Remote Sensing of Environment 215 (2018) 471–481

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Improvement of the Fmask algorithm for Sentinel-2 images: Separating clouds from bright surfaces based on parallax effects

David Frantz*[,1], Erik Haß[2], Andreas Uhl[3], Johannes Stoffels, Joachim Hill

Environmental Remote Sensing and Geoinformatics, Faculty of Regional and Environmental Sciences, Trier University, 54286 Trier, Germany

ARTICLE INFO

Keywords:
Cloud detection
Fmask
MSI
Parallax
Sentinel-2
View geometry

ABSTRACT

Reliable identification of clouds is necessary for any type of optical remote sensing image analysis, especially in operational and fully automatic setups. One of the most elaborated and widespread algorithms, namely Fmask, was initially developed for the Landsat suite of satellites. Despite their similarity, application to Sentinel-2 imagery is currently hampered by the unavailability of a thermal band, and although results can be improved when taking the cirrus band into account, Sentinel-2 cloud detections are unsatisfactory in two points. (1) Low altitude clouds can be undetectable in the cirrus band, and (2) bright land surfaces – especially built-up structures – are often misclassified as clouds when only considering spectral information. In this paper, we present the Cloud Displacement Index (CDI), which makes use of the three highly correlated near infrared bands that are observed with different view angles. Hence, elevated objects like clouds are observed under a parallax and can be reliably separated from bright ground objects. We compare CDI with the currently used cloud probabilities, and propose how to integrate this new functionality into the Fmask algorithm. We validate the approach using test images over metropolitan areas covering a wide variety of global environments and climates, indicating the successful separation of clouds and built-up structures (overall accuracy 95%, i.e. an improvement in overall accuracy of 0.29–0.39 compared to the previous Fmask versions over the 20 test sites), and hence a full compensation for a missing thermal band.

1. Introduction

Cloud detection is inevitably required for any earth surface-related usage of optical remote sensing imagery like Landsat and Sentinel-2 data. If not accounted for, clouds adversely influence virtually any image analysis like atmospheric correction or land cover classification (Zhu and Woodcock, 2012). Nevertheless, the fully automatic detection of clouds is not trivial, partly due to the high variability in reflectance and temperature of both land surfaces and clouds (Irish, 2000). As such, historically, cloud masks were often generated by hand, which is a very labor- and cost-intensive step, only feasible for few images. With the advent of increasing volumes of freely available satellite data (e.g. the opening of the Landsat archive; Woodcock et al., 2008), more and more automatic and accurate cloud detection codes evolved, which simultaneously paved the way for the automatic generation of higher-level earth observation products (e.g. Flood et al., 2013; Frantz et al., 2016; USGS, 2017) and an entirely new usage of the data for both large area and time series analyses simultaneously (Wulder et al., 2012).

In general, cloud detection codes for Landsat-like imagery can be grouped into mono- and multi-temporal approaches. Multi-temporal approaches are advantageous because they can isolate transient changes superimposed on a more stable background signal. Multi-temporal methods include bi-temporal change detection (e.g. Wang et al., 1999) and time series approaches (e.g. Frantz et al., 2015; Goodwin et al., 2013; Hagolle et al., 2010). While detection accuracies are often improved compared to mono-temporal methods (e.g. Goodwin et al., 2013), their inclusion in most Level 2 production systems is not feasible as these are mono-temporal in nature, thus there is still a pressing need for single-date cloud masks.

Mono-temporal cloud detection in Landsat images was initially performed with the automated cloud cover assessment system (ACCA, Irish, 2000; Irish et al., 2006). While the overall cloud contamination was well estimated, ACCA generally failed to identify the exact locations and boundaries of clouds needed for automatic analysis of the data (Zhu and Woodcock, 2012). As such, a number of other techniques were developed over the years (e.g. Choi and Bindschadler, 2004;

* Corresponding author.
E-mail address: david.frantz@geo.hu-berlin.de (D. Frantz).
[1] Present address: Geomatics Lab, Geography Department, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany.
[2] Present address: Robert Bosch GmbH, Future Systems Industrial Technology - Mobile Machinery Systems (CR/AEI1), Robert-Bosch-Campus 1, 71272 Renningen, Germany.
[3] Present address: Image Processing and Data Distribution Services, GAF AG, 80634 Munich, Germany.

https://doi.org/10.1016/j.rse.2018.04.046
Received 7 April 2017; Received in revised form 12 March 2018; Accepted 25 April 2018
Available online 30 April 2018

Hansen et al., 2008; Vermote et al., 2016). However, the probably most successful and elaborated algorithm for Landsat-like data has been the "Function of mask" algorithm (Fmask, Zhu and Woodcock, 2012), which marked an important game changer for the automatic processing and analysis of medium resolution optical imagery and was hence integrated into several Landsat Level 2 production environments (e.g. Flood et al., 2013; Frantz et al., 2016; USGS, 2017) and fully enabled automatic analysis of many data for a wide range of research questions (Bleyhl et al., 2017; Griffiths et al., 2014; Griffiths et al., 2013; Müller et al., 2016; Schmidt et al., 2016; Schneibel et al., 2017a, b; Senf et al., 2017; Zhu et al., 2012).

The accuracy of the Fmask results are generally good on Landsat data: cloud overall accuracy of 96.41%, cloud producer's accuracy of 92.1%, and cloud user's accuracy of 89.4% (Zhu and Woodcock, 2012). In Landsat images, Fmask is generally able to separate land surfaces from clouds. In a first step, a range of spectral tests are used to generate a potential cloud pixel (PCP) layer, which is anticipated to contain all clouds but also some bright clear-sky pixels – merely built-up objects. These false positives are eliminated by computing a cloud probability from all clear pixels in order to estimate a scene-based threshold. In the original Fmask (Zhu and Woodcock, 2012) – hereby defined as $Fmask_{2012}$ – the cloud probability over land is a combination of a temperature and a variability probability. The temperature probability is very effective because clouds are typically colder than the subjacent land surface. The variability probability combines spectral indices from the visible (VIS), near infrared (NIR) and shortwave infrared (SWIR) because clouds have fairly similar reflectance in this part of the spectrum. Since Sentinel-2 is not equipped with a thermal sensor, the cloud probability becomes the variability probability only, hereby defined as CP_{2012}.

Fmask was recently updated (Zhu et al., 2015) – hereby defined as $Fmask_{2015}$ – to also make use of the new cirrus band carried by Landsat 8, and it was demonstrated that the cirrus band can partially account for a missing thermal band. The new cloud probability over land (CP_{2015}) is defined as the sum of the variability probability and a cirrus probability, and can be readily applied to Sentinel-2 imagery.

However, Sentinel-2's cirrus band @ 1.375 µm is located in a strong water absorption band, which only observes the upper layer of the atmosphere (Hagolle et al., 2010). Consequently, the cirrus band is very helpful to detect high altitude cirrus clouds (Zhu et al., 2015), but low to mid altitude clouds are indistinctive and are susceptible to be removed in the cloud probability routine. In addition, many built-up structures – especially artificial materials – end up in the PCP layer and are inseparable because they appear indistinctive in the variability probability. Artificial materials can be very variable in the spectral range covered by Sentinel-2 and – like clouds – can be bright throughout the complete spectrum. In Fmask, bare soils are removed using a NIR to SWIR ratio as reflectance generally increases from NIR to SWIR. However, this is not necessarily the case for bright artificial materials, which results in many false positives in industrial and residential areas.

Hence, without a thermal band, low altitude clouds are susceptible to be omitted and built-up areas often remain as artifacts in the cloud mask. As Sentinel-2's spectral bands fail to succeed at separating artificial materials from clouds with high accuracy, we consequently propose to tackle this problem with an innovative approach that exploits Sentinel-2's unique sensor configuration. This approach is not solely reliant on spectral properties but specifically incorporates view angle effects.

