Science of Remote Sensing 2 (2020) 100010

Contents lists available at ScienceDirect

Science of Remote Sensing

journal homepage: www.journals.elsevier.com/science-of-remote-sensing

Full Length Article

Comparison of cloud detection algorithms for Sentinel-2 imagery

Katelyn Tarrio [a,*], Xiaojing Tang [a], Jeffrey G. Masek [b], Martin Claverie [c], Junchang Ju [b], Shi Qiu [d], Zhe Zhu [d], Curtis E. Woodcock [a]

[a] Department of Earth & Environment, Boston University, 685 Commonwealth Avenue, Boston, MA, 02215, USA
[b] Biospheric Sciences Laboratory, NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD, 20771, USA
[c] Earth and Life Institute, Universitécatholique de Louvain, Croix Du Sud 2, 1348, Louvain-la-Neuve, Belgium
[d] Department of Natural Resources and the Environment, University of Connecticut, 1376 Storrs Road, Storrs, CT, 06268, USA

ARTICLE INFO

Keywords:
Sentinel-2
Cloud
Cloud shadow
Fmask
Tmask
MAJA
Sen2Cor
LaSRC
Cloud detection
Cloud mask

ABSTRACT

Accurate, automated cloud and cloud shadow detection is a key component of the processing needed to prepare optical satellite imagery for scientific analysis. Many existing cloud detection algorithms rely on temperature information to identify clouds, making detection difficult for imagers that lack a thermal band, like Sentinel-2. To get maximum benefit from Sentinel-2 products it is critical to understand which algorithms best identify clouds and their shadows in images. We examined the relative performance of five different cloud-masking algorithms (Sen2Cor, MAJA, LaSRC, Fmask and Tmask) in 6 Sentinel-2 scenes (28 total images) distributed across the Eastern Hemisphere. Expanding on these comparisons, we tested ensemble approaches to improve results. We tested three ensemble approaches to cloud and shadow classification based on the outputs of the five initial algorithms using the cloud masks in: (1) a majority prediction model; (2) a random forests model; and (3) a conditional logic model. Accuracy assessments show a trade-off between omission and commission errors in cloud detection for individual algorithms across all sites, and some algorithms are better at detecting either clouds or cloud shadows. No single algorithm outperforms the others for both clouds and shadows. Aggregating the results from multiple algorithms produces fewer undetected clouds and higher overall accuracy than any single algorithm, with as high as 2.7% improvement over the top-performing algorithm, suggesting an ensemble approach may be the most useful for processing of Sentinel-2 data.

1. Introduction

Clouds and cloud shadows are an unavoidable issue in the acquisition process of optical imagery. Clouds significantly alter the spectral signatures obtained from satellite data, which often leads to the false identification of land cover change (Irish et al., 2006). This situation is problematic for remote sensing analyses of surface properties like cover classification (Zhang et al., 2002), image compositing (White et al., 2014) and change detection (Zhu and Woodcock, 2014a). Issues arising from cloud contaminated data are magnified in studies that use large numbers of images. Since the Landsat archive opened in 2008, granting free access to historical imagery (Woodcock et al., 2008), time series analyses have proliferated across a variety of applications, including land cover and land use classification (Franklin et al., 2015; Zhu and Woodcock, 2014b), disturbance monitoring (Huang et al., 2010; Kennedy et al., 2007; Zhu et al., 2020) and surface water mapping (Pekel et al., 2016; Tulbure and Broich, 2013). Multi-temporal studies often employ all useable images, a process which relies on first identifying cloud and cloud shadow-free observations. Higher-level image-processing analyses like change detection require an efficient, accurate and automated approach for removing observations affected by clouds and shadows.

Though accurate cloud detection in images is difficult, the physical principles underlying most methodologies are well established. The majority of cloud detection algorithms are based on physical rules, making use of spectral indices to separate clear-sky pixels from clouds (Huang et al., 2010; Irish et al., 2006; Zhu and Woodcock, 2014a, 2012). In the optical portion of the spectrum, clouds are distinguished from clear pixels by their high reflectance across almost all wavelengths, giving them their white and bright appearance (Irish, 2000; Zhu et al., 2019). The thermal band is helpful for identifying clouds for the basic reason that clouds are colder than clear pixels; in addition, many cloud detection algorithms use thermal information to calculate cloud height and by

* Corresponding author.
E-mail addresses: ktarrio@bu.edu (K. Tarrio), xjtang@bu.edu (X. Tang), jeffrey.g.masek@nasa.gov (J.G. Masek), martin.claverie@uclouvain.be (M. Claverie), junchang.ju@nasa.gov (J. Ju), shi.qiu@uconn.edu (S. Qiu), zhe@uconn.edu (Z. Zhu), curtis@bu.edu (C.E. Woodcock).

https://doi.org/10.1016/j.srs.2020.100010
Received 28 April 2020; Received in revised form 11 September 2020; Accepted 17 September 2020
Available online 3 October 2020

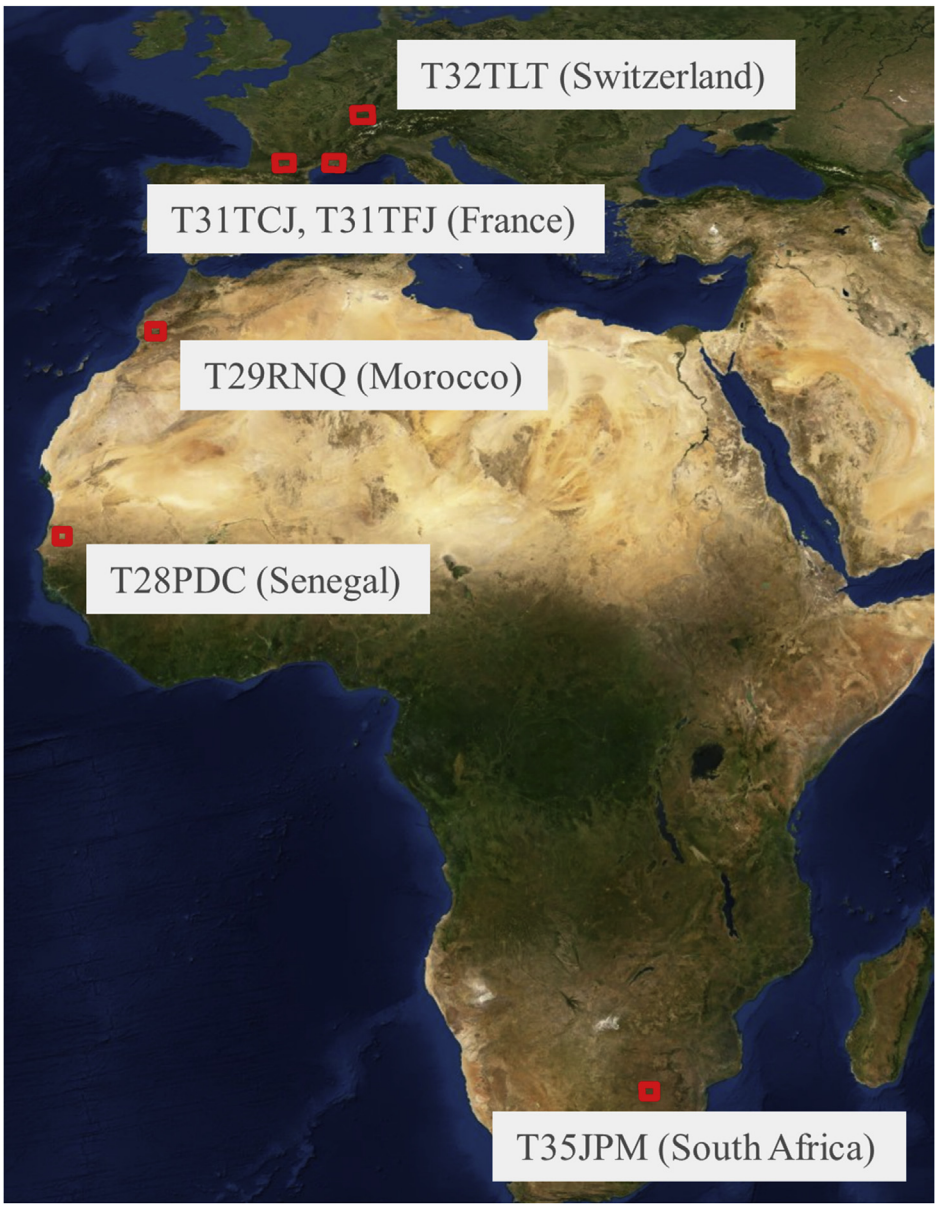

Fig. 1. Study area consists of six sites: T28PDC in Senegal, T29RNQ in Morocco, T31TCJ and T31TFJ in France, T32TLT in Switzerland, and T35JPM in South Africa.

extension cloud shadows (Zhu et al., 2019, 2015). More recently, shortwave infrared (around 1.36–1.39 μm) bands in strong water absorption features have been added to improve the detection of cirrus clouds (commonly referenced as 'cirrus bands') (Zhu et al., 2019, 2015).

As the thermal band is still a key data source for cloud detection, cloud screening in Sentinel-2 is challenging. Though some cloud detection approaches have been developed for Sentinel-2, few have been documented with thorough validation and accuracy assessments; the most cited studies use simulated Sentinel-2 data (Hagolle et al., 2010; Zhu et al., 2015). Some researchers have experimented with using parallax effects to identify clouds, providing a measurable parallax exists (Skakun et al., 2017; Frantz et al., 2018). Recently, machine learning techniques have been employed to address Sentinel-2's cloud issues. Hollstein et al. (2016) applied decision trees and Bayesian models to cloud detection in Sentinel-2 images and released them in ready-to-use formats. Singh and Komodakis (2018) applied generative adversarial networks to model synthetic cloud-free Sentinel-2 data. Liu et al. (2019) used manually-edited cloud masks from Sentinel Level-1C products to construct a residual learning and semantic segmentation network to identify cloud features, Shendryk et al. (2019) trained deep convolutional neural networks (CNNs) to classify clouds, cloud shadows and land cover at the scene-level. These machine learning approaches rely heavily on the cloud and surface conditions of images used for training data, and do not generalize well over a broad range of scenarios (Huang et al., 2010; Zhu et al., 2019).

