Cloud Mask Intercomparison eXercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2

Sergii Skakun, Jan Swevers, Carsten Brockmann, Georgia Doxani, Matej Aleksandrov, Matej Batič, David Frantz, Ferran Gascon, Luis Gómez-Chova, Olivier Hagolle, et al.

Cloud Mask Intercomparison eXercise (CMIX): an evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2

Sergii Skakun [a,b], *Jan Wevers* [c], *Carsten Brockmann* [c], *Georgia Doxani* [d], *Matej Aleksandrov* [e], *Matej Batič* [e], *David Frantz* [f,o], *Ferran Gascon* [g], *Luis Gómez-Chova* [h], *Olivier Hagolle* [i], *Dan López-Puigdollers* [h], *Jérôme Louis* [j], *Matic Lubej* [e], *Gonzalo Mateo-García* [h], *Julien Osman* [k], *Devis Peressutti* [e], *Bringfried Pflug* [l], *Jernej Puc* [e], *Rudolf Richter* [m], *Jean-Claude Roger* [a,b], *Pat Scaramuzza* [n], *Eric Vermote* [b], *Nejc Vesel* [e], *Anže Zupanc* [e], *Lojze Žust* [e]

[a] Department of Geographical Science, University of Maryland, College Park, MD 20742, USA

[b] NASA Goddard Space Flight Center Code 619, Greenbelt, MD 20771, USA

[c] Brockmann Consult GmbH, 21029 Hamburg, Germany

[d] SERCO SpA c/o European Space Agency ESA-ESRIN, 00044 Frascati, Italy

[e] Sinergise LTD, 1000 Ljubljana, Slovenia

[f] Geography Department, Humboldt-Universität zu Berlin, 10099 Berlin, Germany

[g] European Space Agency ESA-ESRIN, 00044 Frascati, Italy

[h] Image Processing Laboratory, University of Valencia, 46980 Valencia, Spain

[i] Centre d'études Spatiales de la Biosphère, CESBIO Unite mixte Université de Toulouse-CNES-CNRS-IRD, 31401 Toulouse CEDEX 9, France

[j] Telespazio France, 31023 Toulouse, France

[k] Thales Services SAS, Labège, France

[l] DLR, German Aerospace Center, D-12489 Berlin, Germany

[m] DLR, German Aerospace Center, D-82234 Wessling, Germany, rudolf.richter@dlr.de

[n] KBR, contractor to the contractor to the U.S. Geological Survey (USGS) Earth Resources Observation and Science Center (EROS), Sioux Falls, SD 57198, USA

[o] Earth Observation and Climate Processes, Trier University, 54286 Trier, Germany

Corresponding author:

Sergii Skakun

Department of Geographical Sciences, University of Maryland, College Park

1153 LeFrak Hall, College Park, MD, USA, 20742

skakun@umd.edu

Highlights

- Ten cloud masking algorithms for Landsat 8 and Sentinel-2 are evaluated
- Algorithm performance varied depending on the reference dataset
- Average overall accuracy for Sentinel-2 was 80.0±5.3% to 89.4±2.4%
- Average overall accuracy for Landsat 8 was 79.8±7.1% to 97.6±0.8%
- Performance of algorithms improved when thin/semi-transparent clouds not considered

Abstract.

Cloud cover is a major limiting factor in exploiting time-series data acquired by optical spaceborne remote sensing sensors. Multiple methods have been developed to address the problem of cloud detection in satellite imagery and a number of cloud masking algorithms have been developed for optical sensors but very few studies have carried out quantitative intercomparison of state-of-the-art methods in this domain. This paper summarizes results of the first Cloud Masking Intercomparison eXercise (CMIX) conducted within the Committee Earth Observation Satellites (CEOS) Working Group on Calibration & Validation (WGCV). CEOS is the forum for space agency coordination and cooperation on Earth observations, with activities organized under working groups. CMIX, as one such activity, is an international collaborative effort aimed at intercomparing cloud detection algorithms for moderate-spatial resolution (10-30 m) spaceborne optical sensors. The focus of CMIX is on open and free imagery acquired by the Landsat 8 (NASA/USGS) and Sentinel-2 (ESA) missions. Ten algorithms developed by nine teams from fourteen different organizations representing universities, research centers and industry, as well as space agencies (CNES, ESA, DLR, and NASA), are evaluated within the CMIX. Those algorithms vary in their approach and concepts utilized which were based on various spectral properties, spatial and temporal features, as well as machine learning methods. Algorithm outputs are evaluated against existing reference cloud mask datasets. Those datasets vary in sampling methods, geographical distribution, sample unit (points, polygons, full image labels), and generation approaches (experts, machine learning, sky images). Overall, the performance of algorithms varied depending on the reference dataset, which can be attributed to differences the reference datasets were produced. The algorithms were in good agreement for thick cloud detection, which were opaque and had lower uncertainties in their identification, in contrast to thin/semi-transparent clouds detection. Not only did CMIX allow identification of

strengths and weaknesses of existing algorithms and potential areas of improvements, but also the problems associated with the existing reference datasets. The paper concludes with recommendations on generating new reference datasets, metrics, and an analysis framework to be further exploited and additional input datasets to be considered by future CMIX activities.

Keywords: cloud, intercomparison, validation, Landsat 8, Sentinel-2, CMIX, CEOS

1 Introduction

Identification of clouds in satellite imagery acquired by passive remote sensing sensors in the visible and infrared parts of the electromagnetic spectrum (EM) is an essential pre-processing step in producing high-quality geoinformation products. Omission of clouds can lead to errors that propagate to high-level products related to Earth surface monitoring, whereas over detection of clouds can lead to a reduced number of valid observations and, therefore, decrease the frequency of cloud-free data. Development of cloud masking algorithms remains an area of active research in the remote sensing community (Foga et al., 2017; Frantz et al., 2018; Hagolle et al., 2010; Hollingsworth et al., 1996; Irish et al., 2006; López-Puigdollers et al., 2021; Qiu et al., 2019; Scaramuzza et al., 2012; Zhu et al., 2015; Zhu and Woodcock, 2012). A range of algorithms utilize satellite image spectral and spatial properties along with decision tree rules to distinguish cloud versus non-cloud regions (Qiu et al., 2019). These algorithms rely mainly on physical properties of cloud reflectance. Utilization of multi-temporal satellite images, where clouds are considered "anomalies" with respect to a cloud-free reference, can generally improve cloud detection (Frantz et al., 2015; Hagolle et al., 2010; Zhu & Woodcock, 2014). With the advancement of machine learning (ML) and deep learning (DL) methods neural networks models are trained to detect clouds in satellite imagery (Chai et al., 2019; Jeppesen et al., 2019; Mateo-García et al., 2020; Segal-Rozenhaimer et al., 2020; Wieland et al., 2019; Xie et al., 2017).

Although a large number of cloud masking algorithms for optical satellite imagery is currently available, there are a limited quantity of studies aiming at their intercomparison. Three studies should be mentioned in this regard. Foga et al. (2017) compared 13 cloud masking algorithms and their variants for cloud detection in Landsat 7 and Landsat 8 data. Their primary objective was to select an algorithm for generating quality assurance (QA) layers when producing operational Landsat data products. They found that CFMask, a C code

version of the Fmask algorithm (Qiu et al., 2019; Zhu et al., 2015), gave the best performance, and this algorithm is currently used within the U.S. Geological Survey (USGS) operational processing chain to generate Landsat Level-1 products (Wulder et al., 2019). Baetens et al. (2018) compared three methods applied to Sentinel-2 data by analyzing 30 images and found large differences in quality, specifically when taking into account the necessary dilation (buffer) of cloud masks. Tarrio et al. (2020) carried out a study comparing five cloud masking algorithms for Sentinel-2 imagery. By analyzing 28 images over six Sentinel-2 tiles using a sample-based approach and analyst-interpreted reference data they found that none of the algorithms yielded the best performance in terms of identifying both cloud and shadow. They also explored ensemble models to integrate outputs from multiple algorithms and found that on average a +2.7% gain can be achieved over the best-performing model, although at the expense of computational performance.

The main objective of this paper is to summarize results of the first Cloud Masking Intercomparison eXercise (CMIX) conducted within the Committee of Earth Observation Satellites (CEOS) Working Group on Calibration & Validation (WGCV). CMIX is an international collaborative effort co-led by National Aeronautics and Space Administration (NASA) and European Space Agency (ESA) aimed at intercomparing state-of-the-art cloud masking algorithms for moderate-spatial resolution (10-30 m) spaceborne optical sensors. CMIX was recommended following the first Atmospheric Correction Inter-comparison eXercise (ACIX) (Doxani et al., 2018), and was conducted in conjunction with ACIX-II-Land and ACIX-II-Aqua (Pahlevan et al., 2021). The focus of this effort is on open and free imagery acquired by Landsat 8 Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS), and Sentinel-2 MultiSpectral Instrument (MSI) sensors, with corresponding cloud masking algorithms applied. Five existing cloud reference datasets for Landsat 8 and Sentinel-2 are utilized to compare ten cloud masking algorithms. Within CMIX, a qualitative

126 definition of “cloud” is adopted, which provides an absolute (spectrally independent)
127 indication of cloudiness in the satellite image. Although rules defining clouds vary across
128 algorithms and reference data, ultimately all data are converted to “cloud” and “non-cloud”
129 classes to perform a consistent intercomparison. Algorithms are compared using the same set
130 of reference data and metrics under identical conditions. Cloud shadows are not considered in
131 this study, since it is typically a cloud-derived product, and its performance heavily depends
132 on accuracy of cloud detection. Consequently, efforts are primarily directed to cloud mask
133 evaluation.

134 The rest of the paper is organized as follows: a brief description of cloud reference
135 data, cloud masking algorithms, and performance metrics is provided in Section 2. Detailed
136 description of results and their implications are respectively presented in Section 3 and
137 Section 4. Section 5 offers recommendations on further activities regarding generation of
138 cloud reference data and intercomparison of algorithms.

139

140 **2 Methods**

141 ***2.1 Cloud reference datasets***

142 Intercomparison of algorithms within CMIX is performed using existing Sentinel-2
143 and Landsat 8 cloud reference datasets (Table 1), which includ Hollstein (Hollstein et al.,
144 2016), PixBox (Paperin et al., 2021a, 2021b), L8Biome (Foga et al., 2017), CESBIO
145 (Baetens et al., 2019) and GSFC (Skakun et al., 2021). These datasets were
146 collected/generated for different purposes using different methodologies and cloud class
147 nomenclatures. Some of the datasets are single-pixel collections (where a minimum mapping
148 unit is a pixel), while others are the collections of connected pixel areas (polygons) or
149 correspond to whole images. For the majority of datasets, pixels were classified manually
150 through photointerpretation by an expert or a group of experts; in others, the labelling process
151 was semi-automatic with extensive manual checking during classification and post-

processing. Geographical distribution of Landsat 8 and Sentinel-2 scenes in the reference datasets is shown in Figure 1.

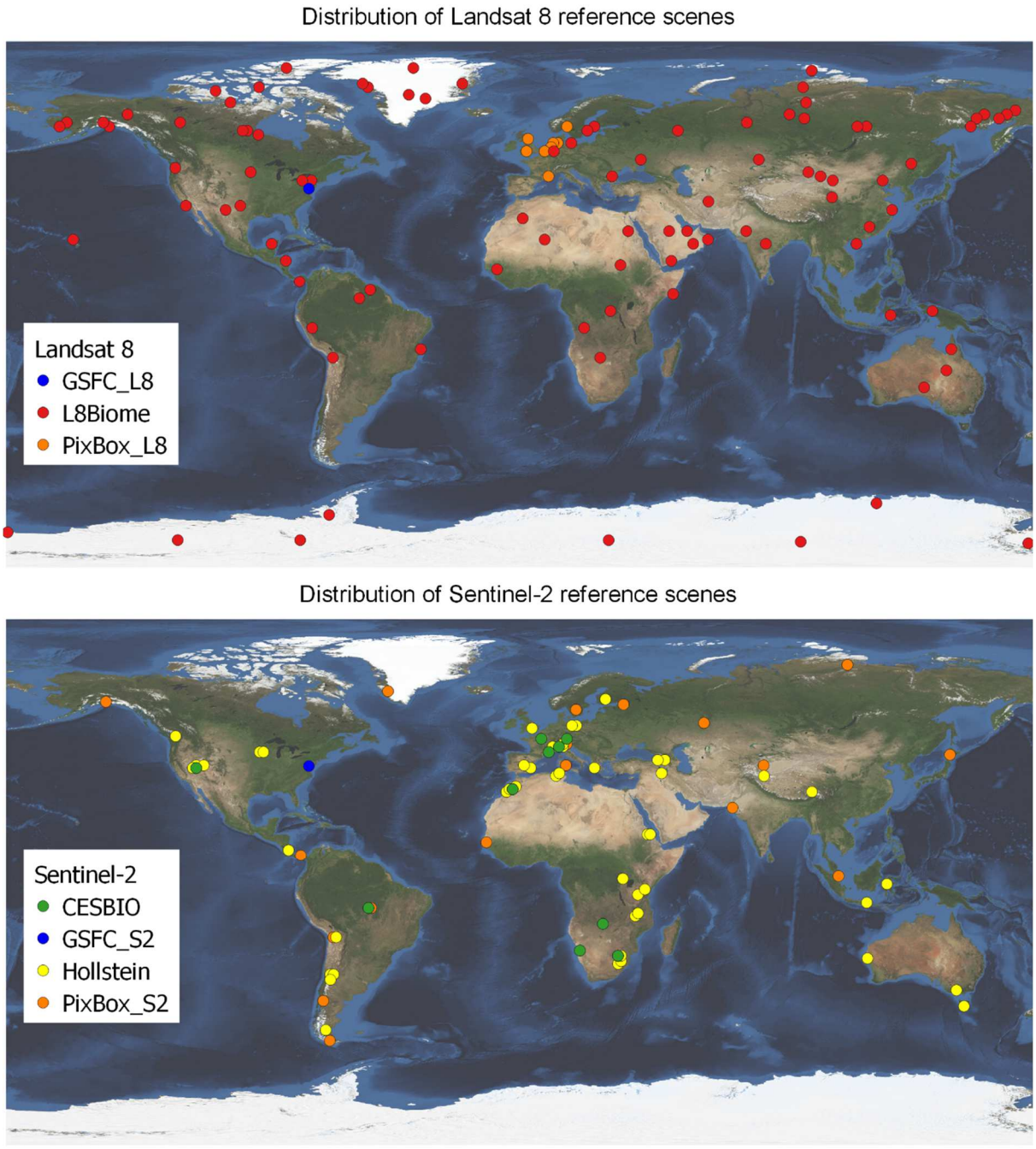


Figure 1. Geographical distribution of the Landsat 8 and Sentinel-2 scenes in the reference datasets used in CMIX.

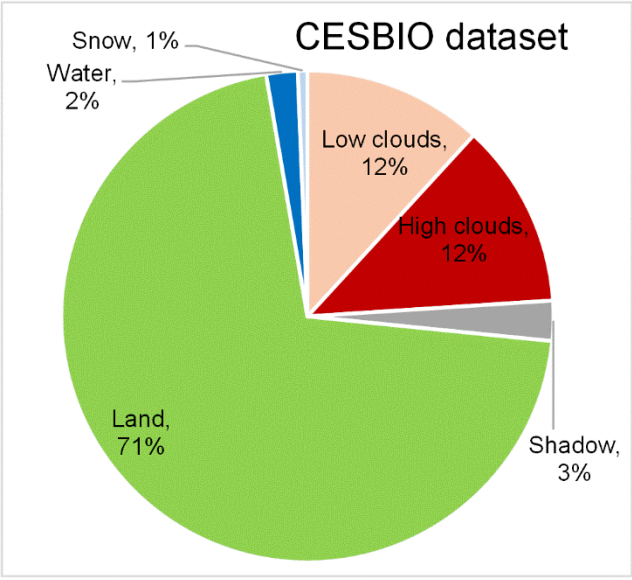
159 Table 1. Summary of cloud reference data (L8: Landsat 8, S2: Sentinel-2).