1.1. Background: the S2A view geometry exploit

Sentinel-2A's Multi Spectral Instrument (MSI) is a push-broom sensor with 13 spectral bands that cover the VIS, NIR and SWIR domains (Drusch et al., 2012). Three of these bands have a spatial resolution of 60 m and are mainly intended for atmospheric characterization. The remaining 10 bands are provided at 10–20 m spatial resolution. Among the spectral domains, the NIR plateau is of special interest as there are three spectrally highly correlated bands available, which partially overlap (bands 7, 8 and 8A with central wavelength at 0.782 µm, 0.835 µm and 0.865 µm, respectively); Table 1 gives the correlation matrix between the NIR bands for a cloud-free acquisition (23 Aug 2016, relative orbit 108); note that the 10 m bands were reduced to 20 m using nearest neighbour resampling. All 15 tiles within the product were analysed in order to cover the complete field-of-view (FOV; upper-left: T31UGS, lower-right: T32UPV). The MSI is characterized by a complex sensor arrangement: for each band, twelve detectors are arranged in a staggered configuration to cover the wide FOV (Drusch et al., 2012); Fig. 1(*a*) displays the viewing vectors (average viewing geometry for each of the 12 detectors) for an across-track scanline; the along-track flight vector and Nadir line are superimposed in black. As a result of the push-broom concept, a parallax exists between odd and even detectors – and although less pronounced, there is also a parallax between bands (Gascon et al., 2017): the viewing vectors of different bands point to different locations on the ground (Fig. 1(*a*)). As this shift is systematic for all stationary objects with known altitude, it can be accounted for during systematic (including flight path adjustment) and geometric correction, as well as in the relative calibration of the focal planes when using a DEM (Gascon et al., 2017). Thus, in Level 1 products, displacement effects are small for objects on the land surface (< 0.3 pixels, Gascon et al., 2017), including mountainous areas. However, non-stationary objects with unknown altitude (like clouds) cannot be corrected this way, hence a displacement is still visible in the final Level 1 products. Table 2 gives the mean (below diagonal) and maximum (above diagonal) of view azimuth differences between the NIR bands. Most strikingly, highly correlated NIR bands 8 and 8A look in different directions (µ: 10.3°, max: 27.5°), whereas bands 8A and 7 are more similar (µ: 1.3°, max: 2.8°); see also Fig. 1(*b*,*c*). For most applications, this sensor design might affect the quantitative analysis of surface reflectance properties. However, we propose to exploit the Sentinel-2 detector arrangement, where three spectrally similar bands are observed under different viewing geometries, for enhanced cloud detection. While objects on the land surface are registered to the same position, objects above the land surface are projected onto slightly different locations in the focal plane – and remain in different positions after systematic and geometric correction.

Table 1
Mean correlation matrix for a cloud-free Sentinel-2 acquisition for the bands on the NIR plateau (23 Aug 2016, West-Germany); individual correlation matrices were computed for all 15 tiles, then averaged.

	λ	7	8[b]	8A
7	0.782		0.948	0.991
8[b]	0.835	0.948		0.949
8A	0.865	0.991	0.949	

[b] 10 m band.

1.2. Objectives

In this study, we propose to

- exploit the Sentinel-2 NIR parallax to separate clouds from artificial surfaces,
- demonstrate the superiority of this approach against the probabilistic approach currently used in Fmask (in absence of a thermal band),
- and propose how to integrate the parallax approach into Fmask.

The next section will outline the data used (Section 2), followed by theoretical considerations on how elevated objects are projected to the

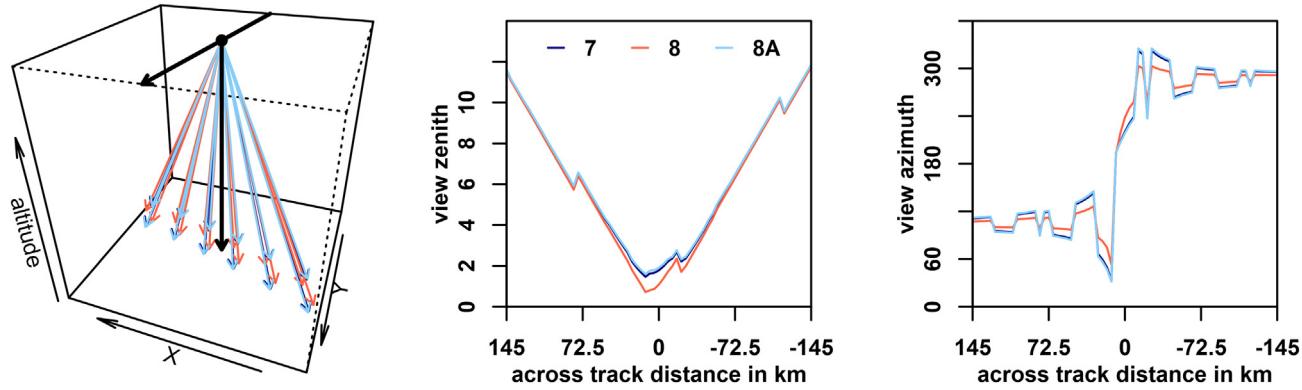

Fig. 1. 3D visualization of viewing vectors (average viewing geometry for each of the 12 detectors) for an across-track scanline for a cloud-free Sentinel-2 acquisition for the bands on the NIR plateau (23 Aug 2016, West-Germany); the along-track flight vector and Nadir line are superimposed in black; view zenith (*b*) and view azimuth (*c*) across satellite track.

Table 2
Mean (below diagonal) and maximum (above) of the view azimuth difference between the NIR bands for a cloud-free Sentinel-2 acquisition (23 Aug 2016, West-Germany) for all 15 tiles covering the complete field-of-view.

		λ	7	8[b]	8A
7	NIR	0.782		25.033	2.799
8[b]	NIR	0.835	9.011		27.543
8A	NIR	0.865	1.321	10.335	

[b] 10 m band.

image focal plane (Section 3.1). Making use of both spectral and angular characteristics, the new index will be presented and compared to existing indices in Section 3.2. Section 3.3 will outline how we integrate the approach into the Fmask toolset, followed by a description of the validation strategy. Results will be presented and discussed in Section 4. The manuscript will close with conclusions.

2. Data

We use 20 globally distributed Sentinel-2 Level 1C acquisitions over several metropolitan areas, where the problem of confusing clouds and built-up structures is most severe. Sites from all continents were selected and systematically cover a wide range of global environments and land surface conditions (Table 3), among them arid, humid, temperate, tropical, cold, mountainous, coastal and continental characteristics. The cloud-contaminated test images were selected based on following criteria: (i) overall cloud contamination between 10% and 50%, (ii) cirrus clouds were avoided since lower altitude clouds are the critical objects, and (iii) both cloud formations and built-up areas should be present. The complete tile was processed, although we will present results only for 100 km^2 subsets. For each of these subsets, cloudy areas were manually classified by a trained expert on the basis of the PCP layer, and thus separated from bright and built-up objects. Pixels that could not be reliably assigned to one of these classes were labelled as uncertain. The relative class proportions of the reference data are shown in Table 3.

3. Methods

3.1. Theoretical basis

In the focal plane, elevated objects are horizontally displaced as a function of object altitude h, view zenith θ and view azimuth Φ. The horizontal distance to the true location, and relative shift in x- and y-direction are given by:

Table 3
Test images.

Location	Code	Tile	Date	Cloud %[b]	Built-up %[b]	Uncer-tain %[b]	Upper left[a]	Lower right[a]
Beijing, China	BJ	T50TMK	10 Sep 2016	51	24	25	5549/7194	7048/7859
Berlin, Germany	BE	T33UUU	02 Sep 2015	90	5	5	8708/7657	10207/8322
Bogotá, Colombia	BO	T18NWL	13 Feb 2017	72	26	2	7706/8967	9205/9632
Cairo, Egypt	CA	T36RUU	13 Feb 2017	65	26	9	1735/7426	3234/8091
Dakar, Senegal	DA	T28PBB	10 Sep 2016	50	40	9	3199/6462	4698/7127
Dhaka, Bangladesh	DH	T45QZG	03 Nov 2016	56	34	9	3622/6193	5121/6858
Jakarta, Indonesia	JK	T48MXU	07 Oct 2016	54	29	17	9170/8119	10669/8784
Johannesburg, RSA	JB	T35JPM	01 Feb 2017	74	17	9	1542/7348	3041/8013
La Paz, Bolivia	LP	T19KEB	21 Oct 2016	67	28	5	8746/1916	10245/2581
Las Vegas, USA	LV	T11SPA	26 May 2016	29	64	7	4778/9236	6277/9901
Mexico City, Mexico	MX	T14QMG	18 Dec 2016	34	59	8	9481/6462	10980/7127
Moscow, Russia	MO	T37UDB	12 Sep 2016	43	56	1	1080/1878	2579/2543
Nairobi, Kenia	NA	T37MBU	21 Oct 2016	72	18	10	5703/4343	7202/5008
New Delhi, India	ND	T43RGM	02 Oct 2016	42	49	9	617/2533	2116/3198
New York, USA	NY	T18TWL	25 Feb 2017	55	25	20	7552/8890	9051/9555
Ottawa, Canada	OT	T18TVR	18 Oct 2016	70	17	13	3661/7079	5160/7744
São Paulo, Brazil	SP	T23KLP	08 Mar 2016	81	12	7	3661/2224	5160/2889
Stockholm, Sweden	ST	T33VXF	23 Feb 2017	40	42	19	5973/1955	7472/2620
Sydney, Australia	SY	T56HLH	26 Nov 2016	76	22	2	2390/4679	3889/5344
Tokyo, Japan	TO	T54SUE	10 Sep 2016	57	28	15	8400/5268	9899/5933

[a] 10 m pixel coordinates, expressed as X/Y.
[b] Percentage of cloud, built-up and uncertain pixels in the potential cloud pixel layer, manually classified.