Few studies have compared the performance of the masks produced by different cloud and cloud shadow detection algorithms. To date, only intercomparison of Sentinel-2 cloud detection algorithms is reported: it featured Fmask, a cloud masking algorithm used by the United States Geological Survey (USGS) for operational processing of Landsat data, Sen2Cor, the algorithm used by the European Space Agency (ESA) to create Sentinel-2 surface reflectance products, and the MACCS-ATCOR Joint Algorithm (MAJA), the cloud detection and atmospheric correction algorithm of the French Center National d'études Spatiales (CNES). This study found that Fmask and MAJA performed similarly, with overall mask accuracies for both cloud and cloud shadows around 90%, while Sen2Cor produced an overall accuracy of 84% (Baetens et al., 2019). Notably, this study relied on reference masks produced from an active

Table 1
A stratified random sample based on agreement and disagreement among cloud masks produced by LaSRC, MAJA, and Sen2Cor. The 24 disagreement combinations were merged into four, together with the three agreement combinations, to create seven strata. A total of 2000 sample units were allocated among the four disagreement strata and 700 sample units among the three agreement strata following the recommendation of Olofsson et al. (2014).

Agreement	Sen2Cor	MAJA	LaSRC	Proportion of area		Stratum	Sample Size
All agree	Clear	Clear	Clear	27.89%	65.22%	1	300
	Shadow	Shadow	Shadow	0.10%		2	100
	Cloud	Cloud	Cloud	37.23%		3	300
One disagrees with other two	Clear	Clear	Shadow	0.39%	4.29%	4	500
	Clear	Shadow	Clear	2.92%			
	Shadow	Clear	Clear	0.11%			
	Clear	Shadow	Shadow	0.18%			
	Shadow	Clear	Shadow	0.03%			
	Shadow	Shadow	Clear	0.25%			
	Shadow	Shadow	Cloud	0.07%			
	Shadow	Cloud	Shadow	0.34%			
	Cloud	Shadow	Shadow	0.0001%			
	Clear	Clear	Cloud	7.41%	16.32%	5	500
	Clear	Cloud	Clear	8.73%			
	Cloud	Clear	Clear	0.18%			
	Clear	Cloud	Cloud	9.06%	11.73%	6	500
	Shadow	Cloud	Cloud	0.17%			
	Cloud	Clear	Cloud	1.54%			
	Cloud	Shadow	Cloud	0.18%			
	Cloud	Cloud	Clear	0.78%			
	Cloud	Cloud	Shadow	0.004%			
All disagree with each other	Clear	Shadow	Cloud	1.19%	2.45%	7	500
	Clear	Cloud	Shadow	0.95%			
	Shadow	Clear	Cloud	0.04%			
	Shadow	Cloud	Clear	0.26%			
	Cloud	Clear	Shadow	0.0006%			
	Cloud	Shadow	Clear	0.01%			

Table 2
Overall accuracy assessment for algorithm results across all sites, weighted by the stratification. Bottom right cell is overall accuracy. Y-axis (rows) are the reference labels ("truth"), X-axis (columns) are algorithm results.

TMASK

	Clear	Not Clear	User's
Clear	0.306	0.206	59.8%
Not Clear	0.064	0.425	87.0%
Producer's	82.8%	67.4%	73.1%

TMASK

	Clear	Shadow	Cloud	User's
Clear	0.306	0.016	0.19	59.8%
Shadow	0.044	0.077	0.066	41.4%
Cloud	0.02	0.005	0.277	91.8%
Producer's	82.8%	78.7%	52.0%	66.0%

FMASK (4.0)

	Clear	Not Clear	User's
Clear	0.351	0.136	72.1%
Not Clear	0.018	0.495	96.6%
Producer's	95.2%	78.4%	84.6%

FMASK (4.0)

	Clear	Shadow	Cloud	User's
Clear	0.351	0.043	0.093	72.1%
Shadow	0.01	0.045	0.023	57.9%
Cloud	0.008	0.011	0.416	95.6%
Producer's	95.2%	45.7%	78.1%	81.2%

LASRC

	Clear	Not Clear	User's
Clear	0.325	0.085	79.3%
Not Clear	0.044	0.546	92.5%
Producer's	88.0%	86.6%	87.1%

LASRC

	Clear	Shadow	Cloud	User's
Clear	0.325	0.059	0.026	79.3%
Shadow	0.009	0.019	0.004	57.7%
Cloud	0.035	0.021	0.503	90.1%
Producer's	88.0%	18.9%	94.4%	84.6%

SEN2COR

	Clear	Not Clear	User's
Clear	0.362	0.206	63.7%
Not Clear	0.007	0.425	98.4%
Producer's	98.1%	67.3%	78.7%

SEN2COR

	Clear	Shadow	Cloud	User's
Clear	0.362	0.073	0.133	63.7%
Shadow	0.002	0.019	0.002	80.8%
Cloud	0.005	0.006	0.397	97.3%
Producer's	98.1%	19.4%	74.5%	77.8%

MAJA

	Clear	Not Clear	User's
Clear	0.286	0.05	85.1%
Not Clear	0.083	0.581	87.5%
Producer's	77.5%	92.1%	86.7%

MAJA

	Clear	Shadow	Cloud	User's
Clear	0.286	0.014	0.036	85.1%
Shadow	0.049	0.051	0.012	45.4%
Cloud	0.035	0.034	0.484	87.5%
Producer's	77.5%	51.3%	90.9%	82.0%

learning software used to minimize manual human operator time, limiting the representativeness of these results to the images used for training data. Further, this active learning method required the existence of at least one cloud-free image to meet MAJA's multi-temporal

Table 3
Overall accuracy assessment for secondary model results across all sites, weighted by stratification. Bottom right cell is total accuracy.

Majority prediction				Majority prediction				
	Clear	Not Clear	User's		Clear	Shadow	Cloud	User's
Clear	0.35	0.095	78.7%	Clear	0.35	0.038	0.057	78.7%
Not Clear	0.019	0.536	96.6%	Shadow	0.005	0.045	0.009	76.2%
Producer's	94.9%	85.0%	88.7%	Cloud	0.014	0.016	0.467	94.1%
				Producer's	94.9%	46.1%	87.6%	86.2%
RF STACKING				RF STACKING				
	Clear	Not Clear	User's		Clear	Shadow	Cloud	User's
Clear	0.317	0.029	91.7%	Clear	0.317	0.017	0.012	91.7%
Not Clear	0.052	0.602	92.0%	Shadow	0.02	0.061	0.005	71.1%
Producer's	85.8%	95.4%	91.9%	Cloud	0.032	0.02	0.516	90.8%
				Producer's	85.8%	62.4%	96.9%	89.4%
CONDITIONAL "ADVANTAGE"				CONDITIONAL "ADVANTAGE"				
	Clear	Not Clear	User's		Clear	Shadow	Cloud	User's
Clear	0.286	0.014	95.3%	Clear	0.286	0.007	0.007	95.3%
Not Clear	0.083	0.617	88.2%	Shadow	0.046	0.071	0.009	56.3%
Producer's	77.6%	97.8%	90.3%	Cloud	0.037	0.021	0.516	89.9%
				Producer's	77.6%	71.9%	97.0%	87.3%

algorithm specifications and thereby excluded the selection of reference sites in persistently cloudy areas like the tropics (Baetens et al., 2019). In general, the remote sensing community has largely focused on comparing these different algorithms in the context of their ability to provide surface reflectance products or other biophysical parameters rather than cloud masks (Doxani et al., 2018). A more thorough comparative analysis of existing cloud detection approaches will help users to use Sentinel-2 imagery more effectively. The Committee on Earth Observation Satellites (CEOS) has recently undertaken a Cloud Masking Intercomparison Exercise (CMIX) for moderate resolution satellite imagery, but results have not been released at the time of this paper.

Given the difficulty of cloud/shadow identification for Sentinel-2, an important question is whether existing algorithms could be altered or combined to improve cloud and cloud shadow detection. Alternative approaches can be organized into two basic methods: modifying an existing algorithm or utilizing a suite of algorithms. The latter approach represents an ensemble model, where the outputs of individual classifiers, or 'base models', become the inputs to a new classifier, or 'secondary model'. Ensemble approaches to classification typically take the form of majority voting, where the class with the most votes is selected; alternatively, they can employ a random forest model (Breiman, 2001) to utilize the outputs of numerous decision trees, a process termed as 'stacked generalization', or 'stacking' (Healey et al., 2018). Stacking smooths the combination of individual algorithms by highlighting correct elements and rejecting incorrect elements through the bootstrapping process (Wolpert, 1992).

A fundamental problem in cloud detection is the tradeoff between omission errors, which represent failures to remove contaminated observations, and commission errors, which unnecessarily remove clear, usable observations. The central idea behind stacking approaches is that the specialized algorithms necessarily contain tradeoffs between omission and commission errors. Wolpert and Macready (1997) explain that for any algorithm, high performance in one dimension often results poor performance in another. It is reasonable to ask whether leveraging the outputs of multiple systems could alleviate the cloud over/under-classification tradeoff. Though ensemble approaches have not been tested for cloud detection algorithms, they have been successfully employed in remote sensing contexts for land cover classification and disturbance detection (Engler et al., 2013; Healey et al., 2018).

We explored the relative performance of multiple cloud masking algorithms to understand how to improve Sentinel-2 cloud and cloud shadow detection. We assessed the advantages and limitations of five algorithms and their performance under varying cloud and environmental conditions. Building on these results, we designed and tested three ensemble classifiers to determine whether they can substantially improve cloud/shadow detection.

Table 4
Summary of algorithm major strengths and weaknesses.