Dataset	Spatial domain	Level of automatization	Purpose	Thematic depth	Satellites	Spatial resolution	# scenes	Data Availability
CESBIO	Fully classified Sentinel-2 scenes	Classification using an iterative and supervised active learning method	Validation	6 classes	S2	60 m	S2: 30	https://zenodo.org/record/1460961
GSFC	Sample polygons	Manually selected and classified by an expert assisted by ground-based images of the sky	Validation	4 classes	L8, S2	Polygons (in vector format)	L8: 6 S2: 28	https://doi.org/10.17632/r7tnvx7d9g.1
Hollstein	Sample polygons	Manually selected and classified by an expert	Training and validation	6 classes	S2	Polygons (at 20 m)	S2: 59	https://git.gfz-potsdam.de/EnMAP/sentinel2_manual_classification_clouds
L8Biome	Fully classified Landsat 8 scenes	Manually classified by an expert	Training and validation	4 classes	L8	30 m	L8: 96	http://doi.org/10.5066/F7251GDH
PixBox	Sample pixels	Manually selected and classified by an expert	Validation	10 classes	S2, L8	S2: 10 m L8: 30 m	S2: 29 L8: 11	https://zenodo.org/record/5036991 https://zenodo.org/record/5040271

160

161 *2.1.1 CESBIO dataset (Sentinel-2)*

162 The CESBIO dataset was generated using an active learning method (Baetens et al.,
163 2019) using the Hollstein dataset (see section 2.1.3) as training samples. The classification
164 method was iterative, the operator constituted a first set of training samples, and iteratively
165 added other samples, where the classification results were wrong or uncertain. It provides
166 fully classified Sentinel-2 scenes into one of the following classes (Figure 2): low-altitude
167 clouds, high-altitude clouds, cloud shadows, land, water, and snow. In addition to the
168 classification map, a QA layer is provided showing the confidence of classification. Overall,
169 30 Sentinel-2 scenes were utilized in CMIX with the total number of labelled pixels
170 85,782,723 (at 60 m spatial resolution). The scenes were acquired from ten sites around the
171 world, five mainly vegetated and five arid sites. The detailed description of the CESBIO
172 dataset is given in Baetens et al. (2019).



173

174 Figure 2. Distribution of labeled pixels in the CESBIO dataset.

175

176 *2.1.2 GSFC dataset (Landsat 8, Sentinel-2)*

177 GSFC cloud reference data were collected over the NASA Goddard Space Flight
178 Center (GSFC) (Skakun et al., 2021). The area is quite heterogeneous with major land cover

classes being forest (~52%) and impervious surfaces (31%) with patches of natural vegetation and cultivated areas (totaling 17%) (Figure 3). NASA GSFC also has an AERONET station (Holben et al., 1998), which provides aerosol optical thickness (AOT) and water vapor. Ground-based images of the sky were collected from 2017 through 2019 using a smartphone camera with a fisheye lens. These data were collected manually during the Landsat 8 and Sentinel-2 overpasses. Reference data were collected for 6 Landsat 8 and 28 Sentinel-2 scenes. The objective was to capture various cloud conditions and seasonal variability. Labeling of satellite imagery was performed into cloud, thin cloud (semi-transparent), shadows, and clear classes (Figure 3). Regions within cloud boundaries were excluded from the reference data due to large uncertainties regarding the exact boundaries of clouds, especially on Sentinel-2 imagery (Skakun et al., 2021). In order to facilitate the labelling process, Sentinel-2 and Landsat 8 images were presented in various spectral combinations including true color (red-green-blue) and false color (NIR-red-green, SWIR1-NIR-red), and using a cirrus band (at 1.38 µm). The detailed description of the GSFC dataset is given in Skakun et al. (2021).

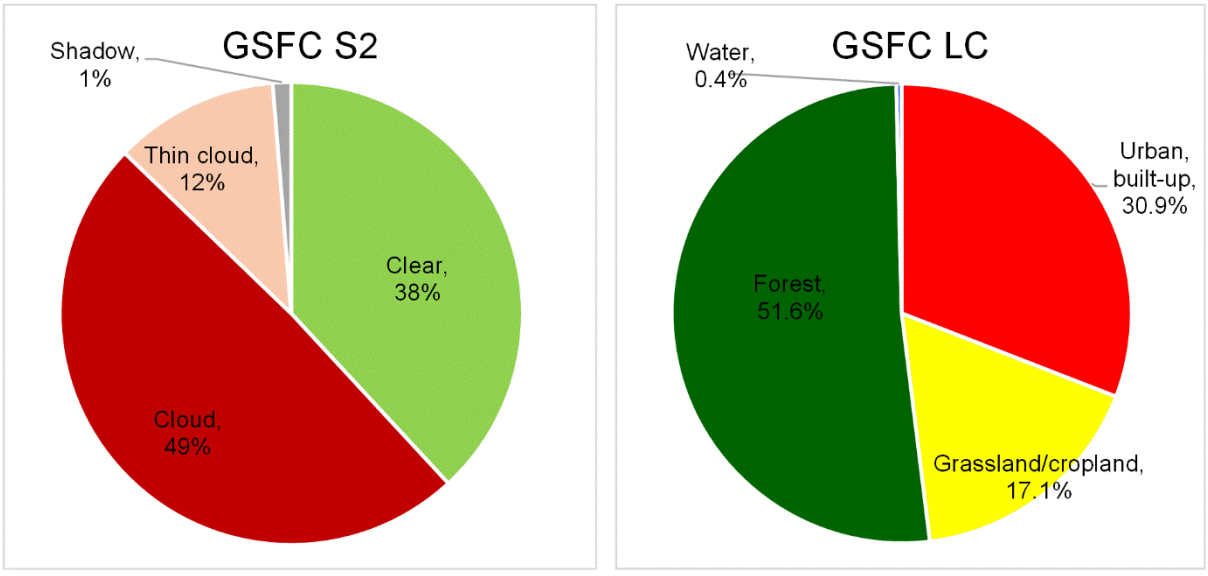
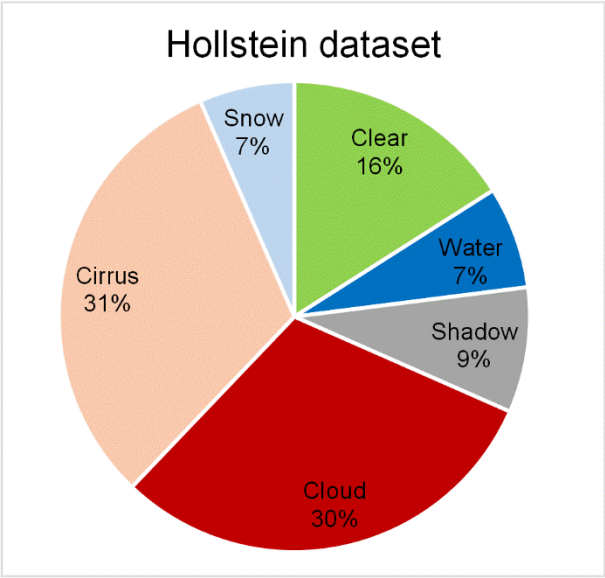


Figure 3. Distribution of labeled pixels in the GSFC S2 dataset (left) and land cover classes (right).

197

198 *2.1.3 Hollstein dataset (Sentinel-2)*

199 The "S2 Hollstein dataset" is a database of manually labeled Sentinel-2A spectra of
200 clouds (Hollstein et al., 2016). By means of different spectral tools, pixels were selected and
201 classified into one of the following six classes (Figure 4): cloud (opaque clouds), cirrus
202 (cirrus, semi-transparent clouds and vapor trails), snow (snow and ice), shadow (shadows
203 from clouds, cirrus, mountains, buildings, etc.), water (lakes, rivers, seas), and clear-sky
204 (other remaining areas). Spectral tools include false-color composites of Sentinel-2 images,
205 image enhancements and graphical visualization of spectra. The aim was to create highly
206 heterogeneous classes with a balanced number of pixels. There were 59 total Sentinel-2
207 scenes and 1,593,911 reference (labelled) pixels.



208

209 Figure 4. Distribution of labeled pixels in the Hollstein dataset.

210

211 *2.1.4 L8Biome dataset (Landsat 8)*

212 The "L8 Biome" cloud validation dataset consisted of 96 Landsat 8 scenes, which
213 were selected using a semi-random sampling by biome (Foga et al., 2017). These biomes
214 included barren, forest, grass/crops, shrubland, snow/ice, urban, water, and wetlands. For

each biome 12 Landsat 8 scenes were selected, and each scene was manually classified by an expert into the following classes (Figure 5): clear, thin cloud, cloud, and cloud shadow. It should be noted that no specific threshold was used to detect thin (semi-transparent) clouds, which were primarily determined by the analyst. Also, the cloud shadow class in the validation dataset was not provided for all the Landsat 8 scenes. The detailed description of the L8Biome dataset is provided in Foga et al. (2017).

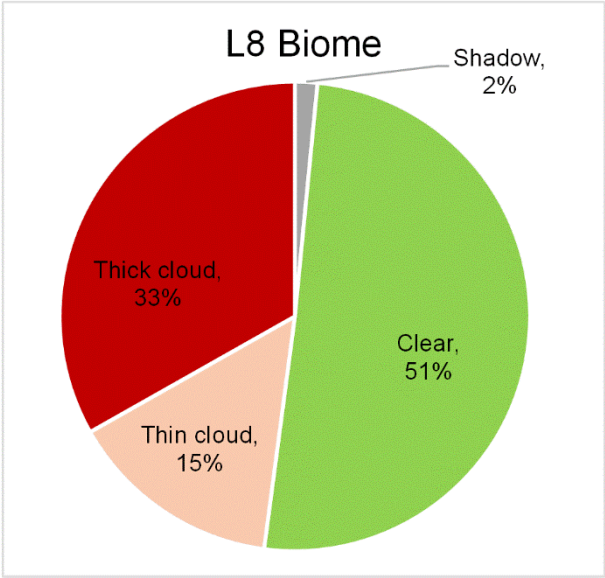


Figure 5. Distribution of labeled pixels in the L8Biome dataset.

2.1.5 PixBox dataset (Landsat 8, Sentinel-2)

The overarching goal of the so called "PixBox" is to enable a quantitative assessment of the quality of a pixel classification produced by an automated algorithm/procedure. Pixel classification is defined as assigning a certain number of attributes to an image pixel, such as cloud, clear sky, water, land, inland water, flooded, snow etc. These pixel classification attributes are typically used to further guide higher level processing. PixBox is not only a dataset but also includes a method comprising a procedure to define the best thematic, spatial and temporal distribution for each collection purpose, a dedicated software for collecting

pixels, the analysis, comparing the collected reference against an automatic classification, as well as the generation of a report.

For the PixBox Reference Dataset, a trained expert(s) manually labels pixels of an image sensor into a detailed set of pre-defined classes. These are typically different cloud transparencies, cloud shadow, and condition of the underlying surface ("semi-transparent clouds over snow", "clouds over bright scattering water"). The collected dataset includes 10's of thousands of pixels because it necessitates representation for all classes, and for various observation and environmental conditions such as climate zones, solar illumination, viewing angles, etc. Prior to the collection process the expert is provided with a detailed list of distribution of categories and classes that needs to be fulfilled. During the collection process the growing database is constantly checked against this reference. Quality control of the collected pixels is important in order to detect misclassifications and systematic errors.

PixBox is a commercially sold product/service of Brockmann Consult GmbH. The following two PixBox datasets have been made freely available to be used for CMIX (Paperin 2021a, Paperin 2021b). The Sentinel-2 PixBox dataset contained 17,351 pixels (at 10 m) manually collected from 29 Sentinel-2A/B Level 1C products (top-of-atmosphere reflectance—TOA reflectance). The Landsat 8 PixBox dataset contained 20,500 pixels (at 30 m) manually collected from 11 Landsat-8 Level 1 products (TOA reflectance). The Sentinel-2 PixBox dataset is spatially, temporally, and thematically evenly distributed, while the Landsat 8 dataset has a strong spatial focus on the Northern European coastal areas. Distribution of labelled pixels and corresponding land clover classes for the PixBox datasets are shown in Figure 6.

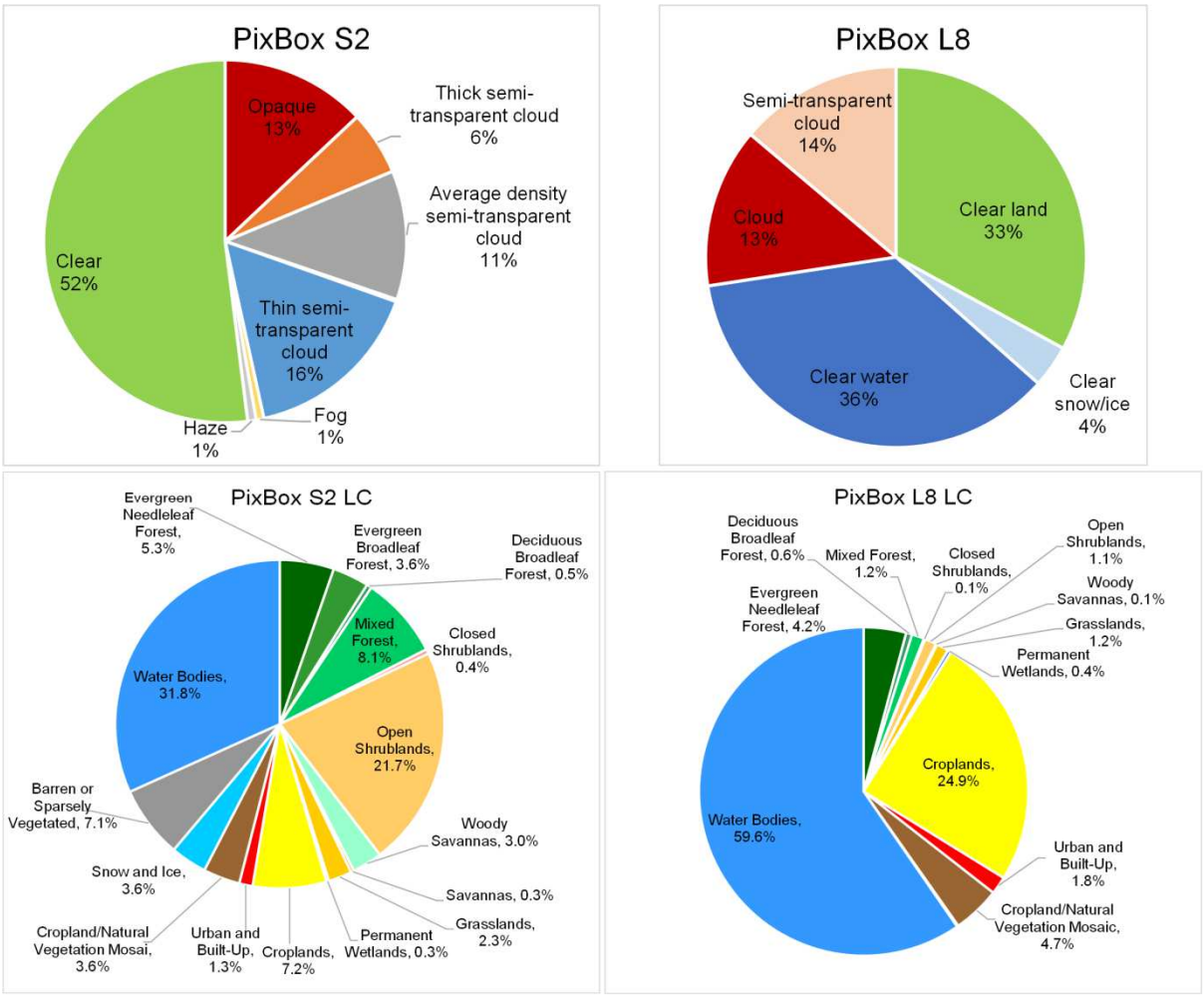


Figure 6. Distribution of labeled pixels and land cover classes in the PixBox dataset.

2.1.6 Summary of strengths and limitations of cloud reference datasets

Table 2 summarizes the strengths and limitations of cloud reference datasets used in this study. Reference data incorporating global coverage and a wide range of image conditions (L8Biome, PixBox, Hollstein) are based on the photointerpretation of images by an expert or a group of experts. This can introduce some subjectivity in labelling clouds, especially for thin/semi-transparent clouds that can be wavelength-dependent and fog (Scaramuzza et al., 2011) (Figure 7), and it is usually difficult to draw the exact boundary between this type of clouds and clear pixels. Another approach is to use high-quality pixels (with no uncertainties in cloud detection) and subsequently apply machine learning

algorithms to extrapolate classification for the whole scene through an iterative process until the classification results assessed by an expert are deemed to be satisfactory (CESBIO) (Figure 8). The quality of the resulting map, however, can still depend on the training data and classification method used. A third approach (GSFC dataset) is to utilize ground-based imagery of the sky to produce a training/validation cloud dataset, either through manual or automatic labelling (Figure 8). While such an approach would potentially decrease subjectivity in identifying clouds, a network of such sites with sky cameras would be required (similar to the Aeronet network) in order to capture various geographical conditions.