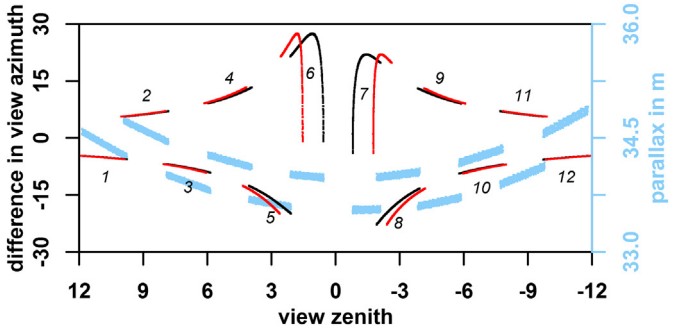

Fig. 2. Difference in view azimuth between bands 8 and 8A, in dependence of view zenith (black: band 8, red: band 8A) for a Sentinel-2 acquisition (23 August 2016, West-Germany) for all 15 tiles covering the complete field-of-view. The numbers indicate the detector arrays. The theoretical parallax in the focal plane for an elevated object (2000 m above surface level) is superimposed in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$d_h = h \bullet \tan\theta, \tag{1}$$

$$d_x = d_h \bullet \sin\Phi, \text{ and} \tag{2}$$

$$d_y = d_h \bullet \cos\Phi. \tag{3}$$

The effective parallax, i.e. the observed shift between the same object observed with different view azimuth angles Φ_1 and Φ_2 and view zenith angles θ_1 and θ_2 can be expressed as:

$$p = \sqrt{(d_{x,1} - d_{x,2})^2 + (d_{y,1} - d_{y,2})^2}. \tag{4}$$

Fig. 2 displays the view azimuth difference between bands 8 and 8A for the data presented in Table 2. Parallaxes (as computed with Eq. (4), assumed cloud height is 2000 m above surface) do exist across the complete FOV, and it is apparent that assumingly lower parallaxes at low view zeniths are counterbalanced by larger view azimuth and zenith differences near the Nadir. In Fmask, cloud heights are assumed to be between 200 m and 12,000 m above surface, for which parallaxes of approx. 3.4 m and 204.7 m exist. Thus misalignments are present for the complete FOV and for all cloud heights, although the uncertainty for the lowest clouds might be higher.

3.2. Cloud and built-up separation

The basic idea to separate clouds and land surfaces is to make use of the spectrally correlated NIR bands, which are additionally affected by a view angle parallax. This parallax results in a displacement of elevated objects in the focal plane. The workflow, as well as the integration into the Fmask algorithm is depicted in Fig. 3.

As the NIR bands 7, 8 and 8A are provided at different spatial resolution, the 10 m band is deconvolved using an approximated Sentinel-2 Point Spread Function (PSF), i.e. a Gaussian lowpass filter. The width of the Gaussian PSF was set to half of the pixel size as suggested by Wang et al. (2016). The PCP separation is done on the 20 m resolution.

For land surfaces, band 8A and band 8 are more similar than band 8A and 7 because the first pair is spectrally overlapping. On the contrary, cloud tops are more similar in bands 8A and 7 because of the high parallax between bands 8A and 8. Hence, we base the PCP separation on two NIR ratios, defined as:

$$R_{8A,8} = B_8/B_{8A}, \text{ and} \tag{5}$$

$$R_{8A,7} = B_7/B_{8A}. \tag{6}$$

An example is shown in Fig. 4 for the Cairo test site; a false color image is depicted in (*e*). The land surface is spatially smooth in $R_{8A,8}$ (*a*), whereas there is more spatial granularity in $R_{8A,7}$ (*b*). On the contrary, clouds appear very flat in $R_{8A,7}$, whereas there is much spatial variability in $R_{8A,8}$, which is due to the parallax effect. It is noted that the spatial variability in $R_{8A,8}$ is not confined to cloud borders, but is also apparent for the center of the clouds, which are highly structured and thus parallax effects can readily be seen. Texture measures are effective to highlight image contrast (e.g. caused by misalignment), thus we convolve each ratio with a focal variance filter of width 7 yielding $V_{8A,8}$ and $V_{8A,7}$. A variance filter was chosen as it is a simple measure of texture, which can be computed in a 1-pass implementation (Pébay, 2008), and thus is cost-effective compared to more advanced texture filters like multi-channel filtering (Jain et al., 1997) or gray level co-occurrence matrices (Haralick et al., 1973). As the internal cloud structuring in $R_{8A,8}$ (see Fig. 4(*a*)) resembles a wave structure with alternating low and high values, a filter width of 7 pixels was found to be optimal to perform spatial textural aggregation such that detectability in both troughs and crests could be ensured. The texture measures of the Cairo test site are depicted in Fig. 4, where high and low image contrast in $V_{8A,8}$ (*c*) are apparent for clouds and land surfaces, respectively. The measure $V_{8A,7}$ (*d*) shows opposite behavior, where clouds typically have much lower spatial contrast than in $V_{8A,8}$. Land surfaces typically have higher or only slightly lower contrast than in $V_{8A,8}$. Please note that the indices as presented in Fig. 4(*a–d,h*) are only for illustration purposes; in practice, the values are only computed for PCPs (not for the complete image), which improves the index behavior at cloud edges and prevents cloud dilation.

Following this, a normalized differenced variance ratio is computed to highlight the opposing nature of $V_{8A,8}$ and $V_{8A,7}$. This new index is hereby denoted as Cloud Displacement Index CDI:

$$CDI = (V_{8A,7} - V_{8A,8})/(V_{8A,7} + V_{8A,8}). \tag{7}$$

Examples of CDI and the Fmask cloud probabilities are shown in Fig. 4 (CP_{2012}, (*f*) CP_{2015} (*g*), CDI (*h*)). Clouds and built-up structures are visibly indistinctive in (*f*), whereas the additional cirrus band partially increases the separability (*g*), although low altitude clouds remain indistinctive. On the contrary, CDI (*h*) assumes very low negative values over clouds – even over low altitude clouds – because the textural and spectral difference in bands 8A and 7 is relatively low compared to the high textural difference caused by the parallax between bands 8A and 8. The land surface assumes higher values because of the higher spectral and textural homogeneity in bands 8A and 8 compared to bands 8A and 7. Fig. 5(*a–c*) depicts density distributions of the land probabilities and CDI, grouped by the reference classification. The land covers cannot be separated with CP_{2012} as both distributions overlap substantially. After adding the cirrus probability (CP_{2015}), the overlap between the distributions is considerably reduced. Nevertheless, many low altitude clouds are still indistinctive (see Fig. 4(*g*)). The corresponding CDI distributions, however, only exhibit a marginal overlap. The CDI makes better use of the available feature space ($-1 - +1$) with opposed skewness. Fig. 6 summarizes the distribution overlap for all test images. The overlap is usually high in CP_{2012} (μ: 55.4%, σ: 19.1%, min: 16.3%, max: 81.0%). In general, separability can be increased when taking the cirrus probability into account (CP_{2015}; μ: 41.8%, σ: 23.2%, min: 3.4%, max: 81.5%), although the improvement can be insignificant (Jakarta, Ottawa), or the separability can even decrease (Moscow). This is a direct result of cloud altitude as separability increases with cloud base altitude. For very low altitude clouds (Moscow), the cirrus band even adds a perturbing element rather than useable information. In all cases, the overlap is smallest in the CDI layer and commonly well below 10% (μ: 4.8%, σ: 3.1%, min: 1.4%, max: 15.6%), which seemingly makes it optimal for this separation task.