Algorithm	Strength	Weakness
Sen2Cor	Lowest commission errors for clear areas	Highest omission of clouds and shadows
LaSRC	Most accurate cloud detection	Underestimates shadows
MAJA	Detects clouds and shadows well	Overestimates cloud presence
Fmask	Detects non-cirrus clouds well	Misses some cirrus, misclassifies some shadows
Tmask	Most accurate shadow detection	Overestimates shadows, misses some cirrus clouds

2. Methodology

2.1. Study areas

Six locations were selected across different landscapes, based in part on image availability early in the Sentinel-2 mission: T28PDC (Senegal), T29RNQ (Morocco), T31TCJ (France), T31TFJ (France), T32TLT (Switzerland) and T35JPM (South Africa) (Fig. 1). Between 3 and 6 images were selected per location at different times of year for a total of 28 images. The sites were selected to include different land cover types (forest, grassland, cropland, water, snow, urban, barren/desert); images were selected to include varying amounts of cloud cover as well as different configurations of clouds (thin cirrus, low cumulus, etc.).

2.2. Algorithms

Five automated cloud detection algorithms were compared in this study, selected for their common operational use by researchers and organizations, broad availability of reference masks and open-access to models:

- **Sen2Cor** – Sentinel 2 Correction (Richter et al., 2011) – a single-date processor designed for land cover classification and atmospheric correction of top-of-atmosphere Level 1C input data; currently used by ESA in Sentinel-2 surface reflectance products. Clouds, cirrus clouds, cloud shadows and snow are included in Sen2Cor's scene classification step, which uses a series of spectral reflectance thresholds, ratios and indices (e.g. NDSI, NDVI) to compute cloud

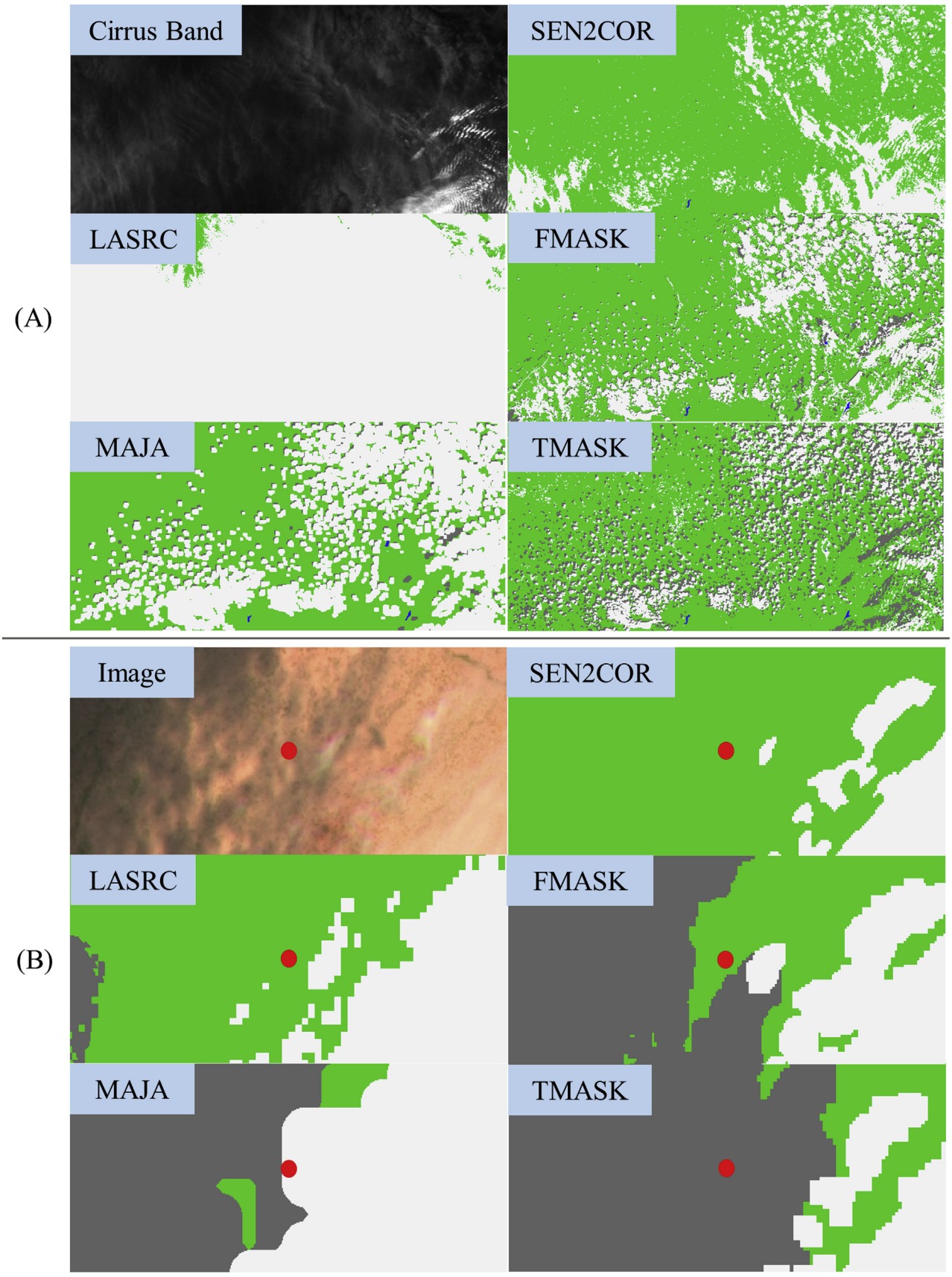

Fig. 2. Panel (A) shows mask results for Sentinel-2 image T29RNQ acquired Apr. 17, 2016 - scenario where only LaSRC identified extremely thin cirrus clouds. Panel B shows mask results for Sentinel-2 image T28PDC acquired Jan. 21, 2016 - scenario where only Tmask identifies cloud shadows. Green area is clear land, grey area is shadow, and white area is cloud. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

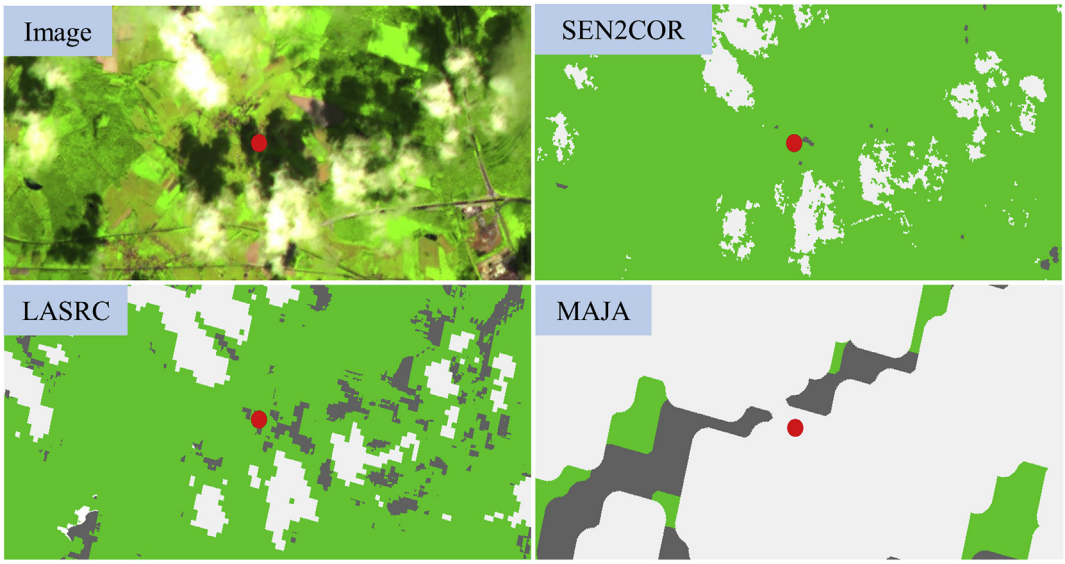

Fig. 3. Comparison of mask results for Sentinel 2 image T32TLT acquired August 4th, 2016 – scenario where LaSRC, Sen2Cor and MAJA failed to detect a shadow—LASRC and Sen2Cor classify the sample point in red as clear and MAJA labels it (and most of surrounding area) a cloud. Green is clear, grey is shadow, and white is cloud. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

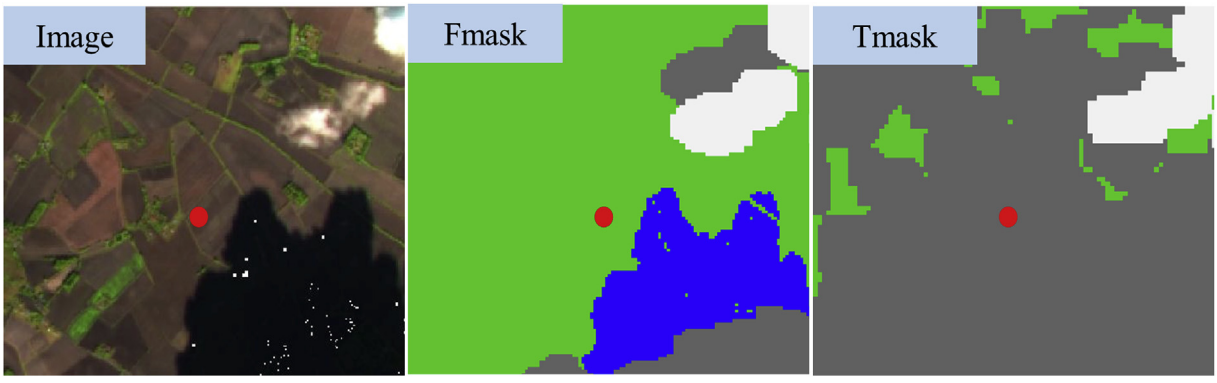

Fig. 4. Comparison of mask results for Sentinel 2 image T31TCJ acquired October 19th, 2016 – scenario where Fmask (3.3) correctly labels a sample point as clear, but misclassifies surrounding shadow as water, while Tmask misclassifies the area at large as shadow. Green is clear land, grey is shadow, white is cloud and blue is water. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

probabilities for each pixel. Thresholding is performed on all bands except the water vapor band (Band 9) and two of the three vegetation red edge bands (Bands 6 and 7). Cirrus clouds are identified using two thresholds in the cirrus band. Cloud shadows are estimated from two probability layers: a geometric probability layer constructed from the sun position, zenith angle, sun elevation and cloud height, and a radiometric probability layer created from a neural network dark area classification. The Sen2Cor version used in this study is 2.4.0, the latest version that was available at the time this study was conducted.