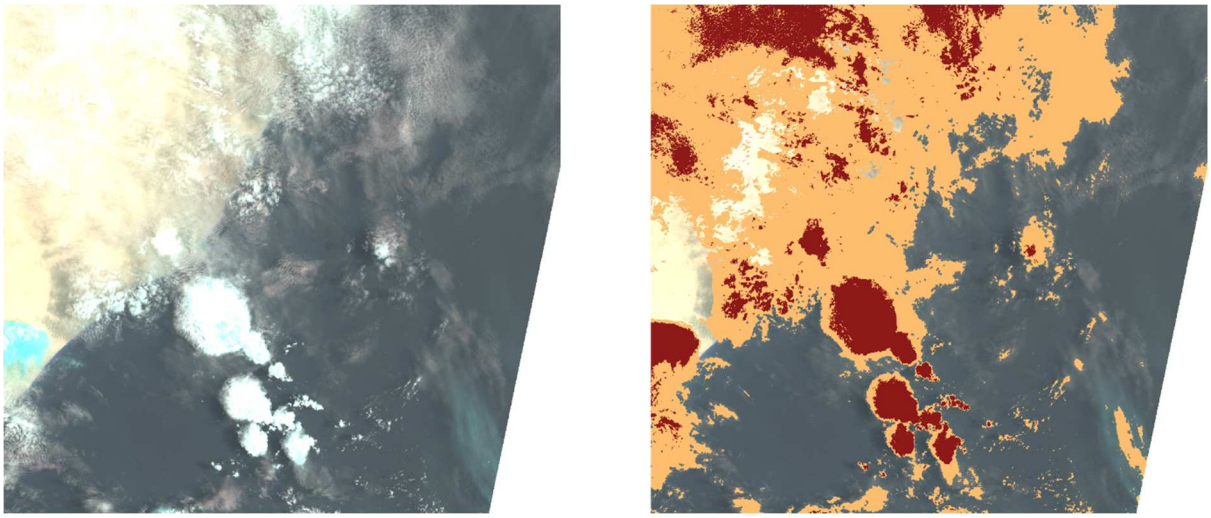


Figure 7. Part of the L8Biome scene (LC81570452014213LGN00) with some thin clouds not labelled. Thin clouds are shown in orange, and thick clouds in maroon.

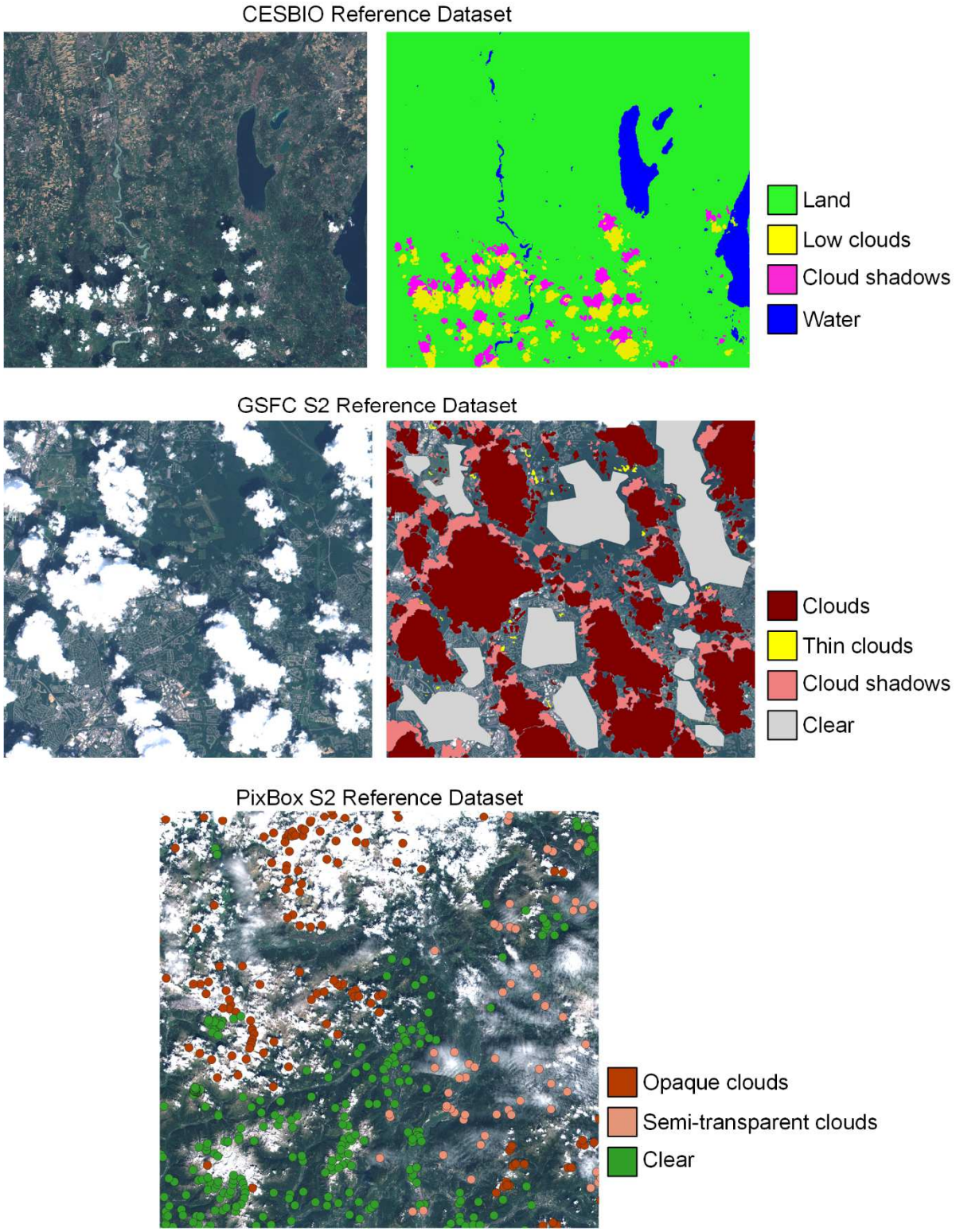


Figure 8. Examples of labeled data in the three datasets: CESBIO (fully labeled images); GSFC (polygons avoiding uncertain areas, such cloud boundaries); PixBox (sample-based approach).

Table A1 (Appendix A) provides a list of classes from the reference datasets that were used to define cloud and non-cloud pixels in the CMIX. Most of the datasets were balanced in terms of cloud and non-cloud pixels, except of CESBIO, which had 24% of cloud pixels (Figure 2). CESBIO, GSFC and Hollstein datasets were primarily over the land surface, while the majority of PixBox datasets was over the water surface: 32% for S2 and 60% for L8.

Table 2. Strengths and limitations of cloud reference datasets.

Dataset	Strengths	Limitations
CESBIO	– All pixels in the scene are classified using an iteratively supervised machine learning approach	– Based on expert knowledge (potential bias). Small number of locations (limited spatial coverage) – Cloud and non-cloud areas unbalanced
GSFC	– Assisted with ground-based imagery – Over the same territory (can be potentially used for temporal consistency analysis)	– Limited field of view and single location – Surface classes limited to the location of sky camera – Cloud boundaries excluded
Hollstein	– Manual classification of polygons using spectral features	– Lack of sample quality – Low level of detail – Based on expert knowledge (potential bias) – Cloud edges not sampled
L8Biome	– Global coverage with stratified sampling – All pixels in the scene are classified	– Based on expert knowledge (potential bias)
PixBox	– High level of detail – High level of classification precision – Global coverage with stratified sampling	– Single pixel, thus a comparably small dataset – Based on expert knowledge (potential bias)

2.2 Cloud masking algorithms

This subsection briefly describes the main concepts utilized in each of the cloud masking algorithms with a summary presented in Table 3.

295 Table 3. Summary of cloud masking algorithms (L8: Landsat 8, S2: Sentinel-2). The "Objective" column shows the intended performance of
296 algorithm in terms of cloud omission/commission errors. "Balanced" means the algorithms aims at balancing omission/commission errors.
297 "Cloud-free conservative" means the algorithm aimed at minimizing cloud omission errors.

Processor	Organization	Methodology	Objective	Spatial resolution, m	Temporality	Buffer for clouds	Shadow detection	References
ATCOR	DLR	Spectral tests	Balanced	L8: 30 S2: 20	Mono	100 m	Yes	Richter & Schläpfer (2019a)
CD-FCNN	University of Valencia	Machine learning	Balanced	L8: 30 S2: 10/20/60	Mono	No	No	Mateo-García et al., (2020), López-Puigdollers et al. (2021)
Fmask 4.0 CCA	USGS	Spectral tests	Balanced	L8: 30 S2: 20	Mono	L8: 90 m S2: 60 m	Yeas	Foga et al. (2017), Qiu et al. (2019), Zhu et al. (2015)
FORCE	Humboldt-Universität zu Berlin / Trier University	Spectral test + parallax (S2 only)	Cloud-free conservative	L8: 30 S2: 10	Mono	300 m	Yes	Frantz (2019), Frantz et al. (2018), Frantz et al. (2016), Zhu et al. (2015), Zhu & Woodcock (2012)
IdePix	Brockmann Consult	Spectral tests	Balanced	S2: 20	Mono	Not used (user-defined)	Yes	Wevers et al. (2021)
InterSSIM	Sinergise	Machine learning + spatio-temporal context	Cloud-free conservative	S2: 10	Multi	160 m	No	Puc & Žust (2019)
LaSRC	NASA / University of Maryland	Spectral tests	Cloud-free conservative	L8: 30 S2: 10	Mono	L8: 150 m S2: 50 m	Yes	Skakun et al. (2019), Skakun et al. (2021), Vermote et al. (2016)
MAJA	CNES / CESBIO	Multi-temporal and spectral tests	Cloud-free conservative	S2: 240	Multi	240 m	Yes	Hagolle et al. (2010), Hagolle et al. (2017)
s2cloudless	Sinergise	Machine learning	Cloud-free conservative	S2: 10	Mono	160 m	No	Zupanc (2017)
sen2cor	ESA / Telespazio France	Spectral test + auxiliary data	Balanced	S2: 20	Mono	No	Yes	Louis et al. (2016), Louis (2021)

298

2.2.1 ATCOR

ATCOR is a generic atmospheric correction algorithm for mono-temporal multi-/hyper-spectral satellite imagery in the solar reflective region (400 – 2500 nm) and thermal region (8-13 µm) (Richter & Schläpfer, 2019b). The code uses MODTRAN5 look-up tables for the radiative transfer functions. Separate codes exist for the processing of flat and rugged terrain imagery. A preprocessing step calculates different masks (water cloud, cirrus cloud, shadow, water) based on spectral tests. The cloud masking uses a buffer of 100 m. For Landsat-8 and Sentinel-2 data the TOA reflectance threshold of the cirrus band is set to 0.01 (reflectance units). The lower threshold for thin cirrus detection was used prevent scenes with very thin cirrus being classified as (thin) cirrus because other classes (e.g., water, shadow) are generally of more interest than very thin cirrus. Cloud detection in ATCOR was aimed to have a balance between commission and omission errors. In CMIX, ATCOR version 9.3.0 (2019) was used. CMIX processing of ATCOR did not use a Digital Elevation Model (DEM) or any other auxiliary data. Some scenes from reference datasets were not processed by ATCOR, since they were acquired with Sun elevation angle values less than 30°.

2.2.2 CD-FCNN

The cloud detection approach based on deep learning, proposed by the Image and Signal Processing (ISP) group of the University of Valencia, is applicable to multispectral images from moderate spatial resolution satellites, including Landsat 8 and Sentinel-2. Training accurate global cloud detection models based on deep learning requires large datasets of annotated images, which must reflect the high variability of clouds, surface, and atmospheric conditions. This is a major difficulty since high-quality labeled datasets usually do not exist or are not publicly available for most satellite sensors. For Landsat 8, the L8Biome dataset matches these requirements (Jeppesen et al., 2019). However, similar global

datasets do not exist for Sentinel-2 yet. (Sentinel-2 Cloud Mask Catalogue (Francis et al., 2020) was made available after CMIX was initiated). Therefore, Landsat 8 datasets (L8Biome, 80%, and L8SPARCS, 20%) were used to train fully convolutional neural networks (FCNN) that may be transferred to perform cloud detection in Sentinel-2 images. L8SPARCS (Spatial Procedures for Automated Removal of Cloud and Shadow) (USGS, 2016) was created for the validation of the cloud detection approach proposed by Hughes & Hayes (2014). It consists of 80 Landsat-8 sub-scenes manually labeled in five different classes: cloud, cloud-shadow, snow/ice, water, flooded, and clear-sky. The size of each sub-scene is 1000x1000 pixels.

After a minimum adaptation of Sentinel-2 data, in terms of band selection and spatial resolution, the models trained on Landsat 8 data are directly applied to Sentinel-2 images. The proposed neural network architecture is based on a modified U-Net with significantly less training parameters and lower computational cost (Mateo-García et al., 2020). It seeks to provide both faster inference time and accurate detection through a lightweight architecture with a moderate number of parameters, i.e., approximately 96,000 parameters, which is around 1% of original U-Net parameters. Moreover, this modified version of U-Net works seamlessly with Landsat-8 and Sentinel-2 images thanks to a transfer learning strategy over both sensors. In this way, all input bands, regardless of the sensor, are homogenized and resampled to 30m overlapping patches of 32x32 pixels, which are used for training the networks in a 64-batch size configuration. Models are trained to minimize a pixel-wise binary cross-entropy cost function, between ground truth and predictions, using the Adam stochastic gradient descent optimization algorithm. An initial learning rate of 10^{-5}, a weight decay of 5 x 10^{-4} and 120 epochs were used to train the final network. The TensorFlow framework was used to implement and train the models on a GPU (average of 800 s/epoch in all configurations). Training and testing details can be found in López-Puigdollers et al. (2021);

in addition, the pre-trained model and a Python-implementation of the proposed cloud detection algorithm for Landsat-8 and Sentinel-2 is provided in a public repository (https://github.com/IPL-UV/DL-L8S2-UV).

Since we propose to use the same model for Landsat-8 and Sentinel-2, we are restricted to bands available in both sensors. In this context, three different bands configurations were tested: "RGBI" corresponds to bands B2, B3, B4 and B5 of Landsat-8 and B2, B3, B4 and B8 of Sentinel-2; "RGBISWIR" to bands B2, B3, B4, B5, B6 and B7 of Landsat-8 and B2, B3, B4, B8, B11 and B12 of Sentinel-2; and "ALLNT" includes all "RGBISWIR" bands plus the coastal aerosols and cirrus bands (B1 and B9 in Landsat-8, B1 and B10 in Sentinel-2, respectively). After internal testing, the network selected for benchmarking in CMIX was the "RGBISWIR" network. Further results about the different band configurations can be found in López-Puigdollers et al. (2021).

The CD-FCNN output is given by a sigmoid activation function that provides continuous values, which could be interpreted as probabilities, between 0 and 1. In order to compare with the rest of the methods, these values are binarized into “non-cloud” (0) or “cloud” (1) classes for each pixel. We set a default 0.5 threshold to obtain the binary cloud mask assuming unbiased data. However, this threshold has a crucial importance in terms of balance between commission and omission and errors. In Landsat-8 images both errors are similar, but performance may decrease in complex scenarios with presence of ambiguous pixels, e.g. over snow, urban areas or coastal lines. Adjusting this threshold for a specific dataset may improve the tradeoff between omission and commission errors depending on the requirements of the application, i.e. cloud or cloud-free conservative applications. The resulting cloud mask is spatially resampled from the native Landsat 8 resolution of 30 m to the corresponding Sentinel-2 resolutions of 10, 20 and 60 m. Throughout the entire process

the work is done at a pixel level, and no spatial dilation of the cloud mask is considered at any stage.

2.2.3 Fmask 4.0 CCA

Function of Mask (Fmask) 4.0 is a cloud assessment algorithm used with Landsat and Sentinel-2 imagery (Qiu et al., 2019). An earlier version, Fmask 3.3, is applied operationally to create cloud masks for USGS Landsat products. The algorithm provisionally identifies cloud pixels using spectral tests, then matches those pixels to provisional cloud shadow pixels using sensor geometry, the Digital Elevation Model (DEM) of the terrain, and an iterative search of altitudes (in Landsat imagery). Fmask was designed to provide a balance between cloud commission and omission errors. Fmask 4.0 is available under an MIT license at https://github.com/GERSL/Fmask.

2.2.4 FORCE

FORCE (Framework for Operational Radiometric Correction for Environmental monitoring, https://github.com/davidfrantz/force) is developed as an 'all-in-one' open-source software solution for the mass-processing and analysis of Landsat and Sentinel-2 image archives (Frantz, 2019). FORCE includes a mono-temporal Level 2 processing system for Analysis Ready Data (ARD) generation which includes: radiometric correction, cloud masking, and data cube generation (Frantz et al., 2016). The cloud masking has branched from Fmask version 1.6.3 (Zhu & Woodcock, 2012), and since then has been developed in parallel (Frantz et al., 2015; Frantz et al., 2016; Frantz et al., 2018). Parts of the updates in Zhu et al. (2015) were incorporated. A darkness filter was implemented to mitigate false positives in bifidly structured dryland areas, where the scene-based temperature distribution tests for Landsat can result in commission errors of cold image parts (Frantz et al., 2015).

