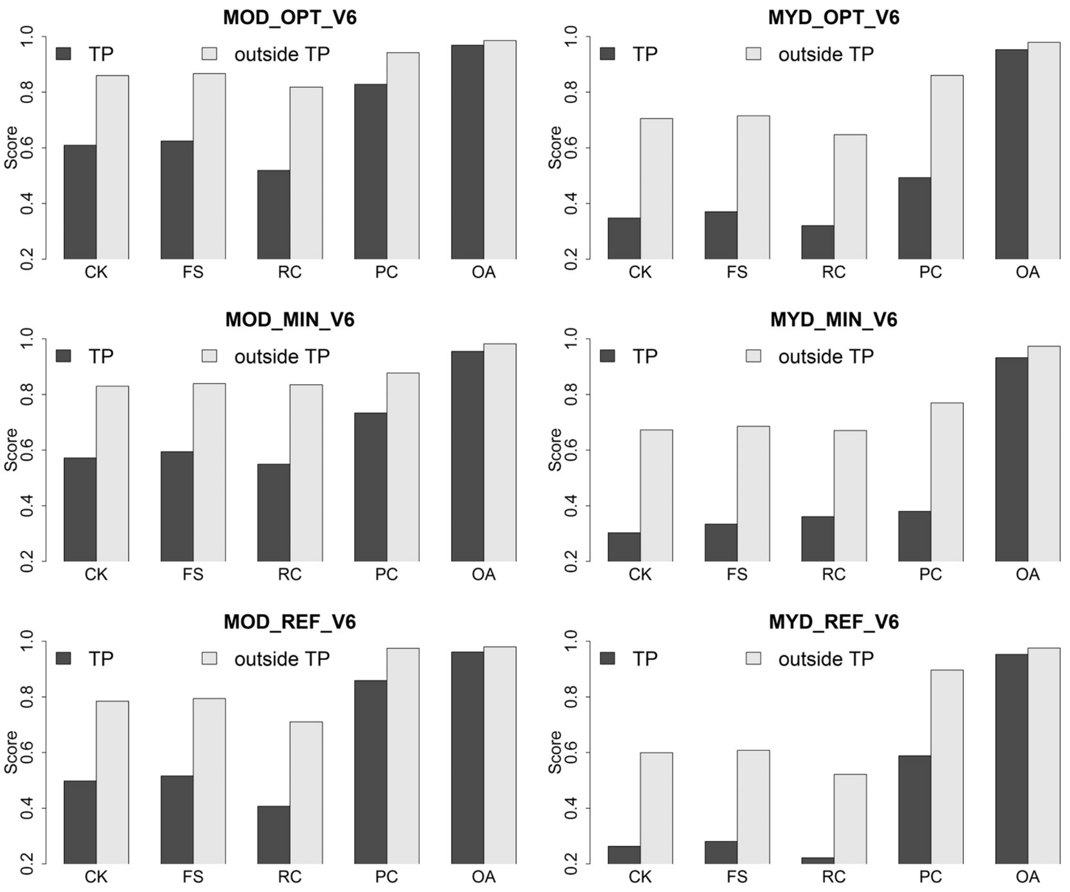


Fig. 11. Comparison of validation accuracies including the CK, FS, RC, PC and OA between stations within the Tibetan Plateau (TP) and those outside the TP.

cover more than half of a pixel, its NDSI value can be very low, and small snow depth is more subject to be affected by snow melt (Parajka and Blöschl, 2006) or snow sublimation (Ma et al., 2015) during the time interval between satellite overpass and ground measurement. These issues make it more difficult for snow detection from MODIS resulting a lower validation accuracy as shown in Fig. 11.

Increasing the snow depth threshold can reduce the uncertainty related to shallow snow, however, even for cases with snow depth ≥ 3 cm, the mean CK values of stations within the TP are much lower than those of the stations outside the TP by 0.12 and even 0.3 for Terra and Aqua, respectively. This indicates that simply modifying the NDSI threshold may have limited effects and fundamental improvements need to be made in the MODIS snow detecting algorithm for complex terrain areas such as the TP. In addition, the Supplementary Material (Fig. S1) shows that most of the stations with high snow commission error rates caused by the "warm snow" flags are located on the TP because of its much higher altitudes. Thus, we consider it challenging to apply the temperature screening algorithm for use in China, especially the TP. The "warm snow" pixels flagged by product V6 may particularly need careful reexamination for the TP and more efficient temperature screening methods need to be investigated in future study.