- **MAJA** – MACCS-ATCOR Joint Algorithm (Hagolle et al., 2017) – a spectro-temporal method for cloud detection and atmospheric correction intended for use with Formosat-2, Landsat, VENμS and Sentinel-2; currently used by the French Theia Land Data Center to deliver Sentinel-2 surface reflectance products. MACCS, the Multi-sensor Atmospheric Correction and Cloud Screening algorithm developed by Center d'Etudes Spatiales de la BIOsphère (CESBIO) and CNES was merged in 2015 with modules from ATCOR, the Atmospheric and Topographic Correction software from the German Aerospace Center (DLR). Pixels are classified as low clouds based on their spectral differences relative to a reference composite image that contains the most recent cloud-free observation for each pixel. Two indicators are used (Hagolle et al., 2010; Baetens et al., 2019): when blue and red band values exceed a threshold and when there is low correlation in neighborhood pixel reflectances. High clouds are detected using a single-date method: a cirrus band threshold that varies linearly with altitude. Cloud detection is performed at 240 m resolution. Cloud shadows are computed by searching for darkened pixels within projections derived from estimated cloud altitudes; red band differencing between dates is compared to an average value to identify a cloud shadow.
- **Fmask** – Function of mask (Zhu and Woodcock, 2012) – a single-date, object-based method for cloud and cloud shadow detection in Landsat and Sentinel-2 data. Version 3.3 (Zhu et al., 2015) has been adopted by the USGS for operational processing of Landsat data. It was recently updated to Version 4.0 (Qiu et al., 2019) by integrating Global Surface Water Occurrence (GSWO) and a global Digital Elevation Model (DEM), which can substantially improve cloud and cloud shadow detection for Sentinel-2 data (Qiu et al., 2019). Specially, clouds are detected using a combination of "potential cloud pixels", identified through spectral tests and indices based on cloud physical properties, and a cloud probability layer computed from probabilities for brightness, spectral variability, Haze Optimized Transformation (HOT)-based cloud, and cirrus cloud. These steps incorporate the visible, near infrared, SWIR, and cirrus bands.

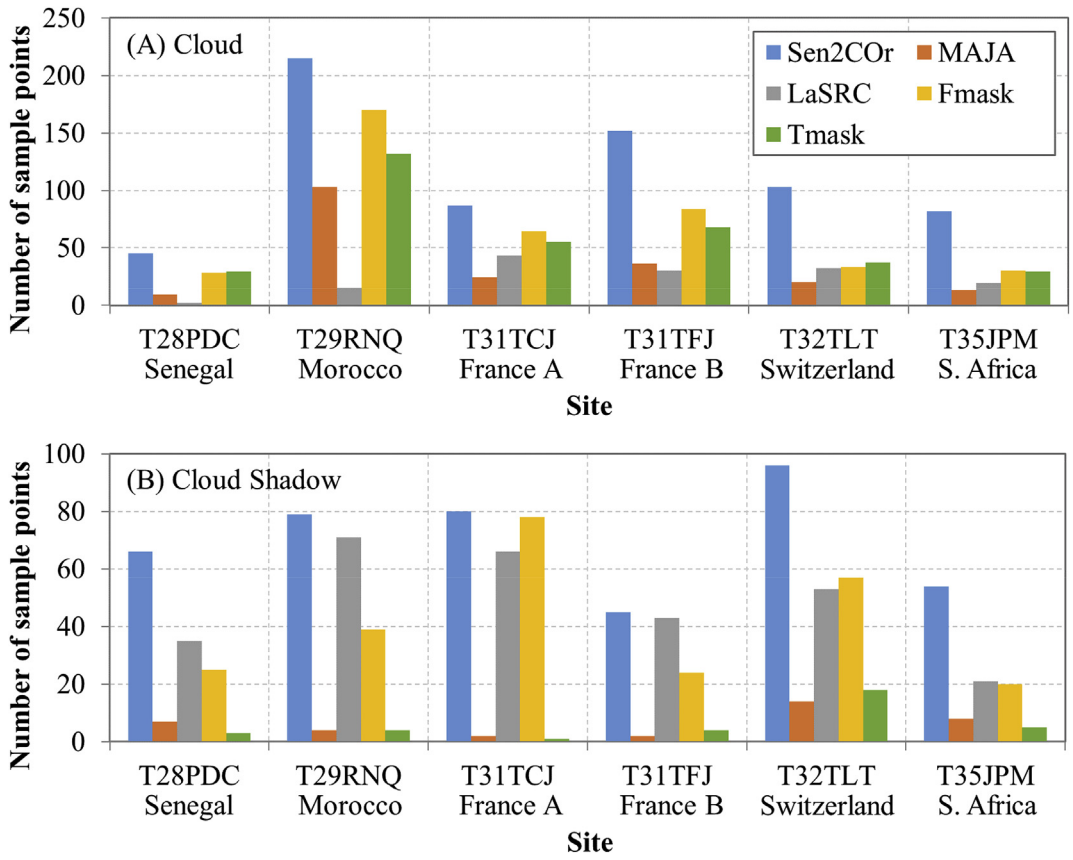

Fig. 5. Distribution of omitted clouds (A) and cloud shadows (B) by site and algorithm.

Moreover, a spectral-contextual optimization is applied to reduce the false identification of clouds from bright and white surfaces (e.g., snow/ice and urban/built-up). Cloud shadows are detected by projecting the segmented cloud objects based on the geometry between sensor view angles, solar angles and cloud height. Here we use Fmask's default parameters of three-pixel cloud dilation and a 20% cloud probability threshold designed for Sentinel-2 imagery (Shi et al., 2019).

- **Tmask** – Multi-temporal mask (Zhu and Woodcock, 2014a) – a multi-temporal method for cloud and cloud shadow detection, which uses the results of Fmask as an input. Fmask is first applied to all available images for initial cloud screening; thresholding in the green band is performed to return potentially misidentified clouds from Fmask. For the remaining clear-sky pixels, time series models of expected surface reflectance for the green, NIR and SWIR bands (Bands 3, 5 and 6 for Landsat-8) are estimated in a Robust Iteratively Reweighted Least Squares (RIRLS) model. Deviations from the predicted reflectance are flagged as either clouds or cloud shadows through spectral thresholding rules. Tmask results for this study were generated using Fmask version (4.0) and Tmask version (1.0).
- **LaSRC** – Land Surface Reflectance Code (Vermote et al., 2016; Skakun et al., 2019) – a single-date, surface reflectance retrieval algorithm for Landsat-8, which includes a cloud mask in its quality assurance (QA) layers produced during the atmospheric correction process. LaSRC uses CFmask, a C programming language implementation of Fmask, to screen clouds prior to estimating reflectance. Clouds are identified as part of the aerosol optical thickness (AOT) retrieval process, which involves calculating an AOT residual using *a priori* ratios of Landsat-8's blue, red and SWIR bands derived from MODIS. Cloud height is estimated from the thermal band information, and cloud shadows are estimated from thresholds for the green, red and SWIR bands (Foga et al., 2017). Cirrus clouds are classified separately from other kinds of clouds. This study used version 3.5.4 of LaSRC.

To adapt to Sentinel-2 specifications, certain modifications were needed for the Landsat-based algorithms. All cloud-masking steps involving thermal bands were removed from Fmask, Tmask and LaSRC. Tmask was applied to Sentinel-2 images with 30 m resolution (given Tmask requires a Landsat time series as an input); the other algorithms were applied to 10 m resolution images. The use of 30 m resolution for Tmask was due to a requirement to fit a time series to observations, and Landsat data were used for this purpose because at the time of this analysis there were not enough years of Sentinel 2 data.

Literature on the performance of these algorithms is largely confined to validation efforts based on subjective cloud reference masks. LaSRC has recently been evaluated as a cloud detection product using the USGS's "L8 Biome" cloud validation dataset, in which it produced commission and omission errors under 4% for thick cumulus clouds but high commission errors for cirrus clouds (Skakun et al., 2019); notably, this validation exercise did not include assessments of cloud shadows and the authors emphasized the subjectivity of reference datasets in interpreting the results (Skakun et al., 2019). Baetens et al. (2019) assessed MAJA's cloud and cloud shadow masks against reference masks produced from an iterative learning approach based on random forest and found MAJA achieves overall accuracies around 90%, with the caveat that small clouds are often omitted, a precision error the authors attribute to the algorithm's dependency on dilation parameters. An early assessment of Sen2Cor's cloud mask product by Coluzzi et al. (2018) found cloud omission errors averaging 37.4% and as high as 87% in areas with high cloud cover and water vapor presence. Fmask has been shown to achieve cloud detection accuracies as high as 96.4% (Zhu and Woodcock, 2012). The algorithm has also been expanded to support Sentinel-2; tests on simulated Sentinel-2 data showed versions of Fmask incorporating