Cirrus masking is based on an elevation-dependent equation (Baetens et al., 2019). The most notable difference to the original Fmask, however, is the complete replacement of the cloud probability module for Sentinel-2 with a new algorithm that makes use of the Cloud Displacement Index, which is formulated to enhance parallax effects in highly correlated NIR bands (Frantz et al., 2018). The FORCE cloud masking aims to aggressively detect clouds and cloud shadows to increase cloud producer's accuracy at the deliberate expense of cloud commission for its safe operation in time-series applications. Circular buffers are used to reduce false negatives (300 m for opaque clouds). FORCE provides quality bits whereby 12 quality indicators with respect to atmospheric conditions are provided (Frantz, 2019). Multiple indicators can be set simultaneously for each pixel, e.g., snow and cloud. This quality product is generated at 30 m and 10 m resolution for Landsat and Sentinel-2, respectively. FORCE v. 3.0-dev was used in CMIX.

2.2.5 Idepix

IdePix (Identification of Pixel properties) is a multi-sensor pixel identification tool available as a SNAP (Sentinel Application Platform) plugin (Wevers, 2021). It provides pixel identification algorithms for a wide variety of sensors such as Sentinel-2 MSI, Sentinel-3 OLCI, MERIS, Landsat-8, MODIS, VIIRS, Proba-V or SPOT VGT. IdePix classifies pixels into a series of categories (flags) for further processing using a mono-temporal approach and background information. Its uniqueness consists of a certain set of flags, which are calculated for all instruments (common flags), complemented by instrument specific flags (instrument flags). The technical design of all IdePix is instrument specific and can include decision trees, probabilistic combination of calculated features or neural networks. The Sentinel-2 IdePix is mainly based on a decision tree technique for cloud calculation as well as geometric calculations for cloud and mountain shadows. In contrast to many other pixel identification

tools the final IdePix classification is non-exclusive and therefore allows multiple classes to be set for a single pixel. This means a single pixel can have multiple properties such as land and cloud (semi-transparent cloud over land), land and snow (land covered with snow), or land, snow and cloud (semi-transparent cloud over snow covered land). This type of implementation allows the most versatile usage of the flagging and combinations according to users' needs compared to a standard integer flag allowing a single status per pixel. Sentinel-2 IdePix derives water cloud flags and cirrus cloud flags (B10>0.01 & elevation < 2000 m) on multiple confidence levels, as well as cloud shadow, mountain shadow, snow/ice and water flags. The pixel identification (IdePix) for Sentinel-2 is only working at single resolution (i.e., 10 m, 20 m, 60 m). Cloud boundary pixels are flagged using a dilation filter. In principle, cloud boundaries are regarded as neighbor pixels of a cloud as identified before by the processor; therefore, a buffer is set around the cloud. The width of this boundary (in number of pixels) can be set by the user. Usage of the buffering functionality was not however utilized for CMIX to validate the sole performance of the cloud detection algorithm.

2.2.6 S2cloudless

The s2cloudless is an automated cloud-detection algorithm for Sentinel-2 imagery (Zupanc, 2017) based on a gradient boosting algorithm. It was developed by the EO Research team at Sinergise and is published under the MIT License on https://github.com/sentinel-hub/sentinel2-cloud-detector. The model was trained on a large training dataset with a global coverage. The algorithm is monotemporal, does not consider any spatial context, and therefore can be executed at any resolution. The s2cloudless algorithm can, unlike many other algorithms, be executed also on averaged Sentinel-2 reflectance values over arbitrary user-defined geometries and still provide meaningful results. The input features are Sentinel-2 Level-1C TOA reflectance values of the following ten bands: B01, B02, B04, B05, B08,

B8A, B09, B10, B11, B12 and output of the algorithm is a cloud probability map. Users of the algorithm can convert the cloud probability map to a cloud mask by thresholding the cloud probability map. The recommended value for the threshold is 0.4 to minimize cloud omission errors. Users can optionally apply additional morphological operations during the conversion of the cloud probability map to the cloud mask. These operations are as follows: convolution of the probability map and dilation of the binary cloud mask with a disk. We recommend convolving cloud probability maps at 10 m (160 m) resolution with a disk with a radius of 22 (2) px and dilate cloud masks with a disk with radius 11 (1) px. Sentinel Hub (https://www.sentinel-hub.com, details in EO Research team (2020)) and Google Earth Engine (https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_CLOUD_PROBABILITY) provide precomputed s2cloudless cloud probability maps and masks to their users for the entire Sentinel-2 archive.

The s2cloudless cloud masks for CMIX were provided in a binary mode (1 – cloud and 0 – non-cloud) using the latest (v0.1) model and default values for threshold and morphological operations.

2.2.7 InterSSIM

The InterSSIM cloud detection algorithm is a multi-temporal extension of the s2cloudless algorithm (section 2.2.6), but unlike s2cloudless, the InterSSIM algorithm takes temporal and spatial contexts into account. The algorithm was developed by the EO Research Team at Sinergise (Puc & Žust, 2019) and integrated into the eo-learn Python library published under the MIT License on https://github.com/sentinel-hub/eo-learn. The input data and parameters for the InterSSIM are same as in s2cloudless (see section 2.2.6) with the addition of prior satellite observations. The algorithm works on the ten Sentinel-2 TOA bands, and in addition to cloud probabilities from the s2cloudless model incorporates

additional features: spatially averaged reflectance values, minimum and mean reflectance values over all prior observations, and maximum, mean, and standard deviation of structural similarity indices computed between the observation for which cloud mask is being predicted and every other prior observations. The output of the algorithm is a cloud probability map for the target timeframe, which can be converted into a cloud mask with the same procedure as in the case of the s2cloudless algorithm.

The InterSSIM cloud masks for CMIX were provided in a binary mode (1 – cloud and 0 – non-cloud) using the latest (v0.1) s2cloudless model with default parameter values.

2.2.8 LaSRC

The Land Surface Reflectance Code (LaSRC) is a generic atmospheric correction algorithm aimed at removing atmospheric effects associated with optical satellite imagery acquisitions (Doxani et al., 2018; Vermote et al., 2016). The code is based on the inversion of the 6SV radiative transfer code (Kotchenova et al., 2006; Vermote et al., 1997). Within the atmospheric correction process, LaSRC generates several quality assurance (QA) layers, including a cloud mask. The main metric for deriving a cloud mask is a per-pixel inversion residual error (Skakun et al., 2019; Skakun et al., 2021; Vermote et al., 2016), which shows the goodness of aerosol optical thickness (AOT) estimation process. For both Landsat 8 and Sentinel-2, we used a threshold of 0.05 for the residual to identify cloudy pixels and to minimize cloud omission errors, so only high-quality pixels will be used for further processing. Pixels adjacent to clouds within 5 pixels are separately masked as “adjacent to clouds”. For S2, a conservative threshold of 0.003 (reflectance units) was used for the cirrus band. Therefore, for LaSRC pixels identified as cloud or adjacent were used as “cloud”, whereas all others were used as “non-cloud”. In CMIX, LaSRC version 3.5.5 was used.

498

499

500

501 *2.2.9 MAJA*

502 MAJA is applicable to satellites which perform repetitive observations at similar
503 viewing angles, such as Sentinel-2. It was developed by CNES with methods designed by
504 CESBIO with a few modules provided by DLR. MAJA is an open-source software.

505 MAJA's cloud and shadow detection methods include several tests, which use the
506 multi-spectral and multi-temporal properties of surfaces, clouds, and shadows to classify
507 different types of pixels. The methods are described in Hagolle et al. (2010) and Hagolle et
508 al. (2017). The main cloud test detects the pixels for which the surface reflectance in the blue
509 band increases sharply. The cloud masks obtained with MAJA are dilated by 240 m, firstly to
510 account for the parallax effects due to differences in observation angles between spectral
511 bands, and secondly for the adjacency effects of clouds and for their 'fuzzy' borders. MAJA
512 aims at a sensible reliability for surface reflectance monitoring, its tests and thresholds are
513 therefore optimized to minimize cloud or cloud shadow omission (aiming at maximizing
514 producer's accuracy for clouds, but balanced for cirrus clouds), without excessively
515 degrading the commission error. Cirrus band is used to detect high clouds using the following
516 equation: Cirrus > $0.007 + 0.007 \times h^2$. where h is the pixel altitude in km above sea level.

517 In CMIX, the cloud masks for Sentinel-2 were computed at 240 m resolution to
518 optimize the computation time, but this can prevent MAJA from detecting very small clouds.
519 In the more recent MAJA versions the clouds and shadows masks are computed at 120 m,
520 which should further improve MAJA's performance. MAJA has been intensively validated
521 and some of its validation data sets (Baetens et al., 2019) were used in the CMIX experiment.
522 Due to the necessity to process times series of data with a processed data volume 10 times

greater than the other algorithms, the MAJA team was not able to process all the data sets submitted to CMIX, and it was decided to only produce the datasets acquired when both Sentinel-2A and -2B satellites were operational.

2.2.10 Sen2Cor

Sen2Cor is a processor for Sentinel-2 Level 2A product generation; it performs the atmospheric correction of the Top-Of-Atmosphere (TOA) Level 1C input data. It is composed of two main modules: an atmospheric correction module and a scene classification module that provides a "Scene Classification Map" (SCL), which is used internally in the atmospheric correction module to distinguish between cloudy, clear and water pixels. The Sen2Cor processor is used by the European Space Agency to generate Sentinel-2 Level-2A products within the Sentinel-2 ground segment. Sen2Cor software is available for download at https://step.esa.int/main/third-party-plugins-2/sen2cor/. The code is open source and written in Python.

The Sen2Cor version 2.8 cloud screening algorithm (Louis et al., 2016; Louis, 2021) uses the reflective properties of scene features (TOA reflectance). Potential cloudy pixels undergo a sequence of filtering based on spectral bands thresholds, ratios, and indexes computations (Normalized Difference Snow Index – NDSI, Normalized Difference Vegetation Index –NDVI). Sen2Cor was designed to provide a balance between cloud omission and commission errors. In addition, it includes a cirrus and cloud shadow detection algorithm. A series of additional steps to improve the quality of the classification are automatically triggered using a priori information: digital elevation model (DEM) information, ESA CCI Water Bodies Map v4.0 (Lamarche et al., 2017), ESA CCI Land Cover Map v.2.0.7 (2015) and a snow climatology.

In CMIX, Sen2cor version 2.8 was used. SCL classes 8, 9 and 10 were used for cloud and the remaining SCL classes for non-cloud.

2.3 Performance metrics

A standard set of classification metrics derived from confusion matrices (Table 4) was used to compare cloud masking algorithms and included (Table 5) overall accuracy (OA) and balanced OA (BOA), producer's (PA) and user's accuracies (UA). BOA (Brodersen et al., 2010) was used in addition to OA since some of the reference datasets were imbalanced in terms of cloud/clear pixels and therefore BOA would be a better indicator of algorithms performance.

Table 4. Confusion matrix for cloud validation.

		Reference	
		Cloud	Non-cloud
Map	Cloud	$n_{\text{cloud\_as\_cloud}}$	$n_{\text{ncloud\_as\_cloud}}$
	Non-cloud	$n_{\text{cloud\_as\_ncloud}}$	$n_{\text{ncloud\_as\_ncloud}}$

Table 5. Main performance metrics.

Metric	Equation
Overall accuracy (OA)	$\frac{n_{\text{cloud\_as\_cloud}}+n_{\text{ncloud\_as\_ncloud}}}{n_{\text{cloud\_as\_cloud}}+n_{\text{ncloud\_as\_ncloud}}+n_{\text{ncloud\_as\_cloud}}+n_{\text{cloud\_as\_ncloud}}}$ (1)
Balanced OA (BOA)	$0.5\left(\frac{n_{\text{cloud\_as\_cloud}}}{n_{\text{cloud\_as\_cloud}}+n_{\text{cloud\_as\_ncloud}}}+\frac{n_{\text{ncloud\_as\_ncloud}}}{n_{\text{ncloud\_as\_cloud}}+n_{\text{ncloud\_as\_ncloud}}}\right)$ (2)
PA (for clouds)	$\frac{n_{\text{cloud\_as\_cloud}}}{n_{\text{cloud\_as\_cloud}}+n_{\text{cloud\_as\_ncloud}}}$ (3)
UA (for clouds)	$\frac{n_{\text{cloud\_as\_cloud}}}{n_{\text{cloud\_as\_cloud}}+n_{\text{ncloud\_as\_cloud}}}$ (4)

Performance metrics were estimated from confusion matrices that incorporated all valid pixels over all scenes available in the dataset. PA is complementary to the omission error, which shows a fraction of missed clouds; UA is complementary to the commission error, which shows a fraction of over detected clouds. High PA (cloud-free, non-cloud or clear conservative) means that after elimination of clouds, the users results will be minimally affected by remaining clouds, while high UA (cloud conservative) means that the cloud masks will not discard supernumerary valid pixels.

3 Results

3.1 Performance of cloud masking algorithms for Sentinel-2

3.1.1 CESBIO reference dataset

Table 6 and Figure 9 show performance metrics when applying cloud masking algorithms on the Sentinel-2 CESBIO dataset. Several observations can be made when analyzing these results. The number of reference pixels varied, since the CESBIO dataset was generated at 60 m spatial resolution, and processors produced masks at various spatial resolution: 10 m (FORCE, InterSSIM, LaSRC and S2cloudless), 20 m (ATCOR, Idepix, Fmask 4.0 CCA, Sen2Cor), 60 m (CD-FCNN, interpolated from 30 m), and 240 m (MAJA). Cloud and non-cloud classes were imbalanced in the reference dataset (of all labelled pixels 24.3% were clouds), therefore it results in the OA to be biased towards the non-cloud (dominant) class. Therefore, the balanced OA (BOA) is a more appropriate metric. Overall, BOA varied from 79.5% to 90.5%, an average of 85.9±3.7%. When not considering MAJA (whose developers generated the CESBIO dataset), the highest cloud PA was 85.6%, with the average being 75.9±8.7%, meaning that most algorithms missed almost 24% of clouds identified in the CESBIO dataset. Average cloud UA without MAJA was 85.1±10.6%,

meaning an average of 15% over detection of clouds, which may lie in the dilated parts of the cloud masks (FORCE, MAJA), or be associated with a stricter detection of cirrus clouds (LaSRC). Overall, the performance of cloud masking algorithms varied for this dataset by an average 11-12% of PA and UA, as measured by the coefficient of variation (CV), which is a ratio between standard deviation and average.

Table 6. Performance metrics of Sentinel-2 cloud masking algorithms for the CESBIO dataset. All algorithms, except MAJA, processed all 30 reference scenes (with 24.3% of clouds in the reference dataset), while MAJA processed 28 references scenes (25.6%). Here, and in Table 7 through Table 14: in bold are the numbers with the highest value for the particular metric (column-wise); * denotes algorithms which did not process the whole dataset; algorithms that are underscored were produced by the same team as the reference dataset.

			Cloud	
Processor	**OA**	**BOA**	**PA**	**UA**
ATCOR	88.6	80.4	64.4	84.9
CD-FCNN	89.5	79.5	60.3	**94.1**
Fmask 4.0 CCA	**93.3**	**88.9**	80.4	90.8
FORCE	91.1	**88.9**	84.7	79.9
Idepix	91.7	86.9	77.5	86.9
InterSSIM	93.2	88.0	77.8	93.1
LaSRC	81.2	82.7	**85.6**	57.6
MAJA* (28/30)	89.2	**90.5**	**92.9**	72.7
S2cloudless	93.1	88.8	80.4	90.2
sen2cor	91	84.7	72.3	88.7
Average	*90.2*	*85.9*	*77.6*	*83.9*
Standard deviation	*3.4*	*3.7*	*9.3*	*10.7*

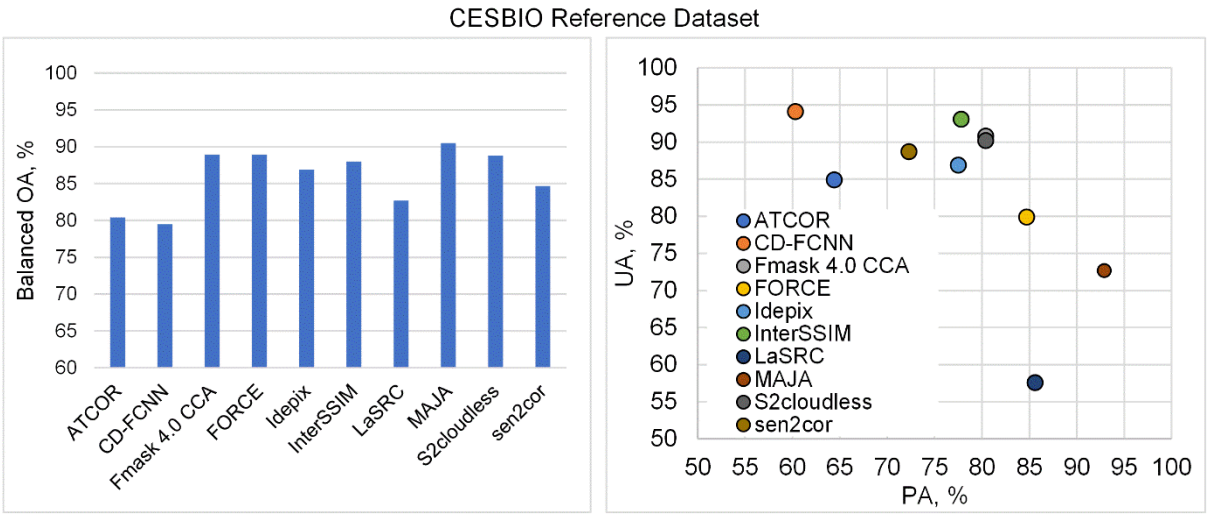


Figure 9. Comparison of BOA values and distribution of PA/UA for the CESBIO reference dataset.