5. Conclusion

This study evaluates the accuracies of the newly released MODIS daily NDSI snow cover product V6 in China, based on daily snow depth observations during 2003–2013 from a total of 279 and 252 stations for evaluation of Terra and Aqua, respectively. With a higher validation accuracy for both Terra and Aqua, and for both versions of V6 and V5, the NDSI threshold of 0.1 is demonstrated to be more reasonable than the global NDSI threshold of 0.4 for use in China. This finding is also supported by the comparison results of their performances in 11-year clear-day SCDs calculation.

Terra product V6 presents a high accuracy with the averaged CK and FS of 0.80 and 0.81, respectively. However, Aqua product V6 shows much lower accuracies with the averaged CK and FS of 0.60 and 0.62, respectively, though Aqua product V6 has made significant revisions. No significant accuracy improvements are found between Terra products V6 and V5, whereas Aqua product V6 truly shows better accuracies than Aqua product V5 due to the use of the restored band 6. The revised temperature screen method employed in product V6 is found to be problematic in high altitude areas of China. The two new screening algorithms of the low NDSI screen and the low SWIR screen are demonstrated to be efficient.

Altitude and snow depth are found to be the two major factors affecting the spatial distribution of validation accuracy that a higher altitude or a smaller snow depth tend to produce a lower accuracy. Thus, product V6 presents much lower accuracy in the TP, due to its higher mean altitude (~4000 m a.s.l.) and shallower daily snow depth (~0.3 cm). The thresholds of snow depth used in evaluation are found to have a large effect on validation accuracy. The highest accuracies are found for the land cover of cereal crops, followed by grassland and urban areas. Accuracies in November and December are higher than those in January, February and March.

Since the forward processing of product V5 has been discontinued in 2016, product V6 would inevitably be applied in further studies. This study, for the first time, provides an independent evaluation of product V6 in China, thereby providing some insights for selecting a reasonable NDSI threshold in China, which will contribute to the optimal use of product V6 in future studies.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Grant No. 41701079, 41771090 and 41761144078) and China Postdoctoral Science Foundation (2017M620069 and 2017M611013). The authors are grateful to the Chinese Meteorology Administration (http://data.cma.cn/) for providing snow depth data. The MODIS LST products can be downloaded from the Earthdata website (http://earthdata.nasa.gov/).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2018.10.128.

References

Andreadis, K.M., Lettenmaier, D.P., 2006. Assimilating remotely sensed snow observations into a macroscale hydrology model. Adv. Water Resour. 29, 872–886.

Arsenault, K.R., Houser, P.R., De Lannoy, G.J., 2014. Evaluation of the MODIS snow cover fraction product. Hydrol. Process. 28, 980–998.

Ault, T.W., Czajkowski, K.P., Benko, T., Coss, J., Struble, J., Spongberg, A., Templin, M., Gross, C., 2006. Validation of the MODIS snow product and cloud mask using student and NWS cooperative station observations in the Lower Great Lakes Region. Remote Sens. Environ. 105, 341–353.

Barnett, T.P., Adam, J.C., Lettenmaier, D.P., 2005. Potential impacts of a warming climate on water availability in snow-dominated regions. J. Nature 438, 303.

Chang, A., Foster, J., Hall, D.K., 1987. Nimbus-7 SMMR derived global snow cover parameters. Ann. Glaciol. 9, 39–44.

Che, T., Li, X., Jin, R., Armstrong, R., Zhang, T.J., 2008. Snow depth derived from passive microwave remote-sensing data in China. In: Schneebeli, M. (Ed.), Annals of Glaciology. Vol 49. Int Glaciological Soc, Cambridge, pp. 145–154.

Che, T., Dai, L., Zheng, X., Li, X., Zhao, K., 2016. Estimation of snow depth from passive microwave brightness temperature data in forest regions of northeast China. Remote Sens. Environ. 183, 334–349.

CMA, 2003. Specifications for Surface Meteorological Observations. China Meteorological Press, Beijing.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46.

Crawford, C.J., 2015. MODIS Terra Collection 6 fractional snow cover validation in mountainous terrain during spring snowmelt using Landsat TM and ETM+. Hydrol. Process. 29, 128–138.

Dai, L., Che, T., 2014. Spatiotemporal variability in snow cover from 1987 to 2011 in northern China. J. Appl. Remote. Sens. 8, 084693.

Dong, J., Ek, M., Hall, D., Peters-Lidard, C., Cosgrove, B., Miller, J., Riggs, G., Xia, Y., 2014. Using air temperature to quantitatively predict the MODIS fractional snow cover retrieval errors over the continental United States. J. Hydrometeorol. 15, 551–562.