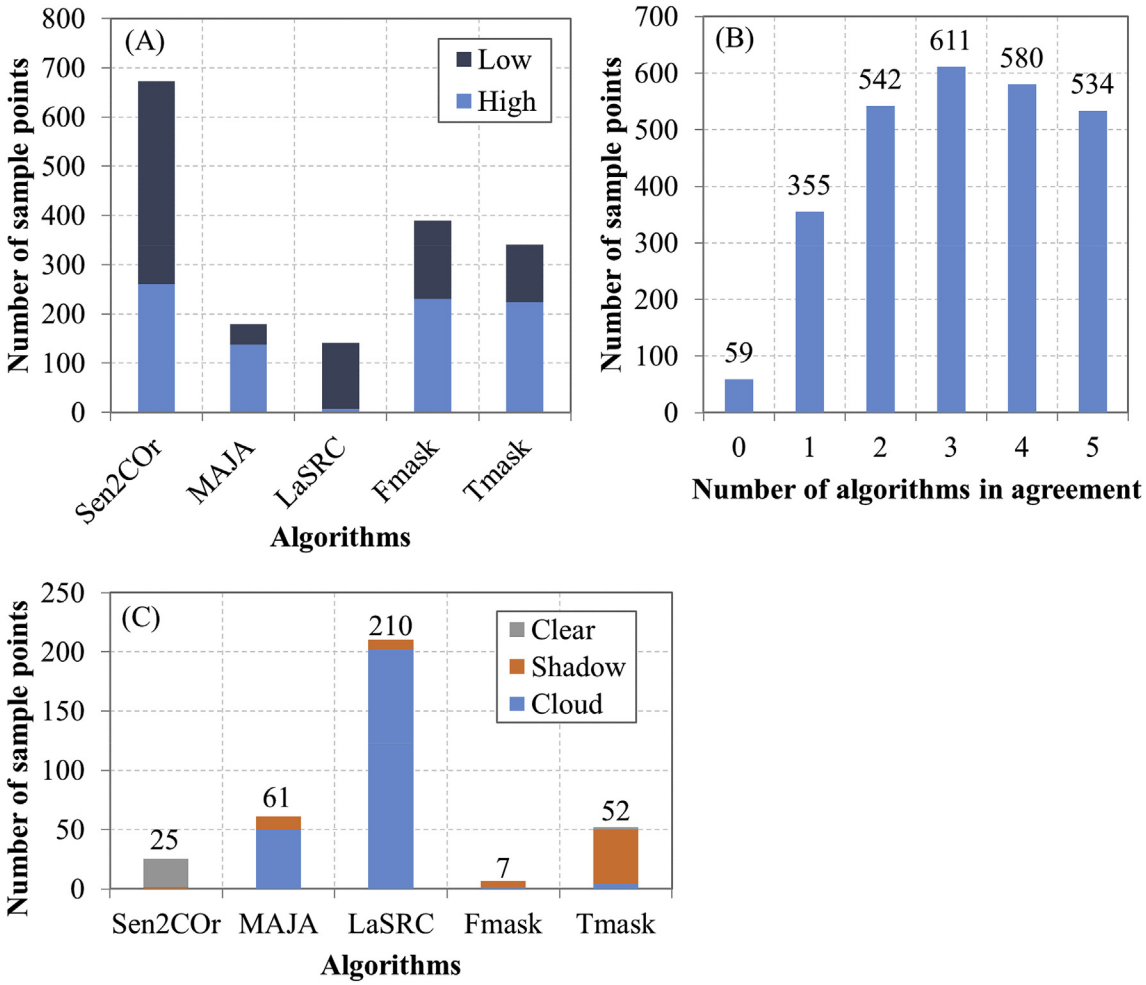

Fig. 6. Panel (A) shows distribution of omitted clouds by algorithm and cloud height (as determined through reference interpretations). Panel (B) shows algorithm agreement with reference data across all sample points. Panel (C) shows class proportions for areas where only one algorithm is accurate. Numbers atop bars are total number of correctly classified sample points unique to each algorithm. All panels represent an aggregate of testing samples across all six sites.

Landsat-8's cirrus band were effective in finding clouds and shadows (Zhu et al., 2015) and the most recent version (4.0) includes a HOT-based cloud probability developed to replace temperature probabilities for Sentinel-2 (Qiu et al., 2019). While Tmask has not been evaluated with a formal accuracy assessment, it has produced robust results in certain scenarios: comparisons with Fmask showed Tmask's pixel-level approach was capable of identifying small clouds often missed by Fmask and significantly out-performed Fmask in detecting cloud shadows (Zhu and Woodcock, 2014a).

2.3. Accuracy assessment

The accuracy assessment used to evaluate the five algorithms is based on a stratified random sample. The population consists of 28 Sentinel-2 images selected from six study sites. Cloud masks produced by LaSRC, MAJA, Sen2Cor, Fmask and Tmask are included in the comparison. Each cloud mask has three classes: *cloud*, *cloud shadow*, and *clear* observations (cirrus clouds are included in the cloud category). The stratification is based on agreement and disagreement among cloud masks produced by LaSRC, MAJA, and Sen2Cor (results from Fmask and Tmask were not available at the time the stratification was created). There is a total of 27 possible combinations, including 3 agreement and 24 disagreement combinations (see Table 1). The disagreement combinations were collapsed into four strata, which together with the three agreement combinations results in seven strata: 1) areas where all three masks found cloud; 2) areas where all three masks found cloud shadow; 3) areas where all three masks found clear land; 4) areas where one mask disagrees with the other two and at least one of the masks indicates cloud shadow; 5) areas where one mask disagrees with the other two and one mask indicates cloud; 6) areas where one mask disagrees with the other two and two masks indicate cloud; and 7) areas where the three masks disagree with each other.

Since the goal was to compare the algorithms, we focused the samples on the places where the results did not agree. The stratum for where they agree was included to allow calculation of overall accuracies. A sample size of 2128 was estimated using Eq. 5.25 in Cochran (1977) targeting a 95% confidence interval of 50% of the area of cloud and cloud shadow combined. A total of 2000 sample units were allocated among the four disagreement strata and 700 sample units among the three agreement strata following the recommendation of Olofsson et al. (2014) (see Table 1)."

Eight analysts visually interpreted sample points to assign class labels of *cloud*, *cloud shadow*, or *clear* to observations. These visually interpreted sample units comprise the reference data used for accuracy assessments of the various algorithms. Though reference labeling was largely informed by the environmental context of images, the Short-Wave Infrared Band (1.365–1.385 μm) was also employed to identify high, thin cirrus clouds that were often not readily visible in the images. For each sample identified as a cloud, interpreters included a label of either *high* or *low* to indicate cloud height (i.e. whether the cloud appeared to be low cumulus/stratus or thin cirrus) and *visible* or *not visible* to indicate if they could be seen in color composites of the spectral bands. As the

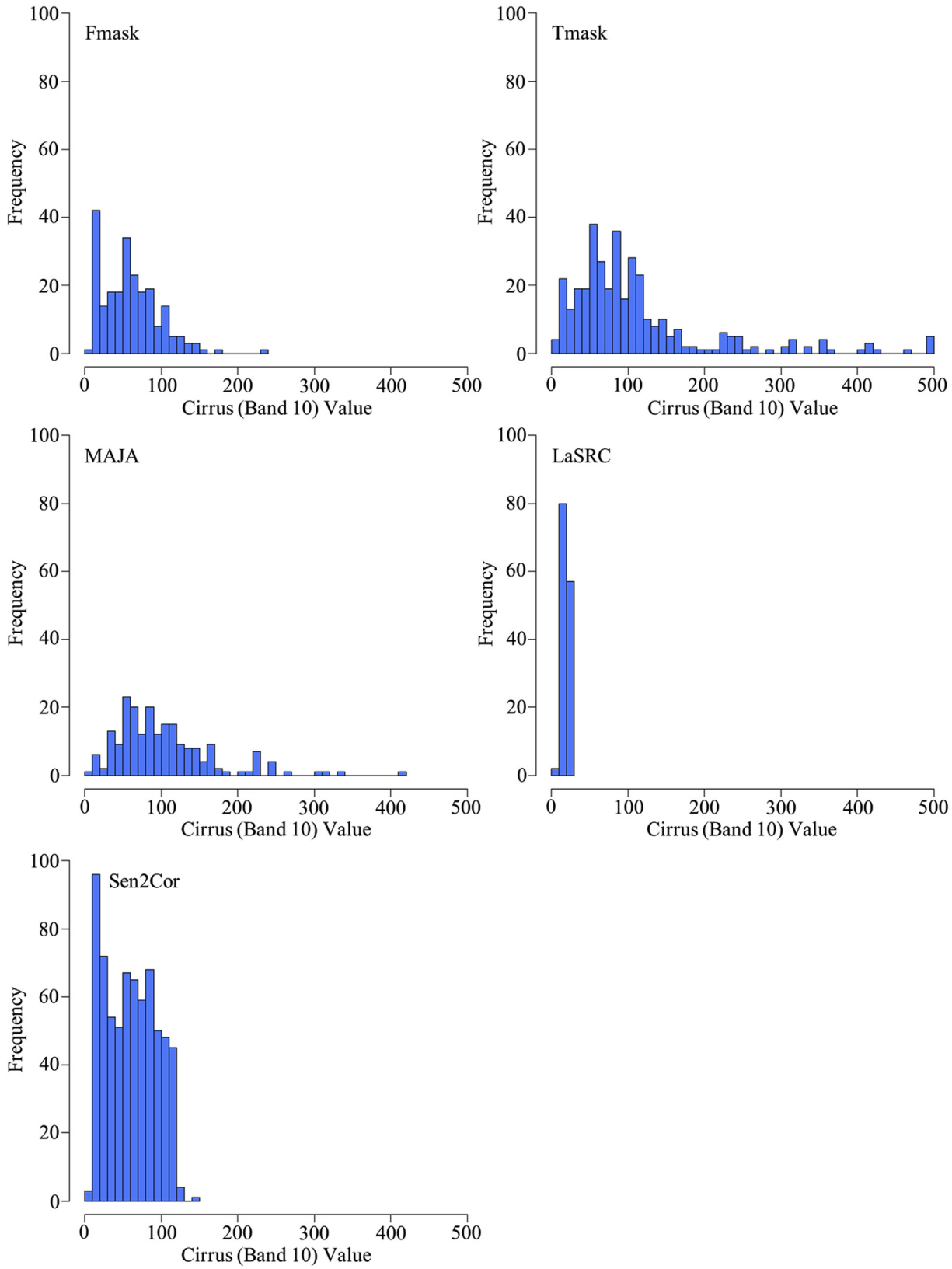

Fig. 7. Distribution of Sentinel-2 band 10 (SWIR/cirrus) values (top of atmosphere or TOA reflectance scaled by 10000) for places where algorithms failed to detect clouds.

central focus of this study is cloud identification, clouds took precedence over shadows in reference labeling, meaning pixels located in areas that could plausibly be considered either clouds or shadows were labeled as *cloud*. Points identified as water or snow by the cloud detection algorithms were considered to be *clear* for the purposes of accuracy assessment. Each sample was interpreted by two different analysts and where

Table 5
"Variance importance" of the stacked/Random Forests model. Percentages reflect predictor variable contribution to classification accuracy.