3.1.2 GSFC S2 reference dataset

Table 7 and Figure 10 show the results of comparing algorithm outcomes against the Sentinel-2 GSFC dataset. MAJA provided only 10 images out of 28 images. In the S2 GSFC dataset, cloud and non-cloud are almost balanced (approx. 61% of reference pixels are identified as clouds), therefore there is minimal difference between OA and BOA. BOA varied from 80.7% to 96.8% with LaSRC being the outlier (developers of LaSRC produced the GSFC data), with average being 85.7±2.8% (not considering LaSRC). Average values of cloud PA and UA not considering LaSRC were 73.7±5.6% and 98.2±2.7%, respectively, meaning large omission errors. It is worth noting that FORCE and MAJA, whose PA was better than the UA for the other reference datasets, have the opposite result for the GSFC reference, due to the strict classification of very thin clouds as clouds in the GSFC reference data set. The reason for all algorithms producing lower accuracies compared to LaSRC is that they did not identify thin (semi-transparent and cirrus) clouds, which, in turn, LaSRC was masking out using a rather conservative threshold (0.003 in reflectance units; for LaSRCv3.5.5) applied for the cirrus band (B10). As the cirrus cloud masking method is very

simple, all methods could obtain similar performances, at the expense of masking an important part of usable pixels. Those clouds were labelled as thin, since they were clearly visible in the ground-based images. If thin clouds are removed from the analysis (Table 7), all algorithms showed much better performance: average BOA was 94.4±2.9% (an average gain +7.4±2.6%) and cloud PA was 90.8±5.9% (an average gain +14.8±5.2%), while cloud-UA remained essentially the same 98.1±2.7%. These results show the differences between algorithms in defining and identifying thin (semi-transparent) cirrus clouds, at the same time mostly agreeing on thick clouds. Variation in algorithms performance was 8% for cloud PA (6% without thin clouds) and 3% for cloud UA.

Table 7. Performance metrics of Sentinel-2 cloud masking algorithms for the GSFC S2 dataset. All algorithms, with exception of MAJA, processed all 28 reference scenes (with 60.6% and 55.5% of clouds in reference data for all clouds and without thin clouds, respectively), while MAJA processed 10 images (49.2% and 40.8%).

	All types of clouds				**Without thin clouds**			
			Cloud				**Cloud**	
Processor	**OA**	**BOA**	**PA**	**UA**	**OA**	**BOA**	**PA**	**UA**
ATCOR	77.9	81.7	63.5	**100**	86.9	88.2	76.4	**100**
CD-FCNN	82.4	85.4	71	99.9	92.9	93.6	87.3	99.9
Fmask 4.0 CCA	86	88.4	77.1	99.7	96.1	96.5	93.3	99.7
FORCE	86.1	88.2	78.2	98.6	95.9	96.1	94	98.5
Idepix	84.8	86.1	80.1	93.9	92.5	92.5	92.9	93.6
InterSSIM	85	87.6	75.4	99.7	95.6	96	92.4	99.7
LaSRC	**96.7**	**96.8**	**96.3**	98.2	**98**	**97.9**	**98.5**	97.8
MAJA* (10/28)	80.9	80.7	66.2	93	92.7	92.2	89.1	92.7
S2cloudless	85.2	87.7	76.1	99.3	95.7	96.1	93	99.3
sen2cor	85.2	87.8	75.8	99.7	95	95.4	91.2	99.7
Average	*85.0*	*87.0*	*76.0*	*98.2*	*94.1*	*94.5*	*90.8*	*98.1*
Standard deviation	*4.6*	*4.1*	*8.4*	*2.4*	*2.9*	*2.7*	*5.6*	*2.6*

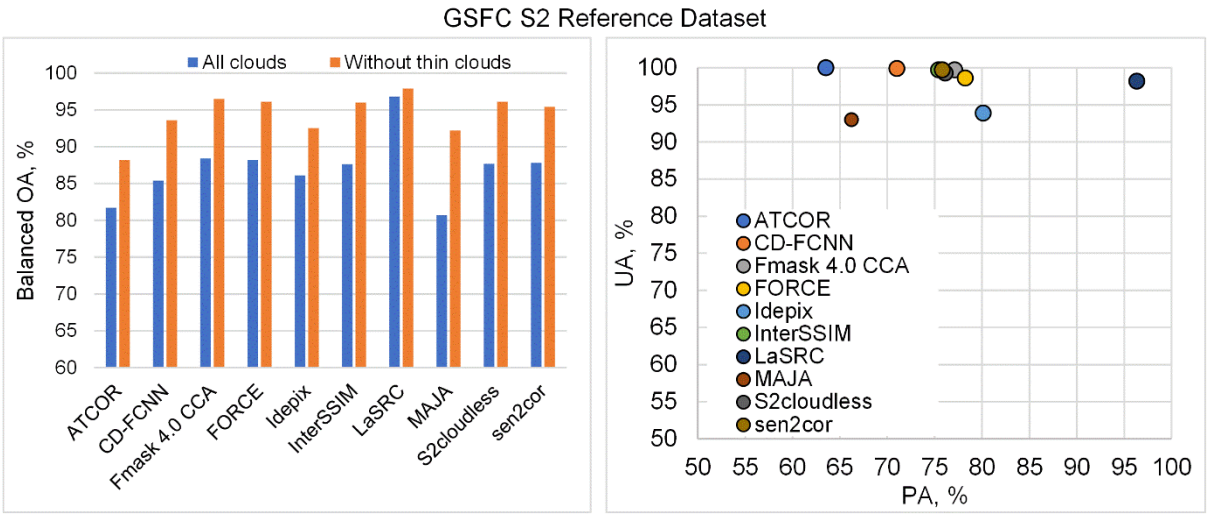


Figure 10. Comparison of BOA values and distribution of PA/UA (for all clouds) for the GSFC S2 reference dataset.

3.1.3 Hollstein reference dataset

Table 8 and Figure 11 show algorithms performance for the Hollstein data depending on the opaque and semi-transparent/cirrus clouds. BOA varied from 84.2% to 92.3% (average 89.4±2.4%) for all cloud types and 86.2 to 97.8% (93.4±3.8%) for opaque clouds only. Not considering semi-transparent/cirrus clouds improved algorithms performance, especially for cloud PA: an average gain +8.0±8.1%. Variation of performance was comparable to the GSFC results with 8% (5% for opaque only) for PA and 4% (7%) for UA. Note that the Hollstein dataset was used to set radii of disks with which the cloud probability mask and binary cloud mask are convoluted and dilated, respectively, by the s2cloudless algorithm. MAJA was not evaluated against the Hollstein data set, as the images were acquired before Sentinel-2B launch.

Table 8. Performance metrics of cloud masking algorithms for the Hollstein dataset. All algorithms processed all 59 reference scenes (with 61.8% and 44.4% of clouds in reference data for all clouds and without thin clouds, respectively).

	Opaque clouds and semi-transparent clouds/cirrus				Opaque clouds only			
			Cloud				Cloud	
Processor	**OA**	**BOA**	**PA**	**UA**	**OA**	**BOA**	**PA**	**UA**
ATCOR	88.6	89.9	84.6	96.5	89.1	88.5	81.8	93.2
CD-FCNN	81	84.2	71.1	97.7	**97.8**	**97.8**	98.3	96.7
Fmask 4.0 CCA	91.2	91.1	91.3	94.2	94.9	95.4	**99.9**	89.8
FORCE	89.1	89.4	88.2	93.8	93.6	94	97.4	89.1
Idepix	91.3	90.5	94.1	92.1	91.9	92.6	98.2	85.7
InterSSIM	90.4	91.9	85.7	**98.6**	97.5	97.4	96.8	**97.5**
LaSRC	89.3	86.7	**97.7**	86.7	85	86.2	96.7	76
S2cloudless	**91.5**	**92.3**	89.2	96.8	96.3	96.5	97.6	94.3
sen2cor	87.9	88.6	85.6	94.3	92.2	92.3	93	89.8
Average	*88.9*	*89.4*	*87.5*	*94.5*	*93.1*	*93.4*	*95.5*	*90.2*
Standard deviation	*3.1*	*2.4*	*7.1*	*3.4*	*3.9*	*3.8*	*5.2*	*6.2*

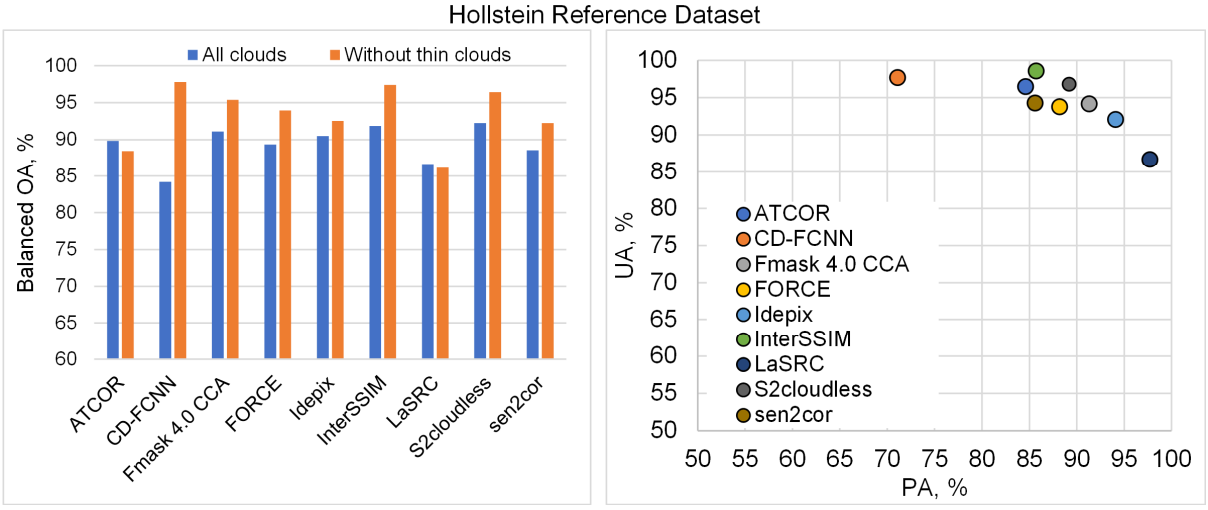


Figure 11. Comparison of BOA values and distribution of PA/UA (for all clouds) for the Hollstein reference dataset.

3.1.4 PixBox S2 reference dataset

Not all algorithms processed all 29 products of the PixBox S2 dataset; the reasons for this were limitations of allowed geometries (ATCOR, 27 processed) or too sparse time-series around the acquisition (MAJA, 14 processed). In order to account for the difference of available products for validation, two different comparisons were made: one using all

664 available products for each algorithm and a second using only the products that all algorithms
665 have been applied to (14 out of 29 reference scenes). We call the second dataset the least
666 common denominator (LCD) subset, while the first is referred to as the "complete dataset".
667 The whole comparison could have been made only on the LCD subset, but this reduces the
668 complete dataset by half, which reduces its utility. Therefore, the complete dataset also was
669 used for comparison. In this comparison using the complete dataset, results for MAJA must
670 be assessed with caution, as they are only based on 14 out of 29 products.

671 Algorithm performance for the complete PixBox dataset is provided in Table 9 and
672 Figure 12. For all types of clouds, BOA varied from 67.5% to 85.9% (average 80.0±5.3%).
673 The top two algorithms (S2cloudless and MAJA) showed a similar performance in terms of
674 BOA; however, the tradeoff between PA and UA varied substantially for those algorithms:
675 S2cloudless yielded PA=80.2% and UA=89.5% (more cloud omissions than commissions)
676 and MAJA yielded PA=88.6% and UA=80.2% (less cloud omissions and more commissions,
677 in part due to the dilation). When thin/semi-transparent clouds were not considered, all
678 algorithms showed a better performance with an average gain in BOA of +5.1±1.6%. Some
679 algorithms (FORCE, Idepix and LaSRC) showed high commission errors (low UA), which
680 were related to identifying snow as clouds.

681

682 Table 9. Performance metrics of cloud masking algorithms for the complete PixBox S2
683 dataset. ATCOR and MAJA processed 27 and 14 reference scenes, respectively, while other
684 algorithms processed all 29 reference scenes. Fraction of cloud pixels was 47.2% and 36.8%
685 for all cloud types and without thin clouds, respectively.

	All types of clouds				**Without thin clouds**			
			Cloud				**Cloud**	
Processor	**OA**	**BOA**	**PA**	**UA**	**OA**	**BOA**	**PA**	**UA**
ATCOR* (27/29)	76.6	76.2	62.5	85.3	82.5	80.4	70.8	81.4
CD-FCNN	80.5	79.7	66	89.9	89.5	88.1	82.7	87.9
Fmask 4.0 CCA	84.5	84.2	79.4	86.5	89.6	89.9	90.8	82.7

FORCE	80.2	80.1	79	78.9	84.6	85.8	90.4	73.6
Idepix	75.7	76.3	85.9	69.7	77.2	81	**95.3**	62.4
InterSSIM	84.6	84	72.7	**93.2**	**91.9**	90.7	86.2	**91.3**
LaSRC	66.4	67.5	**86.8**	59.9	65	71	93.8	51.3
MAJA* (14/29)	85.1	85.5	**88.6**	80.2	86.5	88.3	94.3	74.3
S2cloudless	**86.3**	**85.9**	80.2	89.5	91.6	**91.6**	91.6	86.4
sen2cor	81.2	80.8	74.7	83.6	85.4	84.8	82.7	78.6
Average	*80.1*	*80.0*	*77.6*	*81.7*	*84.4*	*85.2*	*87.9*	*77.0*
Standard deviation	*5.7*	*5.3*	*8.3*	*9.6*	*7.7*	*6.0*	*7.1*	*11.7*

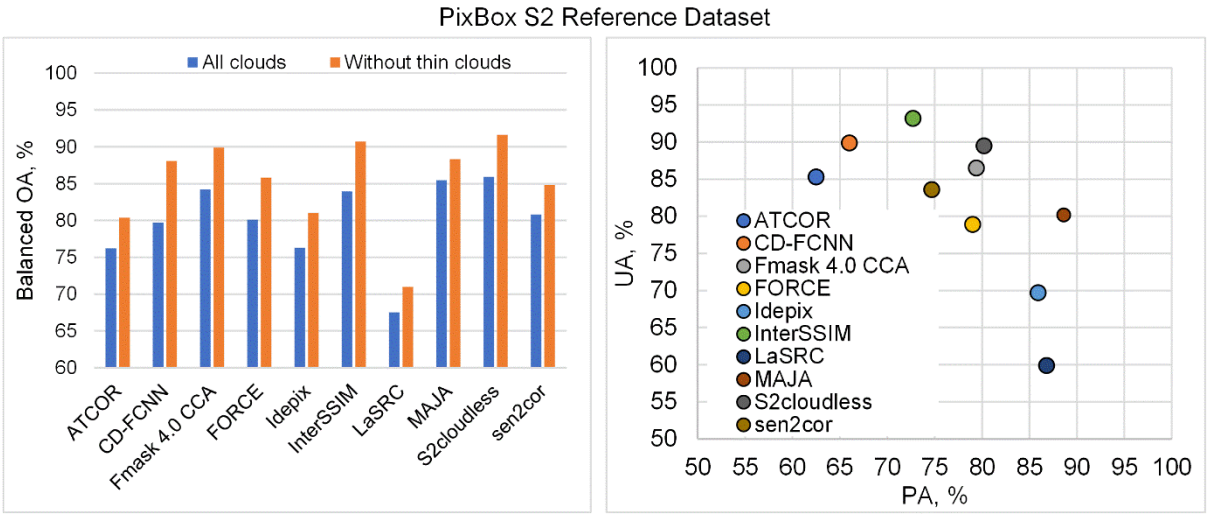


Figure 12. Comparison of BOA values and distribution of PA/UA (for all clouds) for the PixBox S2 reference dataset.