3.3. Integration in Fmask

As the Fmask algorithm is already of high quality and has proven its effectiveness for automatic mass processing of Landsat images, we propose to integrate the CDI functionality into the existing toolset by substituting the cloud probability module for Sentinel-2 processing; the

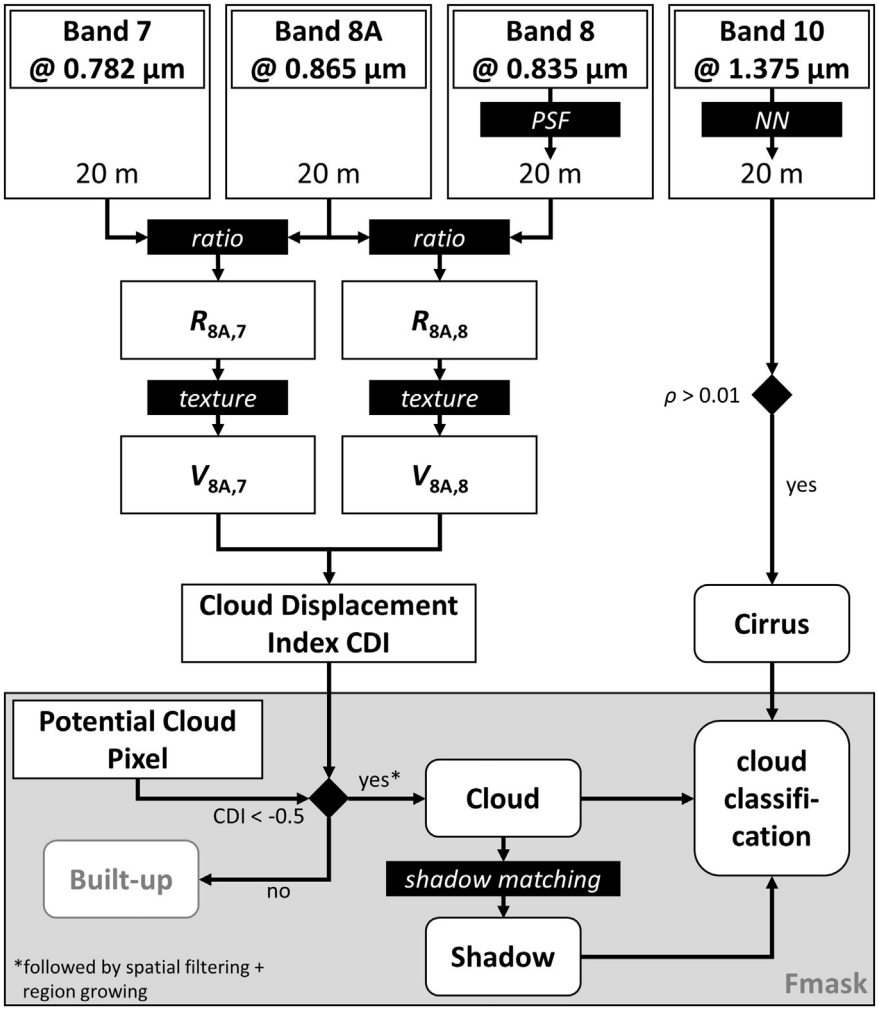

Fig. 3. Workflow of the cloud and built-up pixel separation and its integration into the Fmask algorithm (gray box).

Landsat algorithm remains as is. By doing so, the Fmask PCP layer serves as working mask (see Fig. 3), and thus the proposed method does not add new cloud objects but is rather intended to provide a better separation of PCPs, wherein PCPs with low CDI are considered to be clouds:

$$Cloud = CDI < -0.5 \quad (8)$$

Most clouds are identified considering this simple threshold, and reliably separated from spatially associated bright surface objects. Although the employed NIR bands are highly correlated, spectral differences as well as glint effects can still occur in this spectral domain. The occurrence and strength of directional glint is different in the three NIR bands due to the different viewing directions, as well as due to the larger amount of received energy in the broad band. Similarly to the original Fmask algorithm, these false positives are reduced using spatial filtering. For this purpose we erode the cloud layer by one pixel, which not only removes isolated pixels but removes linear features too. In the case of extremely low altitude clouds, the conservative threshold of 0.5 produces patchy results, i.e. low altitude clouds are only partially detected. Therefore, a subsequent region growing operation is performed on the basis of the remaining cloud pixels: all connected pixels with CDI < -0.25 are added to the cloud layer. This additional region growing was implemented as a compromise when dealing with clouds at different altitude levels. It will somewhat increase the error of commission for clouds in all altitude levels, therefore overall accuracy will likely decrease to a certain degree. However, in case of very low altitude clouds, omission errors will be reduced, which we consider more important from a user perspective.

After removing the non-cloud pixels, processing continues with cloud shadow matching using the modifications described by Frantz et al. (2015, 2016). As the cirrus band does not allow for the identification of low altitude clouds and as the high altitude and thin cirrus clouds produce susceptible results when used in the shadow matching routine, we do not use the cirrus test in the PCP selection as proposed by Zhu et al. (2015). Instead, we incorporated this robust test in the generation of the final cloud classification:

$$Cirrus = B_{10} > 0.01 \text{ AND NOT } Cloud \text{ AND NOT } Shadow. \quad (9)$$

3.4. Validation

Besides visual evaluation, we perform a quantitative assessment on the basis of the manually classified PCP images described in the data section. The new parallax-based approach, as well as both Fmask versions were validated. Uncertain pixels were ignored. For each test image, a confusion matrix was generated and overall accuracy (OA), producer's accuracy (PA), user's accuracy (UA) measures were computed using the following equations; P and N are real positives (cloud) and negatives (other), TP and TN are true positives and negatives, FP and FN are false positives and negatives. Additionally, overall values were computed.