Dozier, J., Painter, T.H., 2004. Multispectral and hyperspectral remote sensing of alpine snow properties. Annu. Rev. Earth Planet. Sci. 32, 465–494.

Dunnett, C.W., 1955. A multiple comparison procedure for comparing several treatments with a control. J. Am. Stat. Assoc. 50, 1096–1121.

Fayad, A., Gascoin, S., Faour, G., López-Moreno, J.I., Drapeau, L., Le Page, M., Escadafal, R., 2017. Snow hydrology in Mediterranean mountain regions: a review. J. Hydrol. 551, 374–396.

Franz, K.J., Hogue, T.S., Sorooshian, S., 2008. Operational snow modeling: addressing the challenges of an energy balance model for National Weather Service forecasts. J. Hydrol. 360, 48–66.

Gao, Y., Lu, N., Yao, T., 2011. Evaluation of a cloud-gap-filled MODIS daily snow cover product over the Pacific Northwest USA. J. Hydrol. 404, 157–165.

Gao, J., Williams, M.W., Fu, X.D., Wang, G.Q., Gong, T.L., 2012. Spatiotemporal distribution of snow in eastern Tibet and the response to climate change. Remote Sens. Environ. 121 (1–9).

Gladkova, I., Grossberg, M., Bonev, G., Romanov, P., Shahriar, F., 2012. Increasing the accuracy of MODIS/Aqua snow product using quantitative image restoration technique. IEEE Geosci. Remote Sens. Lett. 9, 740–743.

Grody, N.C., Basist, A.N., 1996. Global identification of snowcover using SSM/I measurements. IEEE Trans. Geosci. Remote Sens. 34, 237–249.

Hall, D.K., Riggs, G.A., 2007. Accuracy assessment of the MODIS snow products. Hydrol. Process. 21, 1534–1547.

Hall, D.K., Foster, J.L., Salomonson, V.V., Klein, A.G., Chien, J.Y.L., 2001. Development of a technique to assess snow-cover mapping errors from space. IEEE Trans. Geosci. Remote Sens. 39, 432–438.

Hall, D.K., Riggs, G.A., Salomonson, V.V., Digirolamo, N.E., Bayr, K.J., 2002. MODIS snow-cover products. Remote Sens. Environ. 83, 181–194.

Hao, X.H., Wang, J., Hong-Yi, L.I., 2008. Evaluation of the NDSI threshold value in mapping snow cover of MODIS——a case study of snow in the middle Qilian Mountains. J. Glaciol. Geocryol. 30, 132–138.

Huang, X., Liang, T., Zhang, X., Guo, Z., 2011. Validation of MODIS snow cover products using Landsat and ground measurements during the 2001–2005 snow seasons over northern Xinjiang, China. Int. J. Remote Sens. 32, 133–152.

Huang, X., Hao, X., Feng, Q., Wang, W., Liang, T., 2014. A new MODIS daily cloud free snow cover mapping algorithm on the Tibetan Plateau. Sci. Cold Arid Reg. 6, 0116–0123.

Huang, X., Deng, J., Wang, W., Feng, Q., Liang, T., 2017. Impact of climate and elevation on snow cover using integrated remote sensing snow products in Tibetan Plateau. Remote Sens. Environ. 190, 274–288.

Huang, Y., Liu, H., Yu, B., Wu, J., Kang, E.L., Xu, M., Wang, S., Klein, A., Chen, Y., 2018. Improving MODIS snow products with a HMRF-based spatio-temporal modeling technique in the Upper Rio Grande Basin. Remote Sens. Environ. 204, 568–582.

Immerzeel, W.W., Droogers, P., de Jong, S.M., Bierkens, M.F.P., 2009. Large-scale monitoring of snow cover and runoff simulation in Himalayan river basins using remote sensing. Remote Sens. Environ. 113, 40–49.

Karsten, L.R., 2011. Investigation of MODIS Snow Cover Products for Use in Streamflow Prediction Systems. Iowa State University.

Ke, C.-Q., Li, X.-C., Xie, H., Dong-Hui, M., Liu, X., Cheng, K., 2016. Variability in snow cover phenology in China from 1952 to 2010. Hydrol. Earth Syst. Sci. 20, 755.

Kelly, R.E., Chang, A.T., Tsang, L., Foster, J.L., 2003. A prototype AMSR-E global snow area and snow depth algorithm. IEEE Trans. Geosci. Remote Sens. 41, 230–242.

Klein, A.G., Barnett, A.C., 2003. Validation of daily MODIS snow cover maps of the Upper Rio Grande River Basin for the 2000–2001 snow year. Remote Sens. Environ. 86, 162–176.