Table 10 shows BOA values when comparing complete and LCD PixBox dataset. When restricting to the LCD, s2cloudless yielded the highest BOA in all cases Overall, the differences in BOA between complete and LCD sets were below 2%. Also, algorithms performance improved when thin clouds and snow were excluded from the analysis.

Table 10. Performance metrics of cloud masking algorithms for the complete and LCD PixBox dataset for various scenarios.

	All types of clouds		**All types of clouds (excluding snow)**		**Without thin clouds**	
Processor	BOA	BOA	BOA	BOA	BOA	BOA

	complete	LCD	complete	LCD	complete	LCD
ATCOR	76.2	78.3	77.2	79.3	80.4	81.6
CD-FCNN	79.7	78.6	80.4	79.5	88.1	86.0
Fmask 4.0 CCA	84.2	85.1	86.3	86.9	89.9	89.7
FORCE	80.1	83.0	82.1	85.2	85.8	88.2
Idepix	76.3	73.8	84.0	83.0	81.0	78.8
InterSSIM	84.0	84.2	84.9	85.2	90.7	91.1
LaSRC	67.5	70.7	74.2	78.0	71.0	73.4
MAJA	85.5	85.5	86.1	86.1	88.3	88.3
S2cloudless	**85.9**	**87.3**	**86.7**	**87.8**	**91.6**	**93.1**
sen2cor	80.8	82.3	82.1	85.4	84.8	85.3
Average	*80.0*	*80.9*	*82.4*	*83.6*	*85.2*	*85.5*
Standard deviation	*5.3*	*5.1*	*3.9*	*3.3*	*6.0*	*5.7*

Figure 13 shows an example of cloud detection over the Sentinel-2 scene from the PixBox dataset. The scene features opaque clouds as well as semi-transparent clouds over the water. All algorithms were successful in identifying opaque clouds, while majority struggled to identify semi-transparent over the water.

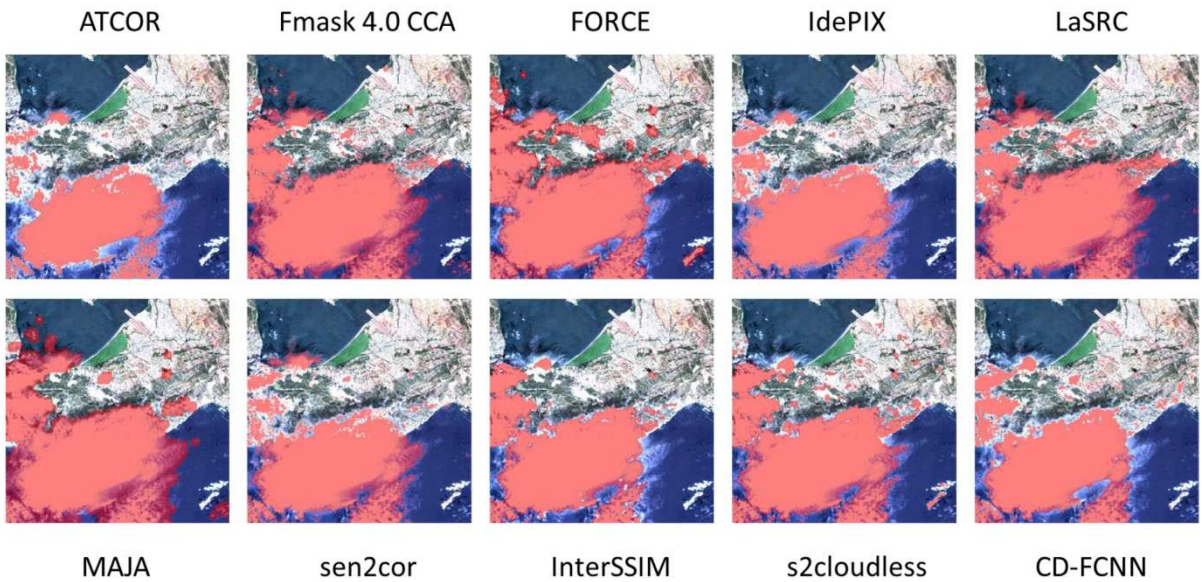


Figure 13. Examples of cloud masking by various algorithms over the Sentinel-2 scene S2A_MSIL1C_20170629T103021_N0205_R108_T31TFJ_20170629T103020.

Figure 14 shows performance of algorithms on clear pixels depending on the major land cover classes (proportion>4%) from the PixBox S2 data. LaSRC, IdePix and FORCE showed the worst performance for the clear snow pixels, which was expected given limitations of these algorithms. Excluding snow, overall performance of algorithms was uniform throughout the land clover classes. All algorithms showed worst performance for the urban area given the presence of bright targets. Even approaches utilizing the Sentinel-2 multi-band parallax (e.g., FORCE, Frantz et al., 2018) over-detected clouds in the urban areas.

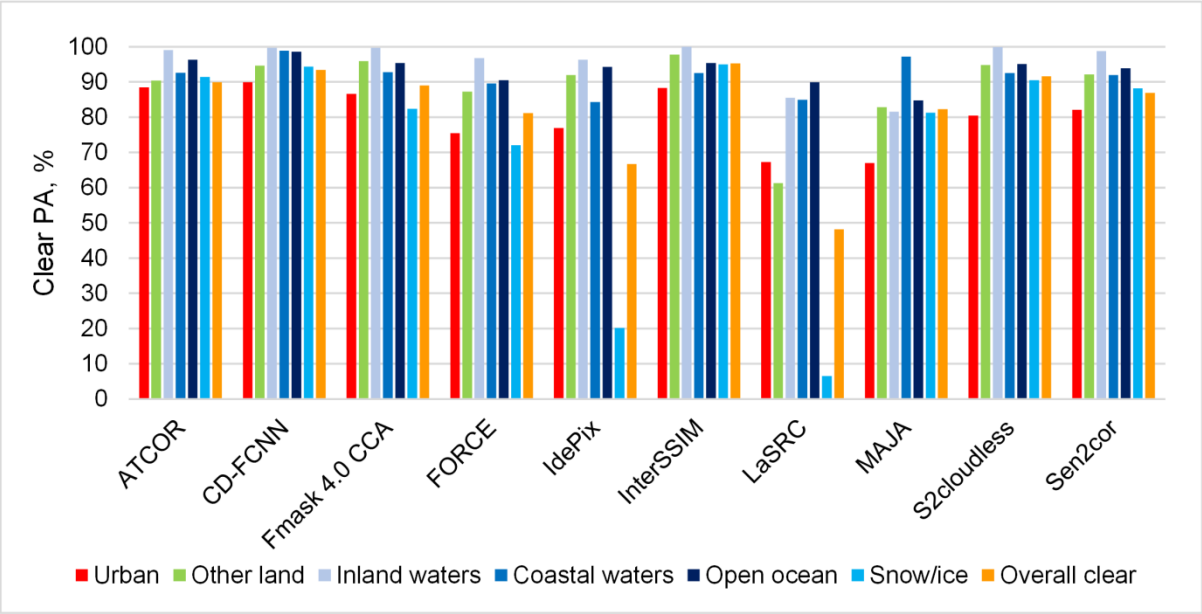


Figure 14. Performance of algorithms in terms of clear producer's accuracy over the non-cloudy regions depending on the land cover types in the PixBox S2 dataset.

3.2 Performance of cloud masking algorithms for Landsat 8

3.2.1 GSFC L8 reference dataset

This dataset included six Landsat 8 scenes and all algorithms showed high performance (Table 11). Fmask showed the highest values of performance metrics. Two algorithms achieved 100% cloud UA, meaning no cloud overdetection in this dataset.

Table 11. Performance metrics of cloud masking algorithms for the GSFC L8 dataset. All algorithms processed six reference scenes (with 49.4% fraction of cloud in reference data).

			Cloud	
Processor	**OA**	**BOA**	**PA**	**UA**
ATCOR	97.3	97.3	94.8	99.8
CD-FCNN	97.3	97.3	94.6	**100.0**
Fmask 4.0 CCA	**98.7**	**98.7**	**97.3**	**100.0**
FORCE	98.2	98.1	96.5	99.7
LaSRC	96.5	96.5	94.8	98.0
Average	*97.6*	*97.6*	*95.6*	*99.5*
Standard deviation	*0.8*	*0.8*	*1.1*	*0.7*

3.2.2 L8Biome reference dataset

Table 12 provides a summary of performance metrics for the L8Biome dataset. Results in this table should not be used directly for intercomparing algorithms for the following reasons: (i) ATCOR processed only 86 images out of 96 images, since images in polar regions were removed due to Sun elevation lower than 30°; (ii) LaSRC processed 80 images, since snow/ice scenes were not considered; (iii) all algorithms, except ATCOR, had on average 2.4% pixels not classified—those pixels are on the boundary of the Landsat 8 scene, and do not have valid values for all spectral bands. In addition, since CD-FCNN was trained on the L8Biome and the L8SPARCS datasets (80% and 20%, respectively), the CD-FCNN results on this dataset are omitted in order to avoid overoptimistic (overfitted) detection results. Fmask partially used L8Biome data to find optimal thresholds for some of the rules, namely weight of cirrus cloud probability, spectral-contextual snow index, and morphology-based post-processing (Qiu et al., 2019; personal communication, Zhe Zhu and Shi Qiu, University of Connecticut, November 2021). Since the foundation of the Fmask algorithm was developed well before the L8Biome dataset release, we still included Fmask 4.0 for the inter-comparison, though with caveats.

Table 12. Performance metrics of cloud masking algorithms for the L8Biome dataset. ATCOR and LaSRC processed 86 (48.3% of clouds in reference data) and 80 (49.4%) scenes, respectively, while Fmask and FORCE processed all 96 scenes (47.9%).

			Cloud	
Processor	**OA**	**BOA**	**PA**	**UA**
ATCOR* (86/96)	86.8	86.7	83.2	88.8
Fmask 4.0 CCA	90.0	90.2	93.6	86.6
FORCE	84.9	85.3	**96.0**	77.7
LaSRC* (80/96)	**90.9**	**90.9**	92.7	**89.2**
Average	*88.1*	*88.3*	*91.4*	*85.6*
Standard deviation	*2.4*	*2.3*	*4.9*	*4.7*

Table 13 provides a correct intercomparison between algorithms since the amount of reference scenes and pixels used was the same. The average BOA was 90.0±1.4% and 91.5±1.8% for all types of clouds and without thin clouds, respectively. Removing thin clouds from the reference increases BOA and Cloud-PA accuracies by +1.5±0.7% and +3.0±1.4%, respectively.

Table 13. Performance metrics of cloud masking algorithms for the L8Biome dataset using the same set of 80 Landsat 8 scenes. Fraction of cloud reference pixels for all types of clouds and without thin clouds was 49.4% and 42.6%, respectively.

	All types of clouds				**Without thin clouds**			
			Cloud				**Cloud**	
Processor	**OA**	**BOA**	**PA**	**UA**	**OA**	**BOA**	**PA**	**UA**
ATCOR	88.2	88.2	84.6	**90.9**	89.6	89.2	86.8	**88.6**
Fmask 4.0 CCA	**91.3**	**91.4**	96.2	87.4	92.1	93.1	**99.7**	84.6
FORCE	89.4	89.5	**96.8**	84.2	89.0	90.2	98.1	80.4
LaSRC	90.9	90.9	92.7	89.2	**92.8**	**93.5**	97.8	86.9
Average	*89.9*	*90.0*	*92.6*	*87.9*	*90.9*	*91.5*	*95.6*	*85.1*
Standard deviation	*1.2*	*1.3*	*4.9*	*2.5*	*1.6*	*1.8*	*5.1*	*3.1*

Analysis of algorithms performance by biomes showed little variability (Figure 15). Exceptions are ATCOR which showed lower cloud PA values over forest and grass/cropland biomes, and Fmask which lower cloud PA values over shrubland. It is worth noting though that those are generic land cover classes and don't enable analysis of the dynamic state of the land cover class during the scene overpass. For example, a cropland can be characterized by multiple physical stages during the year, such as bare land (e.g., fallow or after ploughing), sparse vegetation (during crop emergence), dense vegetation (during peak), snow (during the winter period). Therefore, per-land cover performance of algorithms should be taken cautiously.

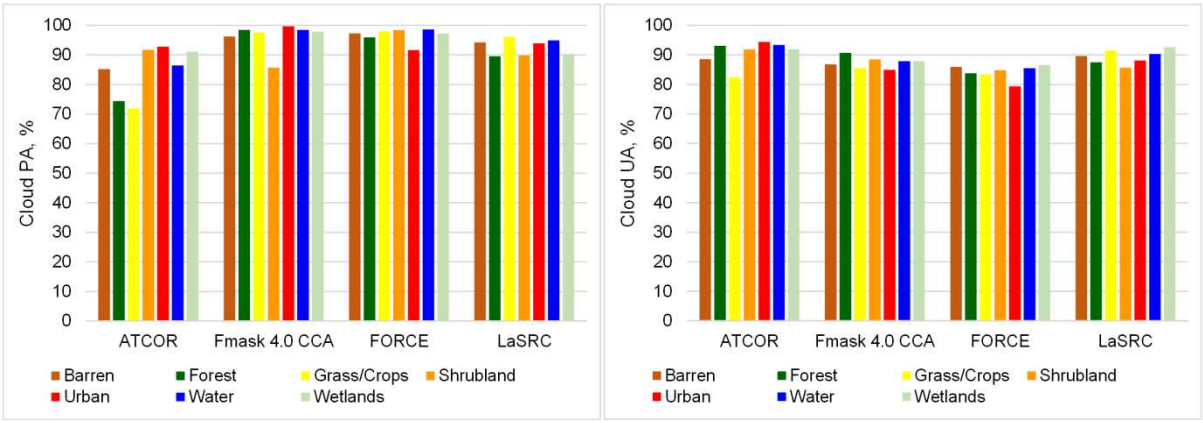


Figure 15. Performance of the Landsat 8 cloud detection algorithms for the L8Biome dataset depending on the biomes. The same set of 80 Landsat 8 scenes was used to calculate PA and UA accuracy values.

3.2.3 PixBox L8 reference dataset

Table 14 shows the algorithm performance for the PixBox dataset. Fmask and ATCOR yielded the best performance in terms of BOA (87.9% and 86.3%, respectively), however PA/UA values exhibited a different behavior: for Fmask, PA and UA were mostly balanced (82.5% and 81.8%), while for ATCOR omission error (26.7%) was much higher

than commission error (2.8%). Overall, performance over the PixBox dataset was lower than for L8Biome and GSFC, as the case with PixBox S2. Performance metrics substantially improved when semi-transparent clouds were removed from the analysis. For all algorithms cloud PA increased on average by 28.1±13.9% reaching 95.9±3.6%. While there was an overall agreement between algorithms on detecting opaque clouds from the PixBox L8 dataset (with average PA 95.9±3.6%) all algorithms failed to detect semi-transparent clouds (average PA was 40.6±27.4%) (Figure 16). It's worth noting that all algorithms showed equally good performance for clear land and water classes. ATCOR and CD-FCNN were also successful in discriminating clouds from snow, while Fmask and FORCE showed intermediate results. LaSRC failed to identify clouds over snow, as expected from the algorithm's design.

Table 14. Performance metrics of cloud masking algorithms for the PixBox dataset. All algorithms processed all 11 Landsat 8 reference scenes. Fraction of cloud reference pixels was 27.4% for all types of clouds and 15.8%, when removing semi-transparent clouds.