$$OA = \frac{TP + TN}{P + N} \quad (10)$$

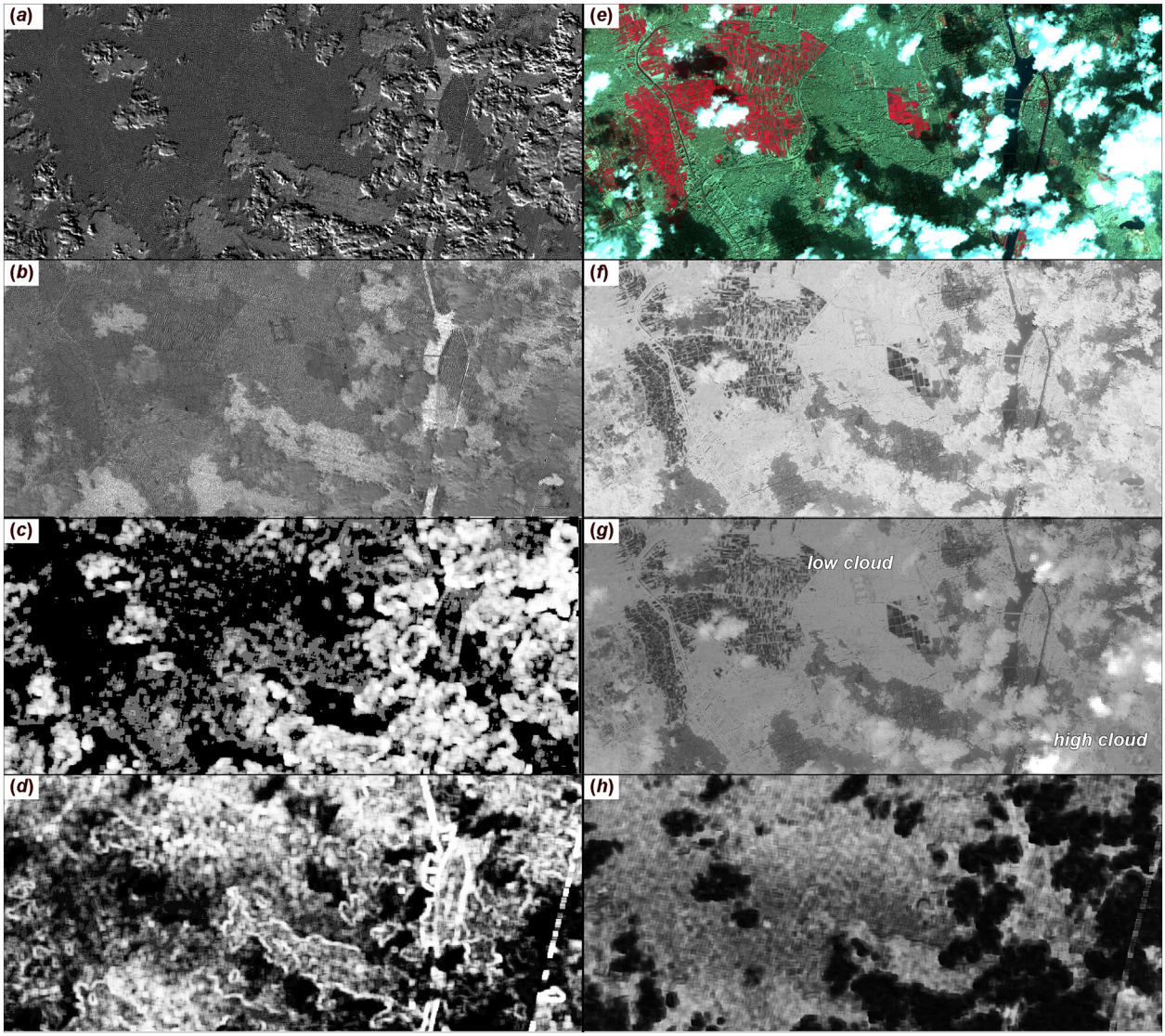

Fig. 4. Cairo test image; (*a*) $R_{8A,8}$, linear stretch from 0.8–1.2; (*b*) $R_{8A,7}$, linear stretch from 0.8–1.2; (*c*) $V_{8A,8}$, histogram equalization stretch; (*d*) $V_{8A,7}$, histogram equalization stretch; (*e*): false color image (bands NIR/red/green); (*f*) Land Probability 2012, linear stretch from 0 to 1; (*g*) Land Probability 2015, linear stretch from 0 to 1.5; (*h*) Cloud Displacement Index, linear stretch from −1–1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

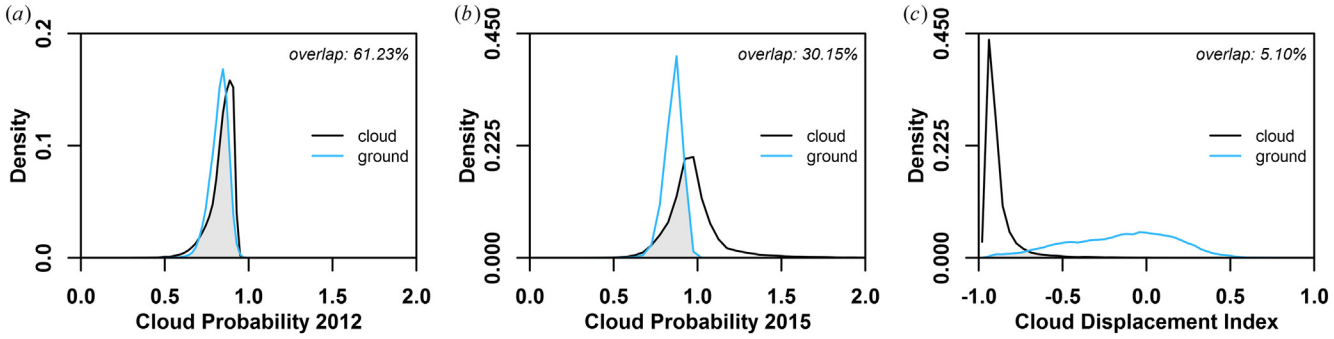

Fig. 5. Density distribution for the Cairo test image based on manually classified PCPs. The Land Probability 2012 (*a*), Land Probability 2015 (*b*) and Cloud Displacement Index (*c*) correspond to the images shown in Fig. 4. The density was estimated on the basis of the histogram counts, and satisfies the sum-to-one condition.

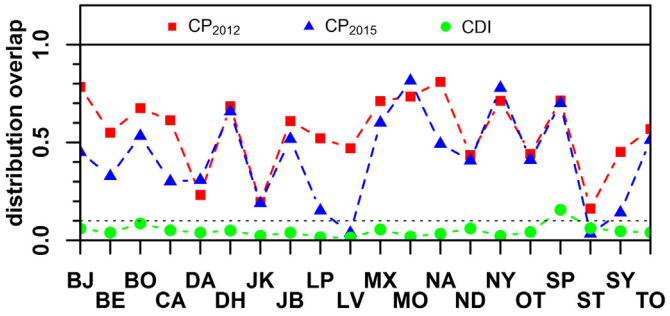

Fig. 6. Density distribution overlap between cloud and built-up pixels for all test images as exemplarily illustrated in Fig. 5. The different signatures refer to the original Fmask's cloud probability (CP_{2012}), updated Fmask's cloud probability (CP_{2015}), and the newly developed Cloud Displacement Index (CDI).

$$PA = \frac{TP}{TP + FN} \tag{11}$$

$$UA = \frac{TP}{TP + FP} \tag{12}$$

4. Results and discussion

Separated PCPs are shown for all test images in Fig. 7 (1st column: false color composites, 2nd–4th column: cloud classification from the $Fmask_{2012}$, $Fmask_{2015}$ and newly proposed method, respectively). Clouds are shown in white, rejected PCPs in blue. The results of the original Fmask version (2nd column) are strongly mixed. Cloud separation can work well (e.g. Berlin), but can also result in significant omission of clouds (e.g. Las Vegas) or a large share of commission errors (e.g. Moscow). As the cloud probability equals the variability probability, the results are a function of the spectral variability of both the land surface and clouds. The variability probability combines spectral indices from the visible (VIS), near infrared (NIR) and shortwave infrared (SWIR) because clouds have fairly similar reflectance in this part of the spectrum. However, this can also apply to very bright surfaces (especially artificial materials), and as the upper range (82.5 percentile) of the non-PCPs is used to derive the cloud threshold, chances are that a threshold is selected that is in the value range of the built-up land surface, which in turn can overlap with the cloud probabilities of the clouds (compare with Fig. 6). As such, it is both possible that clouds are removed from the PCPs or that land surfaces remain in the cloud mask. In general, results are worse towards the edge of clouds, as the cloud probabilities are lower at the edge of the cloud and are thus more susceptible to fall under the selection threshold. In most cases, the results improve with the updated Fmask version (3rd column), which also takes into account the cirrus probability. In this case, the cloud probability is the sum of the variability probability and the cirrus probability. Thus, cloud probabilities increase with cloud altitude and results can be significantly improved if clouds are at higher altitudes (e.g. Las Vegas). On the contrary, the cirrus probability does not add information for low altitude clouds, hence the cloud masks cannot be improved when compared to the original Fmask version. On the contrary, the newly proposed method is able to reliably separate clouds from other surfaces. Cloud shapes are clearly defined and there are distinct borders between clouds and built-up objects, which even allowed for reliable cloud detection over urban areas. This is because the cloud and built-up separation is not solely performed in the spectral feature space but primarily exploits Sentinel-2's unique observation geometry. Even very low altitude cloud formations remain in the cloud layer (e.g. Moscow or São Paulo – compare with Fig. 6).