Klein, A.G., Hall, D.K., Riggs, G.A., 1998. Improving snow cover mapping in forests through the use of a canopy reflectance model. Hydrol. Process. 12, 1723–1744.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 159–174.

Liang, T.G., Huang, X.D., Wu, C.X., Liu, X.Y., Li, W.L., Guo, Z.G., Ren, J.Z., 2008. An application of MODIS data to snow cover monitoring in a pastoral area: a case study in northern Xinjiang, China. Remote Sens. Environ. 112, 1514–1526.

Liu, J., Chen, R., 2011. Studying the spatiotemporal variation of snow-covered days over China based on combined use of MODIS snow-covered days and in situ observations. Theor. Appl. Climatol. 106, 355–363.

Ma, N., Zhang, Y., Guo, Y., Gao, H., Zhang, H., Wang, Y., 2015. Environmental and biophysical controls on the evapotranspiration over the highest alpine steppe. J. Hydrol. 529, 980–992.

Maurer, E.P., Rhoads, J.D., Dubayah, R.O., Lettenmaier, D.P., 2003. Evaluation of the snow-covered area data product from MODIS. Hydrol. Process. 17, 59–71.

Metsämäki, S., 2016. Report on Validation of VIIRS-FSC Products against In-Situ Observations. Finnish Environment Institute.

Musselman, K.N., Clark, M.P., Liu, C., Ikeda, K., Rasmussen, R., 2017. Slower snowmelt in a warmer world. Nat. Clim. Chang. 7, 214.

NSIDC, 2017. MODIS V6 Reprocessing Plan.

Parajka, J., Blöschl, G., 2006. Validation of MODIS snow cover images over Austria. Hydrol. Earth Syst. Sci. Discuss. 3, 1569–1601.

Parajka, J., Blöschl, G., 2012. MODIS-based snow cover products, validation, and hydrologic applications. In: Chang, N.-B., Hong, Y. (Eds.), Multiscale Hydrologic Remote Sensing: Perspectives and Applications.

Powers, D.M., 2011. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation.

Qin, D., Liu, S., Li, P., 2006. Snow cover distribution, variability, and response to climate change in western China. J. Clim. 19, 1820–1833.

Qiu, J., 2008. China: the third pole. Nature News 454, 393–396.

Ramsay, B.H., 1998. The interactive multisensor snow and ice mapping system. Hydrol. Process. 12, 1537–1546.

Riggs, G.A., Hall, D.K., Salomonson, V.V., 2006. MODIS snow products user guide to collection 5. Digital Media 80, 1–80.

Riggs, G.A., Hall, D.K., Román, M.O., 2016. MODIS snow products user guide for Collection 6. https://modis-snow-ice.gsfc.nasa.gov/uploads/C6_MODIS_Snow_User_Guide.pdf.

Riggs, G.A., Hall, D.K., Román, M.O., 2017. Overview of NASA's MODIS and visible infrared imaging radiometer suite (VIIRS) snow-cover earth system data records. Earth Syst. Sci. Data 9, 765.

Rittger, K., Painter, T.H., Dozier, J., 2013. Assessment of methods for mapping snow cover from MODIS. Adv. Water Resour. 51, 367–380.

Romanov, P., Tarpley, D., 2003. Automated monitoring of snow cover over South America using GOES Imager data. Int. J. Remote Sens. 24, 1119–1125.

Salomonson, V., Appel, I., 2004. Estimating fractional snow cover from MODIS using the normalized difference snow index. Remote Sens. Environ. 89, 351–360.

Salomonson, V.V., Appel, I., 2006. Development of the Aqua MODIS NDSI fractional snow cover algorithm and validation results. IEEE Trans. Geosci. Remote Sens. 44, 1747–1756.

Shi, J., Dozier, J., 1997. Mapping seasonal snow with SIR-C/X-SAR in mountainous areas. Remote Sens. Environ. 59, 294–307.

Siderius, C., Biemans, H., Wiltshire, A., Rao, S., Franssen, W., Kumar, P., Gosain, A., van Vliet, M., Collins, D., 2013. Snowmelt contributions to discharge of the Ganges. Sci. Total Environ. 468, S93–S101.

Simpson, J., Stitt, J., Sienko, M., 1998. Improved estimates of the areal extent of snow cover from AVHRR data. J. Hydrol. 204, 1–23.

Tang, B.-H., Shrestha, B., Li, Z.-L., Liu, G., Ouyang, H., Gurung, D.R., Giriraj, A., San Aung, K., 2013a. Determination of snow cover from MODIS data for the Tibetan Plateau region. Int. J. Appl. Earth Obs. Geoinf. 21, 356–365.