	All types of clouds				*Without semi-transparent clouds*			
			Cloud				**Cloud**	
Processor	**OA**	**BOA**	**PA**	**UA**	**OA**	**BOA**	**PA**	**UA**
ATCOR	**92.1**	86.3	73.3	**97.2**	**98.4**	96.7	94.1	**95.6**
CD-FCNN	87.2	78.2	59	89.4	97.8	**98.7**	**99.9**	87.4
Fmask 4.0 CCA	90.4	**87.9**	**82.5**	81.8	94.3	96.6	99.8	72.6
FORCE	80.3	79.1	76.5	61.3	83.5	87.2	92.8	48.7
LaSRC	76.8	67.8	47.8	59.5	88.5	90.4	93.1	58.6
Average	*83.7*	*78.2*	*66.5*	*73.0*	*92.5*	*93.9*	*95.9*	*72.6*
Standard deviation	*5.4*	*7.1*	*13.8*	*12.9*	*6.4*	*4.8*	*3.6*	*19.5*

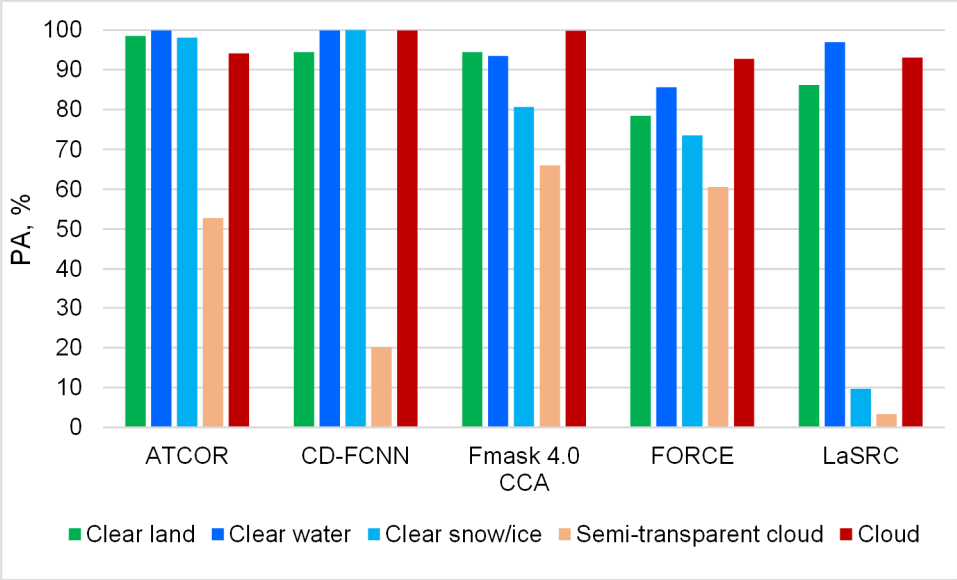


Figure 16. PA values for various types of classes in the PixBox L8 dataset.

4 Discussion

4.1 Algorithm intercomparison

Figure 17 shows the distribution of cloud PA and UA values for Sentinel-2 cloud masking algorithms. Overall, cloud PA/UA values are located in the areas defined by lines PA>80% or UA>80%. While individual values are located in the area of PA>90% and UA>90% (Figure 17, left), suggesting a very good balance of commission and omission errors, however that is not the case for averaged values across all reference datasets (Figure 17, right). No algorithm yielded the PA>90% and UA>90% performance when averaging over reference datasets. Five algorithms (Fmask, FORCE, Idepix, MAJA and S2cloudless) yielded the average performance of cloud PA>80% and UA>80%, providing some balance (within ~10%) between commission and omission errors. Four algorithms (ATCOR, CD-FCNN, InterSSIM and sen2cor) yielded performance with cloud UA>90% (cloud conservative), meaning these algorithms committed less clouds over clear regions, however at the expense of missing clouds. LaSRC yielded the cloud PA>90% performance (non-cloud conservative), detecting most of the clouds, however, at the expense of masking out also

valid non-cloudy observations, and with a large standard deviation in UA across the datasets (potentially, owing to various rules defining the cloud and the use of conservative threshold for the cirrus band).

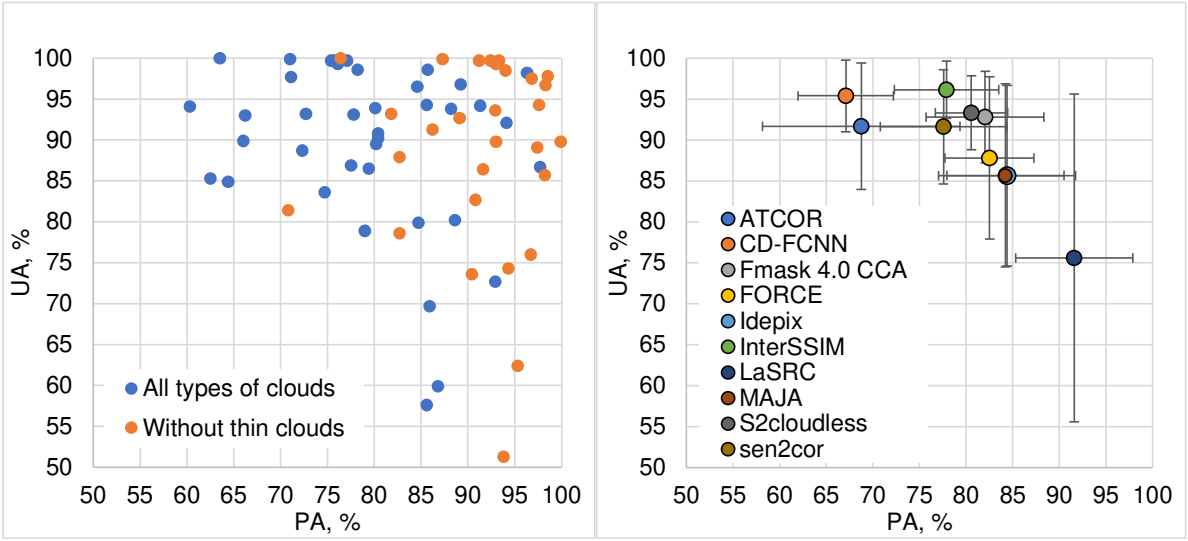


Figure 17. Distribution of cloud PA and UA over all Sentinel-2 cloud masking algorithms and reference datasets (left) and algorithms' average values along with the standard deviation over four reference datasets (right). Averaging was performed using PA and UA values from Table 6, Table 7, Table 8 and Table 9 for all cloud types.

Since only three datasets were used for Landsat 8, we did not perform the averaging (Figure 18). Three distinct clusters corresponding to the three reference datasets were evident with varying performance. The highest performance was for the GSFC dataset with only six Landsat 8 scenes over the same area, which probably is not fully representative of the performance of the algorithms. GSFC L8 had mostly thick and well-identifiable clouds that algorithms were able to classify successfully. L8Biome yielded the second highest performance with PA/UA values distributed over PA>90% (Fmask, FORCE and LaSRC) and UA>90% (ATCOR). Performance for the PixBox dataset was the lowest with algorithms

scattered in the cloud PA/UA space. Fmask yielded PA>80% and UA>80% for PixBox; ATCOR and CD-FCNN yielded UA>90%; while FORCE and LaSRC yielded both cloud PA and UA less than 80%.

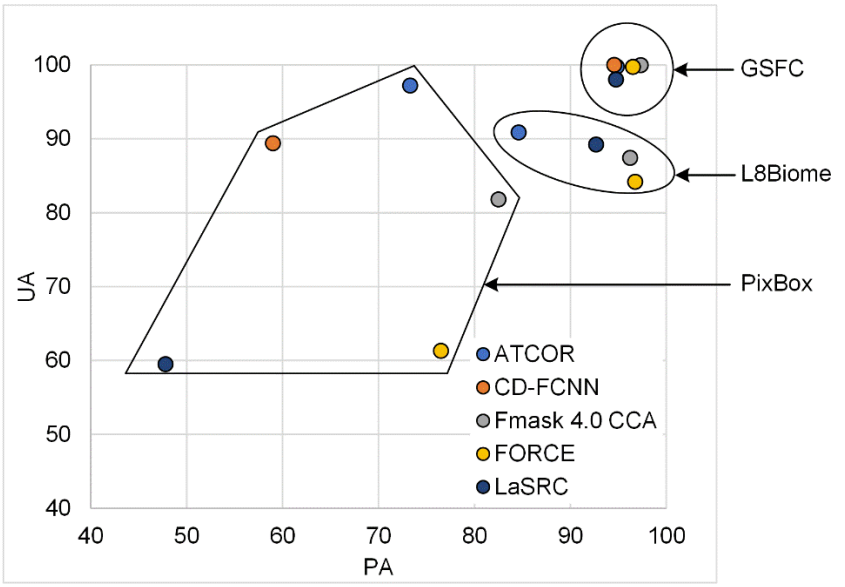


Figure 18. Distribution of cloud PA and UA over all Landsat 8 clouds masking algorithms and reference datasets.

A summary of strengths and weaknesses of cloud algorithms known at the design stage and further identified/elaborated during the CMIX are presented in Table 15.

Table 15. Summary of algorithms parameters to control cloud commission/omission errors along with strengths and weaknesses.

Processor	Parameter	Strengths	Weaknesses
ATCOR	Cloud buffer size (default size is 7 px). Increase will lead to higher cloud PA.	– Water vapor map (S2) is used to reduce cloud pixel commission error – Elevation-dependent cirrus masking	– Conservative cloud mask – Cloud buffer too small – Thin cirrus threshold of ρ(TOA)=0.01 underestimates thin cirrus
CD-FCNN	A posteriori cloud probability (default value is 0.5). Decrease will lead to	– Single architecture to provide global cloud masks for both Landsat-8	– Model can underperform compared to customized algorithms for Sentinel-2

	higher cloud PA (cloud-free conservative). Increase will lead to higher cloud UA (cloud conservative).	and Sentinel-2 images – No ancillary data required – Mitigation of training data requirements: transfer learning from Landsat-8 to Sentinel-2 – General approach directly learnt from available data	– Model performance is fully constrained by the quality of training data – Presence of errors in thin clouds, cloud borders, urban areas, and snow. – It does not provide shadow detection. – It does not provide cloud type classes (e.g. cirrus, thin or thick clouds).
Fmask 4.0 CCA	Cloud dilation (default is 3 px), cloud probability threshold (CPT), and potential false positive cloud (PFPC) extension and erosion. The CPT default value is 17.5% for Landsat 8, and 20% for Sentinel 2. Increase will reduce the number of potential cloud pixels. The PFPC parameters affect how the potential cloud mask is reduced to the final cloud mask. Changing its values will affect the algorithm's performance over bright targets.	– Generic algorithm – Applicable over land and water – Good performance over bright targets (urban, ice/snow)	– Performance decreases when thermal band is not used
FORCE	Cloud probability (default 22.5%). Increase will reduce the number of potential cloud pixels. Clouds were buffered by 300 m. Higher values will increase cloud commission but reduce commission.	– Rigorous cloud mask with emphasis on reducing cloud commission for safe usage in time series applications – Parallax effect is used to reduce bright false positives in Sentinel-2 imagery – Multiple flags can be set, e.g. snow and cloud	– Rigorous cloud mask with emphasis on reducing cloud commission with potential drawbacks for single-scene analysis – Parallax effect may occasionally introduce false positives in bright areas due to micro-vibrations on sensor – Snow and cloud often not mutually exclusively
IdePix	The CLOUD_AMBIGUOUS flag is currently quite probe to clear commission of urban and other very bright surfaces. Cloud buffer was not used, as it would increase cloud commission error.	– Mono-temporal approach – Detects thin clouds quite well – Allows user defined cloud dilation	– Snow detection could be better (bug in code during CMIX) – Commission error of bright (mostly urban) surfaces
s2cloudless	Cloud probability (default is 0.4). Lower values will lead to higher cloud PA (cloud-free conservative). Post-processing:	– Fast single-observation cloud masking – Works on any resolution and even on aggregated values (objects)	– Prone to errors on very bright areas – No spatial context is taken into account – No cloud shadow detection

	convolution (22 px) and dilation (11 px). The convolution smoothens the masks, reducing the amount of salt-and-pepper effect, while the dilation of masks closes small openings and increases the cloud masks on the outside.	– Provides pseudo-probability that user can tweak to get better cloud masks for her use-case	
InterSSIM	Similar to s2cloudless. Number of prior satellite observations. Increase will lead to better performance, especially bright targets, but increase the usage of computational and storage resources.	– Using spatio-temporal context results in lower rate of false positive detections (particularly over consistently bright areas) – Provides pseudo-probability that user can tweak to get better cloud masks for her use-case	– Resource intensive calculation – Higher rate of cirrus misclassifications – Higher rate of misclassifications over large waterbodies – No cloud shadow detection
LaSRC	Threshold for residuals from aerosol retrievals (default is 0.05). Increase will lead to higher cloud UA (cloud conservative).	– Simple, interpretable criteria Easily transferable – Conservative and tune to keep best high-quality data rather than questionable (low-quality)	– Might confuse bad retrievals of aerosol with clouds (high aerosol, urban area) – Not suitable over snow cover region
MAJA	Four major parameters: – Multi-temporal: threshold on increase of surface reflectance in the blue. – Correlation: each neighborhood of a cloud is correlated with previous observations. If the correlation is high, it is not a cloud. – High clouds: threshold for the reflectance of the cirrus band, that depends on the squared altitude of the pixel to account for the fact that mountains may peak above the water vapor layer. – Buffer: all pixels close to a cloud within a buffer of 240 m are classified as clouds, which is rather conservative, and avoids omissions due to the parallax between spectral bands or to fuzzy limits of the cloud.	– Multi-temporal criterion to better detect low clouds that brings a much better separation between cloud / non clouds – Moderate threshold for the cirrus bands, as the multi-temporal threshold already detects clouds which have a significant impact on reflectances – Large buffer (240m), possible thanks to the very low level of cloud commission errors before dilation	– Some very rapid changes of vegetation could be interpreted as clouds – Multi-temporal algorithm is less efficient in places where the cloudiness is extremely high – Working at 120 m resolution (240 m resolution during CMIX, but it has been upgraded since), may cause omissions of very small clouds – The buffer will include some cloud free pixels (but they are in fact are affected by large adjacency effects)
sen2cor	The parameters used to run Sen2Cor version 2.8 for	– Cloud mask at "moderate" resolution (20	– Potential cloud omissions on cloud edges/boundaries

	CMIX were the default parameters used in Sentinel-2 operational ground segment and available in L2A_CAL_SC_GIPP.xml. No cloud mask dilation is applied and cloud boundaries can be omitted.	m) – Robustness. Used operationally in all types of meteorological conditions and solar geometries – Processing time (<5 min for a full Sentinel-2 tile)	– Potential cloud omissions for cloud over water – Potential cloud commissions for bright buildings in urban area or bright surfaces

4.2 Dependence of the performance on the reference datasets

Performance of cloud masking algorithms for Sentinel-2 varied depending on the reference dataset (Figure 19): average BOA was 80.0±5.3% (PixBox) to 89.4±2.4% (Hollstein). Performance of algorithms was the worst for the PixBox dataset compared to datasets. This can be explained by the following. PixBox dataset was sampled in such a way, so non-challenging (e.g., opaque thick clouds) and challenging (e.g., semi-transparent clouds, cloud boundaries) cases are equally present in the dataset. At the same time, other datasets were aimed at labelling the full images (L8Biome, CESBIO) or provide homogeneous polygons (Hollstein, GSFC), where the weight of challenging cases would be lower than for PixBox. In this regard, the question is about whether to weight samples according to the area or not. Both characteristics (based on equal allocation and area proportions) can be valuable to describe separability of classes by a given algorithm (model accuracy) and to estimate probability of a pixel being mapped correctly (map accuracy) (Blickensdörfer et al., 2022; Congalton, 1991).

Across the four reference datasets algorithms showed better performance in terms of cloud UA, which was consistently higher than cloud PA. Removing thin/semi-transparent clouds from the reference datasets improves performance of algorithms (Figure 20), though at the expense of cloud UA. This happens because thin clouds have higher uncertainties and therefore are more challenging to the algorithms in contrast to thick clouds. When thin clouds removed from reference datasets the proportion of correctly detected classes increases and

therefore cloud PA increases. At the same time, cloud UA can experience both increase or decrease depending on the proportion of thin clouds and algorithm's performance on thin clouds.

The issue of thin/semi-transparent cloud detection has a significant impact on the subsequent shadow detection. Figure 21 shows an example of a cloud with different levels of transparency depending on wavelength used and its shadow. While the cloud is semi-transparent in the false color composite (SWIR-NIR-red), its shadow is clearly visible and impacts the reflectance.