	Clear	Shadow	Cloud
LaSRC	54.7%	57.9%	84.9%
MAJA	1.5%	22.9%	26.3%
Sen2Cor	26.2%	15.8%	6.1%
Fmask	41.3%	41.1%	23.1%
Tmask	8.2%	63.3%	23.5%

discrepancies existed in their answers the samples were re-examined by both analysts and a final decision on the label was made. When analysts could not resolve differing interpretations for an observation, the point was excluded from further analysis. Of the initial 2700 sample points used in the analysis, 19 were removed for such reasons.

To estimate the accuracies for each algorithm, the stratification must be taken into consideration. While the methods to assemble a confusion matrix and accuracy are well established when the strata correspond to the map labels, our stratification (based on agreement and disagreement of three masks) is different from the mask labels (*cloud*, *cloud shadow*, or *clear*) for any particular algorithm. In this case, error matrices using sample counts are less informative because the sample units contributing to any cell may come from a different stratum and hence carry a different inclusion probability. We used the method suggested by Stehman (2014) to construct error matrices in terms of proportions of the area instead of sample counts. The confusion matrices and overall accuracies were estimated using Eq. (3) in Stehman (2014), while the user's and producer's accuracies were estimated using Eq. (27) in Stehman (2014). Each cell of the confusion matrices is an estimate of the proportion of total population, and the sum of each confusion matrix is 1 (100% of the population).

2.4. Comparative assessment

In addition to accuracy assessments, comparative analyses were performed to explore discrepancies between reference data and mapped cloud and cloud shadows for each algorithm. To assess environmental influence, we examined each algorithm's distribution of omitted clouds and cloud shadows across both sites and individual images, and checked whether areas of snow and water had more frequent classification errors. To assess the nature of misclassified clouds, we tabulated the proportions of misclassified clouds labeled as 'high' or 'low' in the reference data and analyzed the distribution of cirrus band values for omitted clouds. Samples where only one of the algorithms correctly classified sample points were examined to evaluate whether certain algorithms are better at detecting certain types of clouds or cloud shadows.

2.5. Combinations of approaches

Building upon the individual results of the previous section, three additional approaches were created to test whether the use of multiple algorithms improves cloud and cloud shadow detection. The results from the original five algorithms were used as inputs in the following ways:

1). Majority prediction model

A simple combination rule was applied to the five masks to select the class with the highest number of votes. LaSRC's classification was used to settle cases where there were ties because it had the highest cloud detection accuracy.

2). Stacked random forests model

Stacking was applied with the following model:

$$Class_{reference} \sim LaSRC + MAJA + Sen2Cor + Tmask + Fmask$$

Where Random Forests acts as a meta-classifier, identifying clouds and cloud shadows on the basis of results from the original algorithms. This

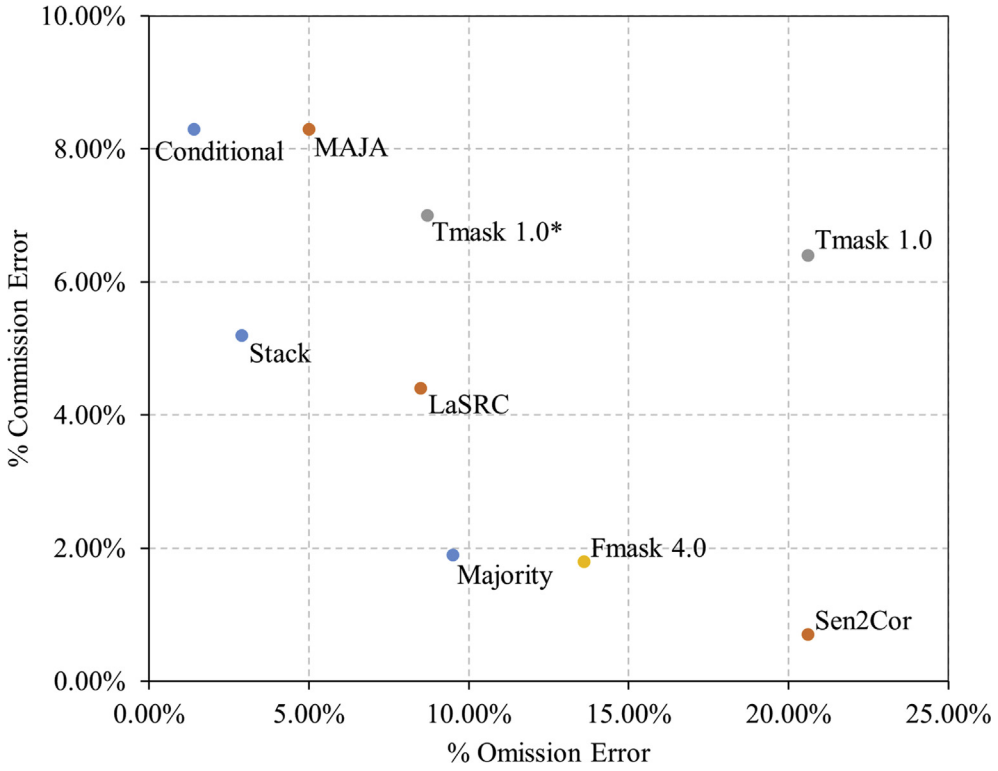

Fig. 8. Comparison of omission and commission error for all algorithms tested (* Tmask with the feature that allows it to correct the Fmask result that it is based on disabled).

analysis was performed with the randomForest package in R (Liaw and Wiener, 2002). Tests showed that class errors stabilized around 500 trees, so this value was used. Half of the reference dataset was used for training and half for testing.

3). Conditional logic/"unique advantage" model

Initial exploratory results suggested that different algorithms had different strengths and weaknesses in their ability to detect clouds or cloud shadows. Based on the results for the individual algorithms, we tried to create a 'logic" based approach that combined the strengths of the various algorithms. LaSRC and Tmask appeared ideal candidates for such optimization, LaSRC for its cloud detection accuracy and Tmask for its shadow detection accuracy. In the interest of conservatively finding contaminated observations, a "unique advantage" model was applied with the following conditional logic:

If *LaSRC = cloud* then *label cloud*
Else if *Tmask = shadow* then *label shadow*
Else *use Fmask label*

In this approach, LaSRC's cloud mask is applied first; Tmask's shadow mask is then applied to the remaining pixels not classified as a cloud by LaSRC; finally, Fmask is used as a base layer for the remaining pixels. Fmask was selected for this final step because: 1) it was most effective at identifying clear observations, as opposed to MAJA, which tended to underestimate clear observations and Sen2Cor, which tended to overestimate clear observations 2) we assumed most of Fmasks errors of omission for clouds and shadows would likely be covered by LaSRC and Tmask's results, and 3) there seemed to be a possibility that Fmask could detect some additional clouds not found by LaSRC.

Accuracies for these approaches that combined the results from other algorithms were assessed using the same sample points and stratification-weighing scheme described in the sampling methods.

3. Results

3.1. Individual algorithms

Table 2 reports each algorithm's results: LaSRC has the highest overall accuracy (87.1%), followed by MAJA (86.7%), Fmask (84.6%), Sen2Cor (78.7%) and Tmask (73.1%). Table 4 presents a descriptive summary of strengths and weaknesses. More specifically, LaSRC identifies clouds particularly well, with the highest producer's accuracy and fewest commission errors for the cloud category. Conversely, LaSRC does not detect shadows well, often misidentifying shadow pixels as clear. MAJA also detects clouds and cloud shadows fairly well, producing the fewest errors of omission for these categories. From a different perspective, this same feature is also a weakness: MAJA frequently misclassifies clear land areas as clouds or cloud shadows. Sen2Cor performs the poorest of all algorithms assessed. Like MAJA, Sen2Cor's apparent advantage is also its disadvantage; though it has the fewest errors of commission for the clear class, it fails to detect many clouds and cloud shadows, instead misclassifying them as clear observations. Fmask performs well in identifying clear areas and some shadows, but there are many instances where it mislabels clouds and cloud shadows as clear areas. Tmask detects cloud shadows most accurately of all the algorithms, with the highest producer's accuracy for shadows; however, it also tends to find too many shadows, often misidentifying clear pixels as shadows. Tmask also frequently misclassifies clouds as clear land.

3.2. Combination approaches

Table 3 shows the accuracy assessment for the approaches that combined the results of the individual algorithms: the RF model produced the highest accuracy (89.4%), followed by the conditional "advantage" model (87.3%) and the majority prediction (86.2%). Both the RF and conditional model accurately found a high proportion of the clouds. The majority prediction model correctly identified fewer clouds than the other two approaches, but also correctly classified more clear areas. All three combinations detected a comparable number of shadows accurately. A notable difference among these approaches is the conditional model has fewer errors of omission for the unclear classes, instead tending to find too many clouds and shadows, while both the stacked and majority approaches show the reverse; higher numbers of missed clouds/cloud shadows and lower numbers of falsely-detected unclear areas. Using Fmask as a base map for the conditional model generated 36 additional misclassifications (meaning errors in cloud and shadow classes outside of those from LaSRC and Tmask) and 79 correct classifications, 74 of which were clouds (meaning Fmask correctly identified 74 samples that were clouds that LaSRC did not find).

4. Discussion

4.1. Algorithms comparison

Sen2Cor performs relatively poorly in identifying cloudy/shadowed observations, failing at the primary task of identifying observations affected by clouds and shadows. MAJA recognizes many clouds and shadows accurately, but it frequently identifies clear observations as clouds or shadows (high errors of commission). While unnecessary removal of clear observations (or errors of commission) may be preferable to errors of omission, commission errors can remove many usable observations. These two algorithms occupy opposite ends of the error-spectrum, with Sen2Cor exhibiting many errors of omission and MAJA over-classifying areas as clouds or shadows.

LaSRC, Tmask and Fmask perform well in certain tasks but are limited in others. LaSRC is clearly the best at identifying clouds. As Fig. 2A shows, LaSRC is able to detect thin, non-visible cirrus clouds that the other algorithms often miss. On the other hand, LaSRC performs poorly for detection of shadows and frequently mislabels them as clear observations. Conversely, shadow detection is an advantage of Tmask, as it can identify even faint shadows that go undetected by the other methods (see the example in Fig. 2B). However, Tmask tends to identify too many areas as shadows. Tmask's major limitation is that it does not identify clouds very well when used alone, often missing cirrus clouds. Fmask appears to be the most balanced algorithm in terms of its commission and omission errors.