The quantitative quality assessment is summarized in Table 4. Overall, all accuracy measures increase from $Fmask_{2012}$ to $Fmask_{2015}$ to $Fmask_{CDI}$ (see last row in Table 4). Overall Accuracies highly resemble the pattern presented in Fig. 6, indicating that the distribution overlap of the CP_{2012}, CP_{2015} and CDI are the main reasons for algorithm performance. In all test sites, results of $Fmask_{2012}$ were improved when using the cirrus band ($Fmask_{2015}$) – or did not change substantially. However, for most test sites, the overall accuracies are still fairly low (on average OA = 0.66) due to the reasons discussed in the last paragraph, i.e. results can only be significantly improved for higher altitude clouds (e.g. Las Vegas). The range of Producer's Accuracies are very large for $Fmask_{2012}$ and $Fmask_{2015}$, which indicates that the algorithm can detect most clouds in some images (e.g. Berlin), but can also discard most PCPs in other circumstances (e.g. New York). In general, User's Accuracies are higher than Producer's Accuracies in the case of $Fmask_{2012}$ and $Fmask_{2015}$, which indicates that the algorithms are less susceptible of detecting false positives than introducing false negatives. The newly proposed $Fmask_{CDI}$ yield the highest accuracies (on average, OA = 0.95; see last row in Table 4). The Producer's Accuracy is highest in all test sites (on average, PA = 0.99), thus the method is very sensitive and rarely misses existing clouds, i.e. omission errors are generally low. As an exception, some cloud pixels were missed in São Paulo, which are mainly patches within the correctly identified cloud formations that were not characterized by a measurable parallax (recall that the parallax is only 3.4 m for cloud heights of 200 m above surface level). Nevertheless, we anticipate this to be uncritical in practice as clouds are generally buffered before analysis or the distance to the cloud taken into account (e.g. Frantz et al., 2017; Griffiths et al., 2013). User's Accuracy is also very high (on average, UA = 0.93), which indicates that false positives are rare, although they occur in specific sites. Most strikingly, some commission effects are apparent over Stockholm, where some confusion between snow and clouds occurred. However, it needs to be noted that the manual classification might also be imperfect to a certain degree as it is sometimes challenging to visually separate clouds from other bright surfaces, especially over snow, when thin clouds are present or at the edge of clouds. In general, the main reason for false positives is the implemented region growing functionality, which attempts to classify less confident, but attached-to confident pixels as cloud. In the extreme case of very densely packed buildings, and the presence of a slight across-track parallax, this approach can cause chaining effects, as e.g. in the Cairo site (other examples can be found throughout the other sites, too). If this feature is unwanted, region growing could be disabled, which would increase User's Accuracy and Overall Accuracy for most sites. Nevertheless, in extreme cases (e.g. the very low altitude clouds in São Paulo), this would be at the expense of missing existing clouds, thus we anticipated that users would rather prefer an utmost sensitive test with fairly high specificity.

The proposed method does not add new cloud objects, but is rather intended to better separate PCPs. In general, missed clouds are not problematic at the PCP stage. As can be seen in Fig. 7 (at least visually), it is rare that clouds are not included in the PCP layer. As such, we refrained from evaluating actual clouds that were not included in the PCP layer in order to enable a more efficient manual cloud screening process, and thus the manual generation of more validation data with a limited budget. The results of $Fmask_{2012}$ and $Fmask_{2015}$ (Fig. 7) show that omitted clouds are rather removed during the second stage of Fmask, i.e. in the cloud probability screening, where especially low altitude clouds and cloud edges are removed from the cloud mask. On the contrary, the new approach is characterized by a very high producer's accuracy of 0.99, thus almost all cloud PCPs remain in the cloud layer.

The presented work mainly improves the performance for low–mid altitude clouds. High altitude cirrus clouds are detected using a simple cirrus band threshold. It is noted that this procedure may introduce false positives in high elevation areas, especially when snow covered. In addition, omission errors may be present for cirrus clouds that do not exceed this threshold. In this regard, it may be advantageous to combine the Cloud Displacement Index and the probabilistic cloud detection of the original Fmask in a future version of either algorithm.

While this work's focus is on the improvement of cloud detection in Sentinel-2 images, it is recognized that a good cloud shadow mask is of equal importance. However, cloud shadow detection capabilities of Fmask, especially with modified cloud shadow matching as described in Frantz et al. (2015, 2016), are already of high quality – especially when considering the more complex problem when compared to cloud detection. Nevertheless, and although not demonstrated here, we anticipate that the cloud shadow masks of Fmask are also improved with the newly proposed method. During shadow matching, each false positive cloud object is matched with local depressions in reflectance, thus many false positive cloud shadows follow naturally from false positive clouds. Thus, an improved cloud mask with fewer false positives

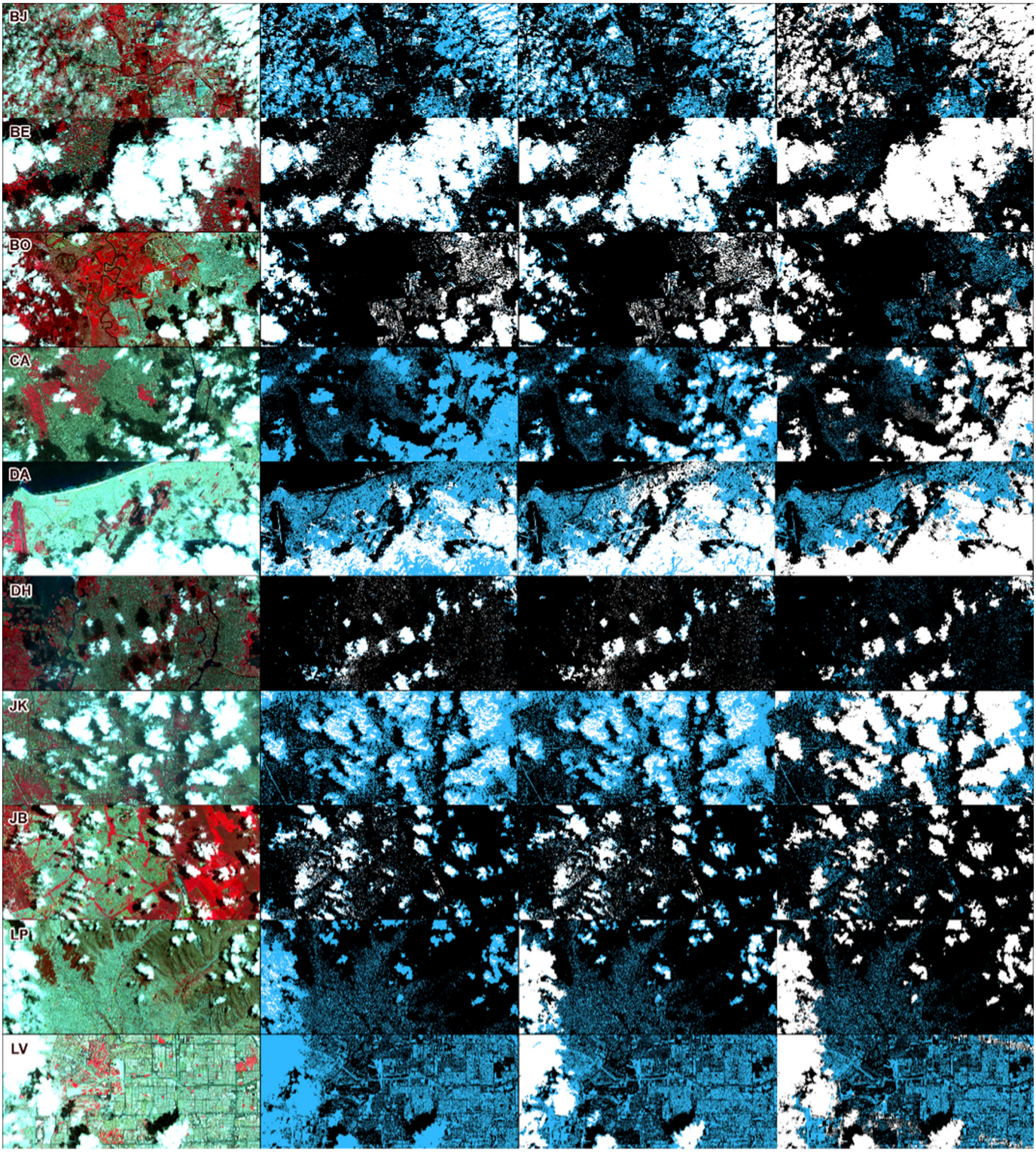

Fig. 7. Cloud and built-up separation for all test images. False color images (bands 8/4/3, same stretch; 1st column) are shown next to the separation results (2nd column: $Fmask_{2012}$; 3rd column: $Fmask_{2015}$; 4th column: $Fmask_{CDI}$). Blue colors are PCPs that were assigned to the built-up class, white pixels are PCPs that remain clouds. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