Tang, Z.G., Wang, J., Li, H.Y., Yan, L.L., 2013b. Spatiotemporal changes of snow cover over the Tibetan plateau based on cloud-removed moderate resolution imaging spectroradiometer fractional snow cover product from 2001 to 2011. J. Appl. Remote. Sens. 7, 14.

Thirel, G., Salamon, P., Burek, P., Kalas, M., 2013. Assimilation of MODIS snow cover area data in a distributed hydrological model using the particle filter. Remote Sens. 5, 5825–5850.

Wang, X., Xie, H., Liang, T., 2008. Evaluation of MODIS snow cover and cloud mask and its application in Northern Xinjiang, China. Remote Sens. Environ. 112, 1497–1513.

Wang, X.L., Wang, W., Feng, Q.S., Zhi-Bang, L.V., Liang, T.G., 2012. A snow cover mapping algorithm based on MODIS data in Qinghai province. Acta Prataculturae Sinica 4, 293–299.

Wang, X., Pang, G., Yang, M., 2017. Precipitation over the Tibetan Plateau during recent decades: a review based on observations and simulations. Int. J. Climatol. 38 (3), 1116–1131.

Xiao, L., Che, T., Chen, L., Xie, H., Dai, L., 2017. Quantifying snow albedo radiative forcing and its feedback during 2003–2016. Remote Sens. 9, 883.

Xu, W., Ma, L., Ma, M., Zhang, H., Yuan, W., 2017. Spatial–temporal variability of snow cover and depth in the Qinghai–Tibetan Plateau. J. Clim. 30, 1521–1533.

Yang, J., Jiang, L., Ménard, C.B., Luojus, K., Lemmetyinen, J., Pulliainen, J., 2015. Evaluation of snow products over the Tibetan Plateau. Hydrol. Process. 29, 3247–3260.

Yao, T., Thompson, L.G., Mosbrugger, V., Zhang, F., Ma, Y., Luo, T., Xu, B., Yang, X., Joswiak, D.R., Wang, W., 2012. Third pole environment (TPE). Environ. Dev. 3, 52–64.

Yeo, S.-R., Kim, W., Kim, K.-Y., 2017. Eurasian snow cover variability in relation to warming trend and Arctic Oscillation. Clim. Dyn. 48, 499–511.

Yu, J., Zhang, G., Yao, T., Xie, H., Zhang, H., Ke, C., Yao, R., 2016. Developing daily cloud-free snow composite products from MODIS Terra-aqua and IMS for the Tibetan Plateau. IEEE Trans. Geosci. Remote Sens. 54, 2171–2180.

Zhang, Y., Ma, N., 2018. Spatiotemporal variability of snow cover and snow water equivalent in the last three decades over Eurasia. J. Hydrol. 559, 238–251.

Zhang, Y.S., Li, T., Wang, B., 2004. Decadal change of the spring snow depth over the Tibetan Plateau: the associated circulation and influence on the East Asian summer monsoon. J. Clim. 17, 2780–2793.

Zhang, G.Q., Xie, H.J., Yao, T.D., Liang, T.G., Kang, S.C., 2012. Snow cover dynamics of four lake basins over Tibetan Plateau using time series MODIS data (2001—2010). Water Resour. Res. 48.

Zhang, B., Wu, Y.H., Lei, L.P., Li, J.S., Liu, L.L., Chen, D.M., Wang, J.B., 2013a. Monitoring changes of snow cover, lake and vegetation phenology in Nam Co Lake Basin (Tibetan Plateau) using remote SENSING (2000–2009). J. Great Lakes Res. 39, 224–233.

Zhang, L., Su, F., Yang, D., Hao, Z., Tong, K., 2013b. Discharge regime and simulation for the upstream of major rivers over Tibetan Plateau. J. Geophys. Res.-Atmos. 118, 8500–8518.

Zhang, F., Zhang, H., Hagen, S.C., Ye, M., Wang, D., Gui, D., Zeng, C., Tian, L., Liu, J., 2015. Snow cover and runoff modelling in a high mountain catchment with scarce data: effects of temperature and precipitation parameters. Hydrol. Process. 29, 52–65.

Zhang, H., Zhang, F., Ye, M., Che, T., Zhang, G., 2016. Estimating daily air temperatures over the Tibetan Plateau by dynamically integrating MODIS LST data. J. Geophys. Res.-Atmos. 121, 11,425–411,441.

Zhou, H., Aizen, E., Aizen, V., 2013. Deriving long term snow cover extent dataset from AVHRR and MODIS data: Central Asia case study. Remote Sens. Environ. 136, 146–162.