Fig. 3 and Fig. 4 illustrate common errors for the different algorithms. Fig. 3 shows how LaSRC and Sen2Cor often miss shadows, while MAJA tends to identify too many observations as clouds. Fig. 4 shows how Fmask frequently misclassifies shadows as clear areas (in this case as water), while Tmask tends to include too many pixels as shadows.

In general, environment type did not appear to significantly influence the ability of various algorithms to detect clouds and cloud shadows; errors are mostly uniformly distributed across the different environments (Fig. 5). The one exception is the site in Morocco (T29RNQ), which had higher rates of missed clouds than other places, most of which were relatively thin cirrus clouds (Fig. 5). This location is at a much higher altitude than other sites and also contains mountains. Elevated, mountainous areas are known to lead to false detection of clouds given both have similar bright and cold features (Selkowitz and Forster, 2016), and thus it is possible that the cirrus thresholding components of some algorithm methodologies mistakenly excluded these areas. Notably, LaSRC did not have reduced performance at this site. Identifying clouds over snow or water did not prove challenging for the algorithms tested: less than 2% of omitted clouds were mapped as clear in snow and water areas.

The properties of undetected clouds vary among algorithms. Fig. 6 shows the number of cloudy samples omitted by each algorithm and the proportion rated as either 'high' cirrus or 'low' cumulus by interpreters across all six sites. Sen2Cor misses twice as many low, thick clouds as high cirrus clouds (Fig. 6A). Conversely, the majority of omitted clouds for MAJA, Fmask and Tmask are high cirrus clouds (Fig. 6A). LaSRC

misses very few high clouds overall, instead misclassifying mostly low clouds (Fig. 6A). These same places of undetected clouds were assessed via their distribution of Band 10 (the "cirrus" band) values (Fig. 7), which reflects these trends. All the algorithms have right-skewed distributions of Band 10 values for misclassified clouds, meaning they typically find pronounced cirrus clouds; however, MAJA, Fmask and Tmask occasionally miss some of these high-value cirrus clouds (Fig. 7). Clouds missed by LaSRC and Sen2Cor have markedly lower Band 10 values, especially LaSRC.

In total, all five algorithms agree with reference interpretations for 20% of sample points (Fig. 6B). Between 2 and 4 algorithms agree with one another and reference data for 65% of sample points (Fig. 6B). For roughly 2% of sample points, no algorithm produces a classification that agrees with reference interpretations (Fig. 6B). Many of these cases are areas of mixed shadows and clouds where either interpretation is reasonable; thus, these are not egregious errors on part of the algorithms and, being identified as "unclear" regardless, are not particularly concerning. Fig. 6C shows the number and class proportions of the 355 cases where only one algorithm was correct. Fig. 6C mirrors the general trends seen in the confusion matrices: LaSRC identifies many clouds where other algorithms failed whereas Tmask detects shadows that other algorithms missed (Fig. 2).

It should be noted that we chose to evaluate the algorithms in areas that are inherently difficult to classify: we sampled from areas of disagreement to highlight differences. Thus, reported overall accuracies are likely low relative to what might be obtained were the entire population of Sentinel-2 images sampled regardless of disagreement.

This study also contains some limitations with regards to methodologies. Our analysis was performed across 28 images from different study sites, and such limited number of images may not be ideal for comparing cloud masking algorithms given cloud type and coverage tend to vary considerably. The decision to use 28 images was based on image availability at beginning of the Sentinel-2 mission, when the reference interpretations were performed. We also constructed our stratification based on agreement and disagreement of cloud masks for LaSRC, MAJA and Sen2Cor, excluding areas where Fmask and Tmask agree or disagree with other classes. As this design was intended to emphasize differences among algorithms, not including Fmask and Tmask in the design may have omitted a more thorough assessment of possible strengths, such as reduced cloud commission errors for Fmask across land covers commonly misclassified as clouds, like urban and snow areas.

4.2. Improvements beyond the original algorithms

An illuminating result from the efforts that combines the results of multiple algorithms is displayed in Table 5, which shows the 'variable importance' output from the stacked RF model. This metric makes use of prediction error rates used to construct the decision trees by measuring the 'mean decrease in accuracy', or the percentage that the prediction error increases, produced by removing a predictor from the model; it represents the extent to which a variable (in this case, the algorithms) contributes to classification accuracy (Liaw and Wiener, 2002). Results from support previous findings: LaSRC is the most significant predictor for cloud and cloud shadows, followed by Fmask for clear and shadow areas; Tmask, while adding little information to clear and cloud classes, is a robust predictor for shadows; MAJA and Sen2Cor contribute relatively less to overall classification accuracy.

Not surprisingly, the models that combine the results from the original algorithms achieve better cloud detection results than any of the individual algorithms. Ensemble accuracies surpass those of Sen2Cor and Tmask by over 10% and are 7–8% higher than for Fmask and MAJA, though these differences are less pronounced when considering the collapsed 'clear/not clear' classes (Table 2). LaSRC is the only standalone algorithm that produces comparable cloud detection accuracies to the combined approaches. Though it has a lower overall accuracy than any of the models tested, LaSRC surpasses the majority prediction model in cloud detection and has fewer omitted clouds and shadows (Table 2). LaSRC, the stacked RF model and the conditional "advantage" model correctly identify a similar number of clouds, but LaSRC ultimately misses more cloud-affected points (particularly shadows) (Tables 2 and 3).

All the models that combine results from multiple algorithms improved the cloud/shadow screening process. While the majority prediction model correctly identified many clear areas, it missed the most clouds—this is likely due to the influence of Sen2Cor's chronic omission errors. Yet the majority model also seems to have benefited from Sen2Cor's predictions, as it generated the fewest commission errors for "clear" observations. Stacking, which proved the most effective approach, appears to have successfully navigated poorer performing masks in its predictions. Stacking has been shown to be more successful when class probabilities, rather than predictions, are used in a Random Forests model, but these were not available for the algorithms tested; only Sen2Cor and LaSrC provide cloud probabilities in the form of low, medium or high confidence levels. Thus, it is possible even better results could be achieved for a stacked cloud detection model were cloud probabilities incorporated. The conditional 'advantage' model also presents an interesting case, as it has the fewest undetected clouds and cloud shadows. A built-in limitation of this conditional combination scenario is that it incorporates commission errors of LaSRC and Tmask. This is why confusion matrix values for areas LaSRC maps as clouds are very similar to the corresponding areas in the combination logic model, and the same is true for areas Tmask maps as shadows (Tables 2 and 3). Still, it attains markedly low omission values.

While better overall accuracies were achieved using ensemble methods, there are practical limitations to implementation. The effort associated with generating masks from the different algorithms is an obvious drawback. Further, the multi-temporal algorithms can be difficult to use and require more processing capacity. Meeting individual algorithm input requirements increases required time and effort, especially preparing multiple images for time series-based algorithms.

If the ideal cloud detection algorithm represents a balance between omission and commission errors, the random forest approach appears the most suitable. On the other hand, if the ideal approach weighs removing bad observations (either cloud or shadow) over removing good (clear) observations, the conditional advantage model (making use of LaSRC's cloud masks and Tmask's shadow masks) seems the most useful. Ultimately, the error ratios illustrated in Fig. 8, which represent aggregate testing samples across all 28 images, seem to support the argument that combining approaches can alleviate limitations inherent to the individual original algorithms, a process which may be a viable solution for Sentinel-2 cloud masking in the future.

4.4. Future direction

An important question arises within the framework of this study: what constitutes a cirrus cloud? The interpretation of high clouds is often subjective; there is a continuous spectrum between haze and clouds, both visually and spectrally. As previously mentioned in the Methods Section, interpreters for this study classified high clouds by both image context and cirrus band values, yet our collective results do not establish a definitive threshold for cirrus cloud identification. As Fig. 7 demonstrates, there is much variation in cirrus band reflectances for clouds missed by the algorithms. The median value of these distributions is 64.6 (reflectance $\times$ 10000); the analog to this metric, the median value of band 10 in places where all five algorithms correctly found a cloud, is 61.5 (reflectance $\times$ 1000). Thus, a clear optimal cirrus threshold is not apparent. Should SWIR/cirrus band thresholds, which multiple algorithms in this study make use of to classify high clouds, be reduced? Or would algorithms begin to over-classifying high clouds in high altitude areas? MAJA developers have recently experimented with changing the algorithm's cirrus thresholds based on user feedback, citing this very dilemma as motivation (Hagolle, 2018). The lack of a firm definition of a

cirrus cloud makes it challenging to identify which algorithm would have highest overall accuracy for detecting non-cirrus clouds, which would be of interest to some users that may tolerate cirrus clouds omission for their applications.

5. Conclusion

No one algorithm produces the best results for detecting both clouds and shadows in Sentinel-2 images. The LaSRC algorithm is the most effective at finding clouds, especially for thin cirrus clouds, and the Tmask algorithm for finding shadows. Combining the results from multiple algorithms improves the overall accuracy of the results, but clearly requires more computational effort. The stacked algorithm approach that combined LaSRC, Tmask and Fmask may be a good solution as it produces relatively high accuracies given Tmask and Fmask can detect most clouds well except for thin cirrus clouds, which LaSRC can detect, and requires only 3 of the 5 algorithms tested.

Declaration of interest

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

Acknowledgements

This research was funded by NASA through both the Harmonized Landsat Sentinel effort and the Making Earth System Data Records for Use in Research Environments (MEaSUREs) Program, as well as the USGS through the Landsat Science Team. The Authors would like to thank Shijuan Chen, Qiyuan Fu, Yihao Liu, Xianfei Shen, Yetianjian Wang, and Yingtong Zhang, Chongyang Zhu for the help on the collection of reference data for this study.