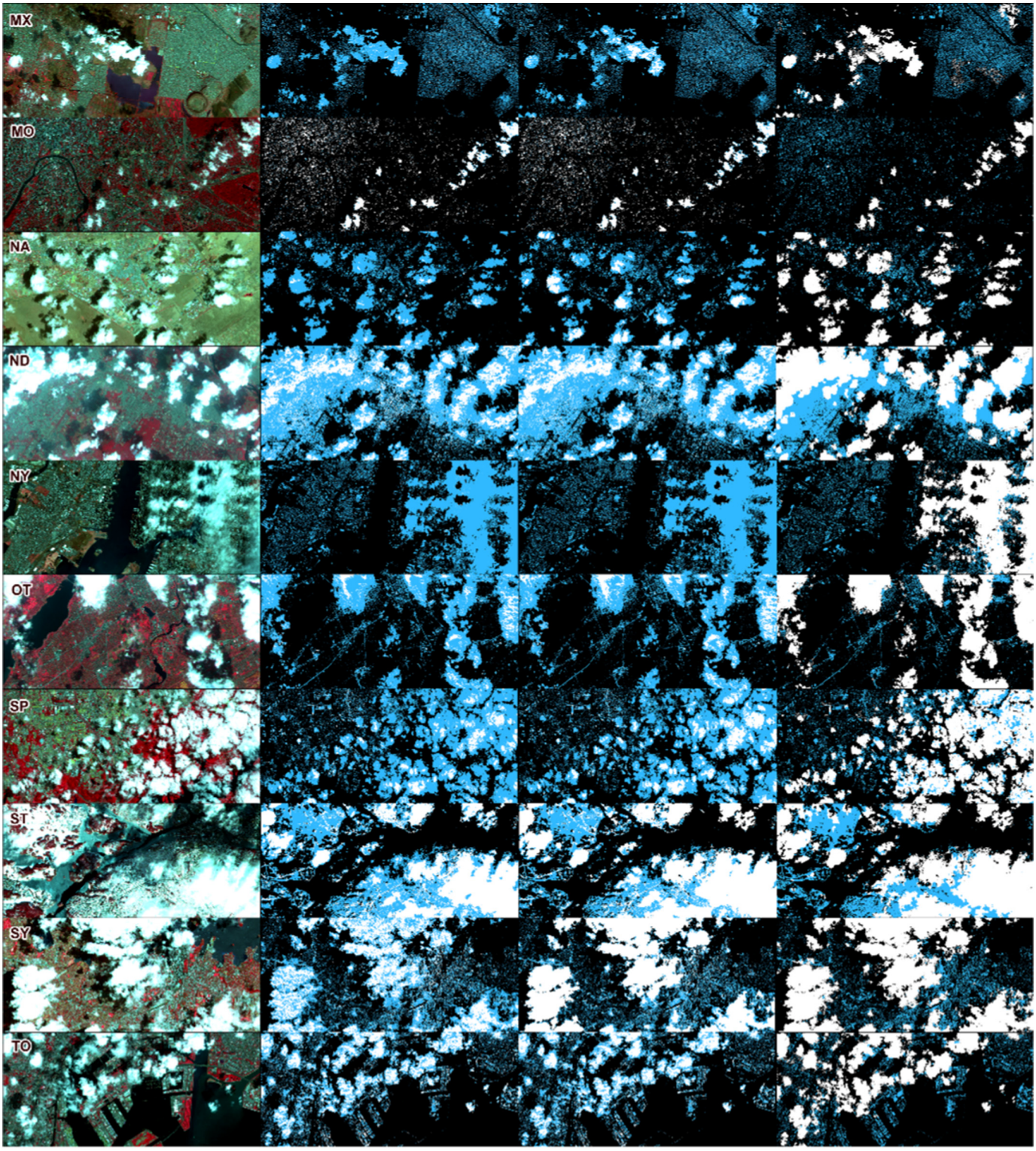

Fig. 7. (*continued*)

naturally results in an improved cloud shadow mask. However, a quantitative evaluation of this chain of thought would be highly beneficial in the future.

While the effective parallax between NIR bands has proven effective for improved cloud detection, it is also imaginable that this effect can be leveraged for other problems, too. As an example, cloud shadow matching could be improved if the cloud altitude would be known. In theory, it could be estimated from the parallax with the equations given in Section 3.1. However, this is not trivial in a number of respects. First of all, the exact viewing vectors would be needed. However, the viewing geometry for Sentinel-2 is only given for 5 km grid cells, but in such a cell, two sensor arrays with quite different view angles can be present. Thus, the exact border between the detectors would need to be estimated or the viewing geometry would need to be given at full resolution, which would tremendously increase Level 1 data volume. In addition, the horizontal displacement would need to be measured,

Table 4
Accuracy assessment.

Fmask	Overall accuracy			Producer's accuracy			User's accuracy		
	2012	2015	CDI	2012	2015	CDI	2012	2015	CDI
BJ	0.466	0.681	0.932	0.330	0.639	0.997	0.738	0.855	0.911
BE	0.862	0.887	0.985	0.892	0.918	0.999	0.959	0.960	0.986
BO	0.674	0.678	0.945	0.859	0.865	0.972	0.739	0.740	0.953
CA	0.289	0.614	0.921	0.008	0.461	0.995	0.991	1.000	0.904
DA	0.792	0.787	0.903	0.713	0.885	0.986	0.890	0.768	0.860
DH	0.679	0.684	0.947	0.817	0.878	0.992	0.710	0.695	0.928
JK	0.658	0.653	0.960	0.473	0.465	0.998	0.997	0.999	0.943
JB	0.613	0.654	0.976	0.621	0.684	0.984	0.863	0.860	0.987
LP	0.375	0.784	0.988	0.111	0.694	0.994	0.998	0.999	0.989
LV	0.684	0.972	0.942	0.000	0.913	0.998	0.903	1.000	0.847
MX	0.645	0.726	0.939	0.054	0.267	0.985	0.671	0.935	0.865
MO	0.502	0.512	0.983	0.872	0.859	0.994	0.464	0.469	0.967
NA	0.232	0.408	0.983	0.050	0.267	0.991	0.883	0.981	0.988
ND	0.735	0.758	0.912	0.554	0.574	0.989	0.811	0.853	0.846
NY	0.310	0.310	0.986	0.003	0.003	0.993	0.661	0.639	0.987
OT	0.333	0.396	0.977	0.183	0.259	0.997	0.959	0.976	0.976
SP	0.246	0.314	0.877	0.147	0.226	0.868	0.932	0.953	0.990
ST	0.845	0.850	0.871	0.916	0.994	0.996	0.796	0.768	0.793
SY	0.631	0.879	0.965	0.564	0.882	0.992	0.936	0.959	0.964
TO	0.662	0.683	0.964	0.609	0.620	0.992	0.847	0.874	0.956
Σ	0.562	0.662	0.948	0.439	0.618	0.986	0.837	0.864	0.932

which would necessitate image-to-image matching and which would have a huge impact on processing time and algorithmic complexity. Please note, in our proposed cloud detection method, the actual view angles and the actual horizontal displacement are not needed since we compute a pixel-based index, which is based on the effect of the parallax.

The proposed method relies on the availability of several highly correlated, spatially highly resolved bands in the same spectral domain, and an observation geometry, which produces parallax effects for objects located above the land surface. In the case of Sentinel-2, the NIR domain is optimal to accomplish this task as it was densely populated with spectral bands for the computation of enhanced vegetation related indicators, which had to be arranged in a staggered sensor configuration. We presume that the approach can be directly transferred to the upcoming Sentinel-2B–D sensors as these are anticipated to be identically constructed; first tests with recent Sentinel-2B data have confirmed this. While the approach is not suited for whiskbroom sensors like the Landsat 7 ETM+, it may be possible to apply it to the Landsat 8 OLI pushbroom sensor - either as a supplement to the existing algorithm, or as a potential replacement for the thermal band in the event of a future sensor failure. Regarding the spectral configuration, the panchromatic versus visual bands may be utilized as they are spectrally overlapping, and first tests revealed that there is a slight parallax, too. However, additional research needs to be carried out in order to verify if the Cloud Displacement Index will work in another spectral domain, and if the observed parallax is sufficiently large to supplement the cloud detection capabilities of the Fmask Landsat algorithm.

5. Conclusion

Cloud identification in Sentinel-2 imagery has been hampered by the unavailability of thermal information, often used to reliably separate clouds from other bright objects like built-up structures, snow covered areas or bright bare soils. Cloud masking can be improved when taking the cirrus band into account, however, missed low altitude clouds and false positive bright land surface remain problematic. Consequently, we developed a new index – the Cloud Displacement Index (CDI) – which not solely relies on spectral information but additionally takes into account effects introduced by the band-specific Sentinel-2 observation geometries. The technique relies on the parallax between near infrared bands that is apparent in Sentinel-2 imagery across the complete Field-of-View. This parallax results in a horizontal displacement of objects located above the land surface, while built-up structures are registered to identical image positions. The CDI was successfully tested on 20 globally distributed Sentinel-2A Level 1C datasets focusing on metropolitan areas – regions where the problems of the existing Sentinel-2 cloud masking algorithms are most severe – and considering a wide range of global environments and its typical land cover, including bright bare soils and snow. Using the CDI, clouds can be reliably separated from other objects, and even low altitude clouds – which cannot be identified in the cirrus band – were successfully retained, even over bright land covers like built-up, snow or soil. The method was demonstrated to have substantially fewer false negatives and fewer false positives, too. Compared to the existing Fmask versions, a substantial performance boost was apparent in all accuracy measures, yielding an overall accuracy of 0.95, producer's accuracy of 0.99, and user's accuracy of 0.93. Hence, this parallax-based approach fully compensates for a missing thermal band, and outperforms the addition of a cirrus band. The validation and demonstration was performed using Sentinel-2A imagery, however, recent tests have revealed that it can be equally applied to the newly available Sentinel-2B data as they share the same sensor design. The method is readily integrated into our implementation of the Fmask algorithm, which is incorporated into the FORCE Level 2 Processing System (Framework for Operational Radiometric Correction for Environmental monitoring; available from https://www.uni-trier.de/index.php?id=63673 under the terms of the GNU General Public License v. ≥3), and does not use any ancillary data. Fully automatic preprocessing frameworks that generate higher-level satellite products, as well as the subsequent thematic analyses of these data, will greatly benefit from these improved cloud masks.

Acknowledgements

Sentinel-2A data courtesy of ESA. This work was supported by the Federal Ministry of Transport and Digital Infrastructure and National Aeronautics and Space Research Centre of the Federal Republic of Germany under contract number FKZ 50EW1605 as part of the Sentinel4GRIPS project. The authors would also like to thank the three anonymous reviewers whose very constructive comments significantly improved the quality of the final paper.

References

Bleyhl, B., Baumann, M., Griffiths, P., Heidelberg, A., Manvelyan, K., Radeloff, V.C., Zazanashvili, N., Kuemmerle, T., 2017. Assessing landscape connectivity for large mammals in the Caucasus using Landsat 8 seasonal image composites. Remote Sens. Environ. 193, 193–203.

Choi, H., Bindschadler, R., 2004. Cloud detection in Landsat imagery of ice sheets using shadow matching technique and automatic normalized difference snow index threshold value decision. Remote Sens. Environ. 91, 237–242.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, 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. 120, 25–36.

Flood, N., Danaher, T., Gill, T.K., Gillingham, S.S., 2013. An operational scheme for deriving standardised surface reflectance from Landsat TM/ETM+ and SPOT HRG imagery for eastern Australia. Remote Sens. 5, 83–109.

Frantz, D., Röder, A., Udelhoven, T., Schmidt, M., 2015. Enhancing the detectability of clouds and their shadows in multitemporal Dryland Landsat imagery: extending Fmask. IEEE Geosci. Remote Sens. Lett. 12, 1242–1246.

Frantz, D., Röder, A., Stellmes, M., Hill, J., 2016. An operational radiometric Landsat preprocessing framework for large-area time series applications. IEEE Trans. Geosci. Remote Sens. 54, 3928–3943.

Frantz, D., Röder, A., Stellmes, M., Hill, J., 2017. Phenology-adaptive pixel-based compositing using optical earth observation imagery. Remote Sens. Environ. 190, 331–347.

Gascon, F., Bouzinac, C., Thépaut, O., Jung, M., Francesconi, B., Louis, J., Lonjou, V., Lafrance, B., Massera, S., Gaudel-Vacaresse, A., Languille, F., Alhammoud, B., Viallefont, F., Pflug, B., Bieniarz, J., Clerc, S., Pessiot, L., Trémas, T., Cadau, E., De Bonis, R., Isola, C., Martimort, P., Fernandez, V., 2017. Copernicus Sentinel-2A calibration and products validation status. Remote Sens. 9, 584.

Goodwin, N.R., Collett, L.J., Denham, R.J., Flood, N., Tindall, D., 2013. Cloud and cloud shadow screening across Queensland, Australia: an automated method for Landsat TM/ETM+ time series. Remote Sens. Environ. 134, 50–65.

Griffiths, P., van der Linden, S., Kuemmerle, T., Hostert, P., 2013. A pixel-based Landsat compositing algorithm for large area land cover mapping. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6, 2088–2101.
Griffiths, P., Kuemmerle, T., Baumann, M., Radeloff, V.C., Abrudan, I.V., Lieskovsky, J., Munteanu, C., Ostapowicz, K., Hostert, P., 2014. Forest disturbances, forest recovery, and changes in forest types across the Carpathian ecoregion from 1985 to 2010 based on Landsat image composites. Remote Sens. Environ. 151, 72–88.
Hagolle, O., Huc, M., Pascual, D.V., Dedieu, G., 2010. A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images. Remote Sens. Environ. 114, 1747–1755.
Hansen, M.C., Roy, D.P., Lindquist, E., Adusei, B., Justice, C.O., Altstatt, A., 2008. A method for integrating MODIS and Landsat data for systematic monitoring of forest cover and change in the Congo Basin. Remote Sens. Environ. 112, 2495–2513.
Haralick, R.M., Shanmugam, K., Dinstein, I., 1973. Textural features for image classification. IEEE Trans. Syst. Man Cybern. SMC-3, 610–621.
Irish, R.R., 2000. Landsat 7 automatic cloud cover assessment. In: SPIE Proceedings 4049 - Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery VI, pp. 348–355 (Orlando, FL, USA).
Irish, R.R., Barker, J.L., Goward, S.N., Arvidson, T., 2006. Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm. Photogramm. Eng. Remote. Sens. 72, 1179–1188.
Jain, A.K., Ratha, N.K., Lakshmanan, S., 1997. Object detection using gabor filters. Pattern Recogn. 30, 295–309.
Müller, H., Griffiths, P., Hostert, P., 2016. Long-term deforestation dynamics in the Brazilian Amazon—uncovering historic frontier development along the Cuiabá–Santarém highway. Int. J. Appl. Earth Obs. Geoinf. 44, 61–69.
Pébay, P.P., 2008. Formulas for Robust, One-pass Parallel Computation of Covariances and Arbitrary-order Statistical Moments. Sandia National Laboratories.
Schmidt, M., Pringle, M., Devadas, R., Denham, R., Tindall, D., 2016. A framework for large-area mapping of past and present cropping activity using seasonal Landsat images and time series metrics. Remote Sens. 8, 312.
Schneibel, A., Frantz, D., Röder, A., Stellmes, M., Fischer, K., Hill, J., 2017a. Using annual Landsat time series for the detection of dry Forest degradation processes in south-Central Angola. Remote Sens. 9, 905.
Schneibel, A., Stellmes, M., Röder, A., Frantz, D., Kowalski, B., Haß, E., Hill, J., 2017b. Assessment of spatio-temporal changes of smallholder cultivation patterns in the Angolan Miombo belt using segmentation of Landsat time series. Remote Sens. Environ. 195, 118–129.
Senf, C., Pflugmacher, D., Heurich, M., Krueger, T., 2017. A Bayesian hierarchical model for estimating spatial and temporal variation in vegetation phenology from Landsat time series. Remote Sens. Environ. 194, 155–160.
USGS, 2017. Landsat Surface Reflectance Level-2 Science Data Products. https://landsat.usgs.gov/landsat-surface-reflectance-data-products (In.).
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. 185, 46–56.
Wang, B., Ono, A., Muramatsu, K., Fujiwara, N., 1999. Automated detection and removal of clouds and their shadows from Landsat TM images. IEICE Trans. Inf. Syst. 82, 453–460.
Wang, Q., Shi, W., Li, Z., Atkinson, P.M., 2016. Fusion of Sentinel-2 images. Remote Sens. Environ. 187, 241–252.
Woodcock, C.E., Allen, R., Anderson, M., Belward, A., Bindschadler, R., Cohen, W., Gao, F., Goward, S.N., Helder, D., Helmer, E., 2008. Free access to Landsat imagery. Science 320, 1011a.
Wulder, M.A., Masek, J.G., Cohen, W.B., Loveland, T.R., Woodcock, C.E., 2012. Opening the archive: how free data has enabled the science and monitoring promise of Landsat. Remote Sens. Environ. 122, 2–10.
Zhu, Z., Woodcock, C.E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. 118, 83–94.
Zhu, Z., Woodcock, C.E., Olofsson, P., 2012. Continuous monitoring of Forest disturbance using all available Landsat imagery. Remote Sens. Environ. 122, 75–91.
Zhu, Z., Wang, S., Woodcock, C.E., 2015. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. Remote Sens. Environ. 159, 269–277.