References

Baetens, L., Desjardins, C., Hagolle, O., 2019. Validation of copernicus Sentinel-2 cloud masks obtained from MAJA, Sen2Cor, and FMask processors using reference cloud masks generated with a supervised active learning procedure. Rem. Sens. 11, 433. https://doi.org/10.3390/rs11040433.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Cochran, W.G., 1977. Sampling Techniques. John Wiley and Sons, New York.

Coluzzi, R., Imbrenda, V., Lanfredi, M., Simoniello, T., 2018. A first assessment of the Sentinel-2 Level 1-C cloud mask product to support informed surface analyses. Remote Sens. Environ. 217, 426–443. https://doi.org/10.1016/j.rse.2018.08.009.

Doxani, G., Vermote, E., Roger, J.C., Gascon, F., Adriaensen, S., Frantz, D., Hagolle, O., Hollstein, A., Kirches, G., Li, F., Louis, J., Mangin, A., Pahlevan, N., Pflug, B., Vanhellemont, Q., 2018. Atmospheric correction inter-comparison exercise. Rem. Sens. 10, 352. https://doi.org/10.3390/rs10020352.

Engler, R., Waser, L.T., Zimmermann, N.E., Schaub, M., Berdos, S., Ginzler, C., Psomas, A., 2013. Combining ensemble modeling and remote sensing for mapping individual tree species at high spatial resolution. For. Ecol. Manag. 310, 64–73. https://doi.org/10.1016/j.foreco.2013.07.059.

Foga, S., Scaramuzza, P.L., Guo, S., Zhu, Z., Dilley, R.D., Beckmann, T., Schmidt, G.L., Dwyer, J.L., Joseph Hughes, M., Laue, B., 2017. Cloud detection algorithm comparison and validation for operational Landsat data products. Remote Sens. Environ. 194, 379–390. https://doi.org/10.1016/j.rse.2017.03.026.

Franklin, S.E., Ahmed, O.S., Wulder, M.A., White, J.C., Hermosilla, T., Coops, N.C., 2015. Large area mapping of annual land cover dynamics using multitemporal change detection and classification of Landsat time series data. Can. J. Rem. Sens. 41, 293–314. https://doi.org/10.1080/07038992.2015.1089401.

Frantz, D., Haß, E., Uhl, A., Stoffels, J., Hill, J., 2018. Improvement of the Fmask algorithm for Sentinel-2 images: separating clouds from bright surfaces based on parallax effects. Remote Sens. Environ. 215, 471–481. https://doi.org/10.1016/j.rse.2018.04.046.

Hagolle, O., 2018. Improvement of high cloud detection in MAJA for Sentinel-2 images. Series Temporalles. CESBIO. http://www.cesbio.ups-tlse.fr/multitemp/?p=12894.

Hagolle, O., Huc, M., Desjardins, C., Auer, S., Richter, R., 2017. MAJA Algorithm Theoretical Basis Document. CNES, CESBIO & DLR Report ref MAJA-TN-WP2-030 Issue 1.0. https://www.theia-land.fr/sites/default/files/imce/produits/atbd_maja_071217.pdf.

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. https://doi.org/10.1016/j.rse.2010.03.002.

Healey, S.P., Cohen, W.B., Yang, Z., Kenneth Brewer, C., Brooks, E.B., Gorelick, N., Hernandez, A.J., Huang, C., Joseph Hughes, M., Kennedy, R.E., Loveland, T.R., Moisen, G.G., Schroeder, T.A., Stehman, S.V., Vogelmann, J.E., Woodcock, C.E., Yang, L., Zhu, Z., 2018. Mapping forest change using stacked generalization: an ensemble approach. Remote Sens. Environ. 204, 717–728. https://doi.org/10.1016/j.rse.2017.09.029.

Hollstein, A., Segl, K., Guanter, L., Brell, M., Enesco, M., 2016. Ready-to-use methods for the detection of clouds, cirrus, snow, shadow, water and clear sky pixels in Sentinel-2 MSI images. Rem. Sens. 8, 1–18. https://doi.org/10.3390/rs8080666.

Huang, C., Thomas, N., Goward, S.N., Masek, J.G., Zhu, Z., Townshend, J.R.G., Vogelmann, J.E., 2010. Automated masking of cloud and cloud shadow for forest change analysis using Landsat images. Int. J. Rem. Sens. 31, 5449–5464. https://doi.org/10.1080/01431160903369642.

Irish, R.R., 2000. Landsat 7 automatic cloud cover assessment. Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery VI, p. 348. https://doi.org/10.1117/12.410358.

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. Rem. Sens. 72, 1179–1188.

Kennedy, R.E., Cohen, W.B., Schroeder, T.a., 2007. Trajectory-based change detection for automated characterization of forest disturbance dynamics. Remote Sens. Environ. 110, 370–386. https://doi.org/10.1016/j.rse.2007.03.010.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R. News 2, 18–22.

Liu, C.C., Zhang, Y.C., Chen, P.Y., Lai, C.C., Chen, Y.H., Cheng, J.H., Ko, M.H., 2019. Clouds classification from Sentinel-2 imagery with deep residual learning and semantic image segmentation. Rem. Sens. 11 https://doi.org/10.3390/rs11020119.

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.a., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sens. Environ. 148, 42–57. https://doi.org/10.1016/j.rse.2014.02.015.

Pekel, J.-F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature 540, 1–19. https://doi.org/10.1038/nature20584.

Qiu, S., Zhu, Z., He, B., 2019. Fmask 4.0: improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. Remote Sens. Environ. 231, 111205.

Richter, R., Louis, J., Berthelot, B., 2011. Sentinel-2 MSI – Level 2A Products Algorithm Theoretical Basis Document. ESA Report, ref S2PAD-ATBD-0001 Issue 1.8. https://earth.esa.int/c/document_library/get_file?folderId=349490&name=DLFE-4518.pdf.

Selkowitz, D.J., Forster, R.R., 2016. An automated approach for mapping persistent ice and snow cover over high latitude regions. Rem. Sens. 8 https://doi.org/10.3390/rs8010016.

Shendryk, Y., Rist, Y., Ticehurst, C., Thorburn, P., 2019. Deep learning for multi-modal classification of cloud, shadow and land cover scenes in PlanetScope and Sentinel-2 imagery. ISPRS J. Photogrammetry Remote Sens. 157, 124–136. https://doi.org/10.1016/j.isprsjprs.2019.08.018.

Singh, P., Komodakis, N., 2018. Cloud-GAN: cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks. In: International Geoscience and Remote Sensing Symposium. IGARSS) 2018-July, pp. 1772–1775. https://doi.org/10.1109/IGARSS.2018.8519033.

Skakun, S., Vermote, E.F., Roger, J.-C., Justice, C.O., Masek, J.G., 2019. Validation of the LaSRC cloud detection algorithm for Landsat 8 images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12, 2439–2446. https://doi.org/10.1109/jstars.2019.2894553.

Skakun, S., Vermote, E., Roger, J.C., Justice, C., 2017. Multispectral misregistration of Sentinel-2A images: analysis and implications for potential applications. Geosci. Rem. Sens. Lett. IEEE 14, 2408–2412. https://doi.org/10.1016/j.rse.2018.04.046.

Stehman, S.V., 2014. Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. Int. J. Rem. Sens. 35, 37–41. https://doi.org/10.1080/01431161.2014.930207.

Tulbure, M.G., Broich, M., 2013. Spatiotemporal dynamic of surface water bodies using Landsat time-series data from 1999 to 2011. ISPRS J. Photogrammetry Remote Sens. 79, 44–52. https://doi.org/10.1016/j.isprsjprs.2013.01.010.

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. https://doi.org/10.1016/j.rse.2016.04.008.

White, J.C., Wulder, M.A., Hobart, G.W., Luther, J.E., Hermosilla, T., Griffiths, P., Coops, N.C., Hall, R.J., Hostert, P., Dyk, A., Guindon, L., 2014. Pixel-based image compositing for large-area dense time series applications and science. Can. J. Rem. Sens. 40, 192–212. https://doi.org/10.1080/07038992.2014.945827.

Wolpert, D., 1992. Stacked generalization. Neural Network. 5, 241–259.

Wolpert, D.H., Macready, W.G., 1997. No free lunch theorems for optimization. IEEE Trans. Evol. Comput. 1, 67–82. https://doi.org/10.1109/4235.585893.

Woodcock, C.E., Allen, R., Anderson, M., Belward, A., Bindschadler, R., Cohen, W., Gao, F., Goward, S.N., Helder, D., Helmer, E., Nemani, R., Oreopoulos, L., Schott, J., Thenkabail, P.S., Vermote, E.F., Vogelmann, J., Wulder, M.A., Wynne, R., 2008. Free access to Landsat imagery. Science 320, 1011–1012. https://doi.org/10.1126/science.320.5879.1011a.

Zhang, Y., Guindon, B., Cihlar, J., 2002. An image transform to characterize and compensate for spatial variations in thin cloud contamination of Landsat images. Remote Sens. Environ. 82, 173–187. https://doi.org/10.1016/S0034-4257(02)00034-2.

Zhu, Z., Qiu, S., He, B., Deng, C., 2019. Cloud and cloud shadow detection for Landsat images: the fundamental basis for analyzing Landsat time series. In: Remote Sensing Time Series Image Processing, pp. 3–23. https://doi.org/10.1201/9781315166636-1.

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. https://doi.org/10.1016/j.rse.2014.12.014.

Zhu, Z., Woodcock, C.E., 2014a. Automated cloud, cloud shadow, and snow detection in multitemporal Landsat data: an algorithm designed specifically for monitoring land cover change. Remote Sens. Environ. 152, 217–234. https://doi.org/10.1016/j.rse.2014.06.012.

Zhu, Z., Woodcock, C.E., 2014b. Continuous change detection and classification of land cover using all available Landsat data. Remote Sens. Environ. 144, 152–171. https://doi.org/10.1016/j.rse.2014.01.011.

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

Zhu, Z., Zhang, J., Yang, Z., Aljaddani, A.H., Cohen, W.B., Qiu, S., Zhou, C., 2020. Continuous monitoring of land disturbance based on Landsat time series. Remote Sens. Environ. 238, 111116 https://doi.org/10.1016/j.rse.2019.03.009.