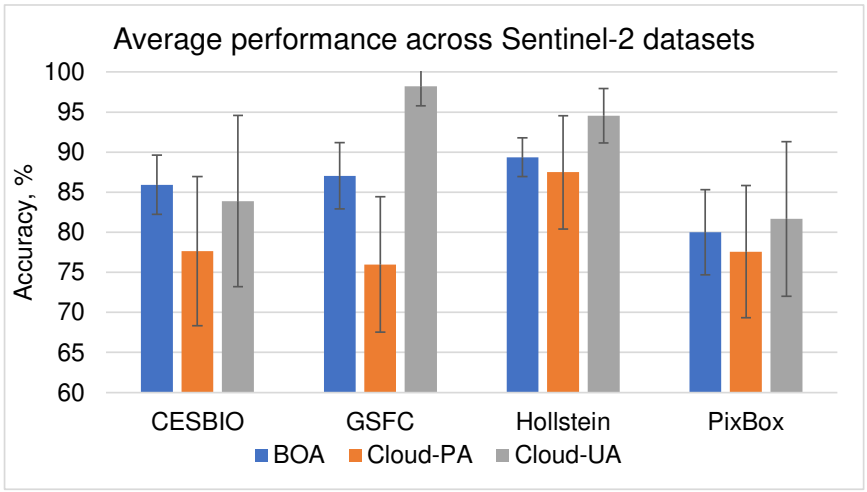
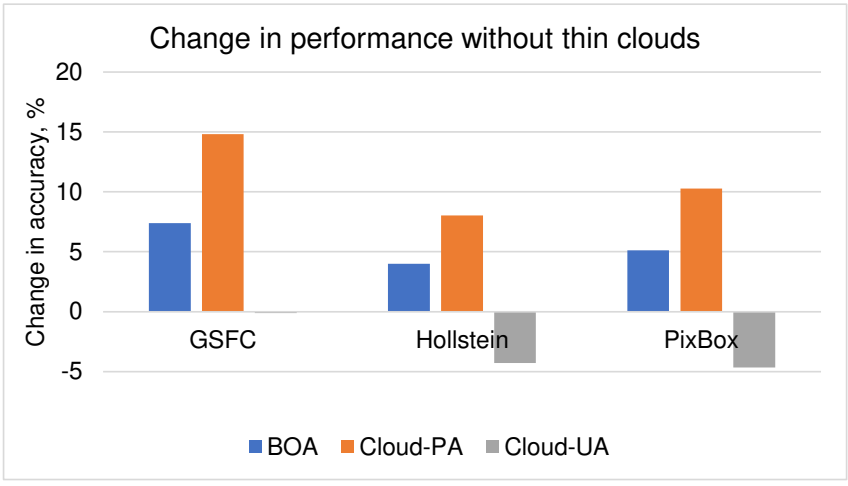
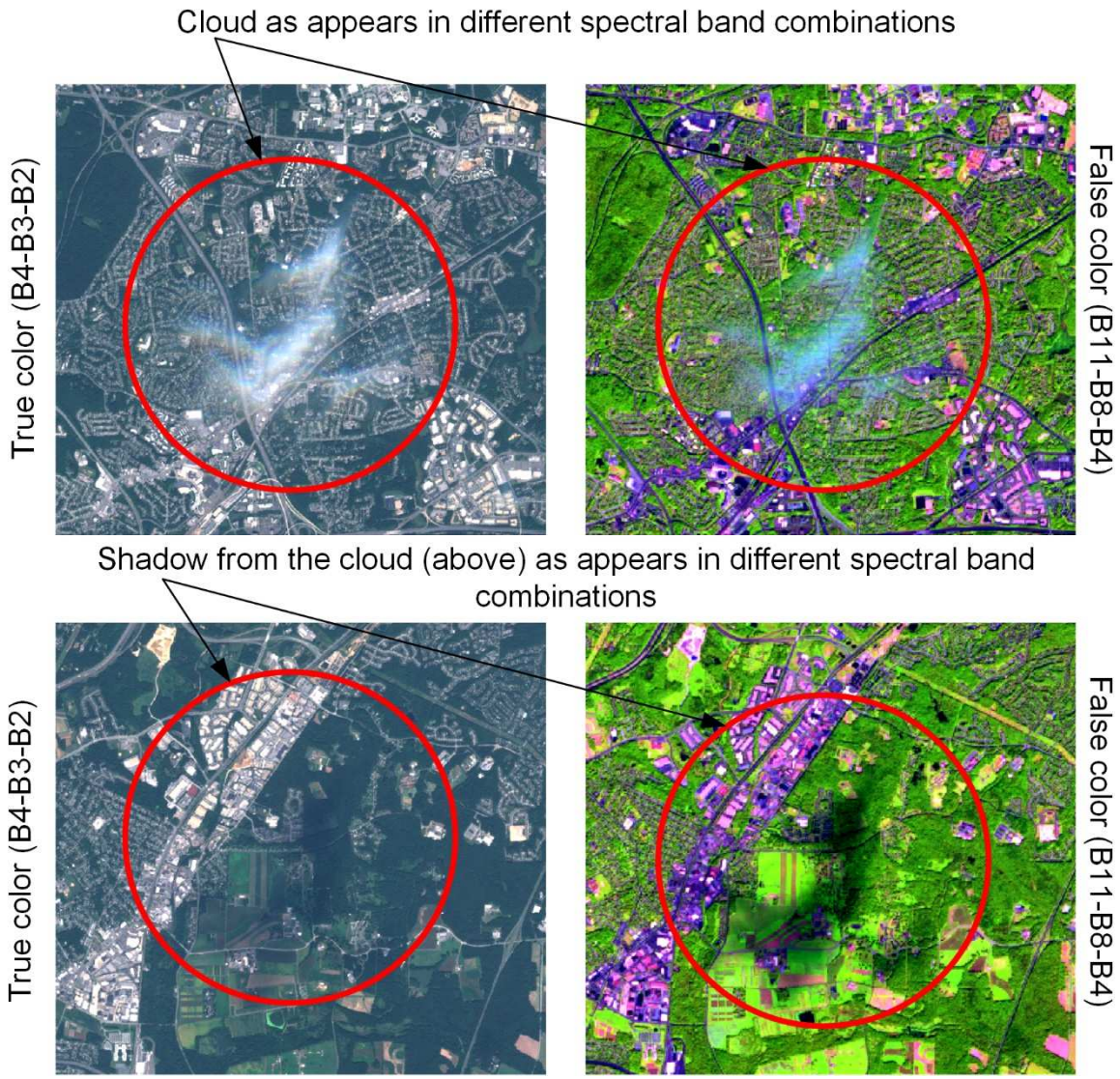


Figure 19. Average performance of algorithms for Sentinel-2 for four cloud reference datasets.



878 Figure 20. Change in performance of Sentinel-2 cloud masking algorithms, when thin/semi-
879 transparent clouds removed from the reference datasets.
880



881
882 Figure 21. Example of thin/semi-transparent cloud in various band combinations (true color
883 and false color in top-of-atmosphere reflectance) along with the shadow from that cloud
884 (Sentinel-2 scene, L1C_T18SUJ_A011777_20170923T160124).
885
886 Figure 22 shows averaged BOA values across multiple Landsat 8 algorithms. As with
887 Sentinel-2, the performance varied across datasets yielding BOA of 97.6±0.8%, 90.0±1.3%

and 79.8±7.1% for GSFC, L8Biome and PixBox, respectively. As with Sentinel-2, cloud PA was higher than cloud UA for GSFC and PixBox datasets, but not for L8Biome.

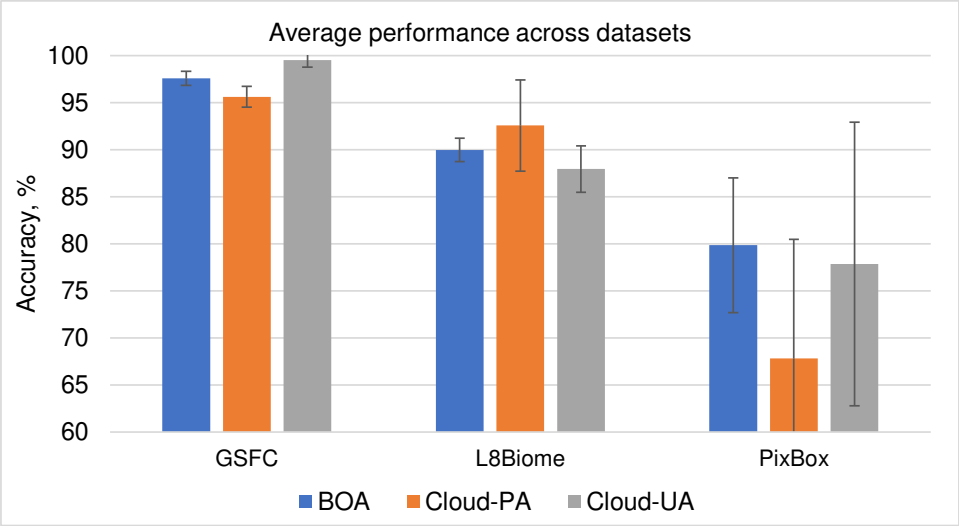


Figure 22. Average performance of algorithms for Landsat 8 for three cloud reference datasets.

In terms of various land cover classes, it is difficult to draw conclusions since only generic "static" information on land cover was available for some of the datasets. We did not observe any substantial differences in algorithm's performance over various land cover classes, except for urban areas in the PixBox S2 data, which is expected. Sentinel-2 does not have a thermal band and, therefore, detection of clouds over bright targets in urban areas remains a challenging task. The use of multi-spectral parallax (Skakun et al., 2017) only partially addresses this problem (Frantz et al., 2018).

5 Recommendations

Results and lessons learned from CMIX-I provide a good foundation for future activities for improving practices related to the development and validation of cloud masking algorithms for passive optical satellite imagery.

907 The first area for improvement should aim at initially providing an agreed upon
908 definition of "cloud" (Mejia et al., 2016; Stubenrauch et al., 2013) that is passed beforehand
909 to intercomparison participants and validation dataset originators. Ideally this would be an
910 objective (quantitative) definition of clouds, which would include a numerical metric. As
911 results from CMIX-I showed, existing validation datasets varied in how a cloud was defined
912 through mostly photointerpretation, and it influenced the performance of the algorithms. For
913 example, one potential metric to define the cloud would be the cloud optical thickness.
914 However, this poses the questions at which wavelength the thickness should be defined, what
915 threshold to apply, and how it could be estimated for sizeable quantity of images. For
916 example, Mejia et al. (2016) use a radiative transfer model to estimate cloud optical depth (τ_c)
917 from ground-based sky images and define thick clouds with τ_c>30, thin clouds with τ_c~1, and
918 clear sky with τ_c ~0 (all in the visible spectrum). While there was a consensus between
919 algorithms and developers in defining thick non-transparent clouds, there was a disagreement
920 (sometimes by design and depending on the intended applications) in transparent (semi-
921 transparent) clouds, such as cirrus, stratus and cloud edges. Also, the effect of those clouds
922 can vary with wavelengths, which adds complexity to the analysis.
923 Based on the cloud definition, the second area for improvement would include
924 generation of new reference/validation datasets. The strengths and weaknesses of existing
925 cloud reference datasets were thoroughly analyzed and discussed within this study, and new
926 datasets should substantially address those weaknesses. A special attention should be paid to
927 ensure a balanced statistical distribution of surface and cloud types, as well as the need to
928 cover a wide range of environmental conditions, in order to thoroughly test the performance
929 of the algorithms at global scale. Some of the recommendations include:
930 • Consistently implementing the cloud definition, and adding cloud shadows to the
931 analysis. Recommended practices for labelling clouds should be developed and

implemented for new datasets, whether through visual interpretation or ground measurements or ancillary data (e.g. geostationary satellites). Cloud shadows should be also part of the analysis, since an inaccurate cloud shadow mask can lead to substantial artifacts in the downstream products.

- Defining a proper dilation of cloud masks to be applied, taking account the effect of parallax between spectral bands, smooth variation of clouds at their borders, and adjacency effects. The dilation could then be applied to the reference datasets and to the algorithm results.
- Increasing the number of sites collecting ground-based imagery of the sky and use them in coordination with Aeronet measurements. Some limitations of the use of ground-based sky imagery include radiance contrast which could yield better detection of thin clouds; furthermore, the geometrical matching between sky-camera and satellite pixel may introduce some errors, which are related to the cloud height.
- Acquire multiple datasets (time-series) over the same area to analyze consistent errors in cloud detection. This would enable temporal metrics to be exploited when assessing the efficiency of cloud masks.

The third set of activities should focus on expanding the analysis framework, which would include:

- A sample-based approach versus an area-based approach, when comparing reference cloud mask with a predicted one. The problem with an area-based approach is that more weight would be given to large clouds (which cover the larger area), whereas smaller clouds might have a small impact on the performance metrics. At the same time, sampled-based approaches can also miss some specific land cover features (unless a stratification scheme can be constructed with strata describing those features), and often do not address the boundaries of the clouds or more broadly

segmentation aspects. Area-based approaches are likewise necessary to study the effects of cloud dilation. Therefore, both approaches should be considered.

- Temporal analysis of cloud masks over the same area. Originally planned for CMIX-I, the idea of using temporal metrics was abandoned, since no reference data (except GSFC, which were assisted with sky imagery and Aeronet measurements) was available for these purposes. As undetected clouds add noise on time-series, it is possible to evaluate the noise on time-series and compute the contribution of different cloud masks to this noise.
- Application-based approach to cloud validation. One way to analyze efficiency of the cloud/shadow masks is to "validate" them indirectly within the downstream products. An example could include a generic land cover mapping workflow, when the same set of satellite data will be processed by various cloud detection algorithms and used as input to the classification algorithm. The derived land cover maps will be validated using the same validation data and intercompared.

And finally, CMIX-I was limited to Landsat 8 and Sentinel-2 data. Future activities could include adding hyper-spectral data (such as PRISMA or DESIS), coarse resolution data (such as MODIS, VIIRS, Sentinel-3), and commercial very high spatial resolution satellites, such as Planet or hyperspectral sensors.

6 Conclusion

The Cloud Mask Intercomparison eXercise (CMIX) was a community-wide effort to intercompare the state-of-the-art and commonly-used cloud masking algorithms, with a focus on moderate spatial resolution data acquired by Landsat 8 and Sentinel-2 missions. Ten algorithms developed by nine teams from fourteen organizations representing universities, industry and space agencies were evaluated within CMIX using existing cloud reference data.

Overall, the performance of algorithms varied depending on the reference dataset, which can be attributed to differences in which reference datasets were generated. Average overall accuracy (across algorithms) varied 80.0±5.3% to 89.4±2.4% for Sentinel-2, and 79.8±7.1% to 97.6±0.8% for Landsat 8, depending on the reference dataset. An overall accuracy of 90% yields twice less errors than an overall accuracy of 80%. The study highlighted algorithms that provided a balance between commission and omission errors, as well as algorithms which are cloud conservative (high UA) and non-cloud (clear) conservative (high PA). With repetitive observations like those of Sentinel-2, it seems reasonable to favor cloud conservative approaches, with maybe the exception of very cloudy regions where every cloud free observation is critical. When thin/semi-transparent clouds were not considered in the reference datasets algorithms' performance generally improved: overall accuracy values increased from +1.5% to 7.4%. It should be noted though that these clouds are commonly occurring and are often present in optical imagery. We concluded the paper with recommendations for further activities, which include provision of a quantitative definition for clouds (targeting moderate spatial resolution imagery by Landsat 8 and Sentinel-2), generation of new reference datasets, and expansion of the analysis framework (for example, multi-temporal analysis and application-driven validation). Such intercomparison studies will hopefully help the community to improve the algorithms and move towards standardization of cloud masking. Given the importance of cloud masking in optical imagery we encourage CEOS to continue the CMIX activities.

Acknowledgment

We would like to thank to Chris Justice (University of Maryland) for helpful comments on an earlier draft of paper and Gasmine Myers (University of Maryland) for

proof-reading the paper. L.G.C., D.L.P. and G.M.G. (University of Valencia) were supported for this work by the Spanish Ministry of Science and Innovation (project PID2019-109026RB- I00, ERDF) and the European Social Fund. S.S., J.C.R. (University of Maryland) and E.V. (NASA GSFC) were supported by NASA grants 80NSSC19K1592, 80NSSC19M0222 and 80NSSC21M0080.

Appendix A.

Table A1. Cloud and non-cloud classes that were used from the original reference datasets.

Dataset	Cloud	Non-cloud
CESBIO	Low clouds, high clouds	Shadow, land, water, snow
GSFC	Cloud, thin cloud	Clear, cloud shadow
Hollstein	Cloud, cirrus	Clear, water, shadow, snow
L8Biome	Thin cloud, thick cloud	Shadow, clear
PixBox S2	Opaque, thick semi-transparent cloud, average density semi-transparent cloud, semi-transparent cloud, thin semi-transparent cloud, fog, haze	Clear
PixBox L8	Cloud, semi-transparent cloud	Clear land, clear snow/ice, clear water, mixed snow_ice/water

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. Remote Sens. 11, 433.

Blickensdörfer, L., Schwieder, M., Pflugmacher, D., Nendel, C., Erasmi, S., Hostert, P., 2022. Mapping of crop types and crop sequences with combined time series of Sentinel-1, Sentinel-2 and Landsat 8 data for Germany. Remote Sens. Environ. 269, 112831.

Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution, in: Proc. 2010 20th International Conference on Pattern Recognition. IEEE, pp. 3121–3124.

Chai, D., Newsam, S., Zhang, H. K., Qiu, Y., Huang, J., 2019. Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks. Remote Sens. Environ. 225, 307–316.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ. 37(1), 35-46.

Doxani, G., Vermote, E., Roger, J.C., Gascon, F., Adriaensen, S., Frantz, D., Hagolle, O., Hollstein, A., Kirches, G., Li, F., Louis, J., 2018. Atmospheric correction inter-comparison exercise. Remote Sens. 10(2), 352.

EO Research Team, 2020. Cloud Masks at Your Service. https://medium.com/sentinel-hub/cloud-masks-at-your-service-6e5b2cb2ce8a (accessed 11 July 2021).

Foga, S., Scaramuzza, P.L., Guo, S., Zhu, Z., Dilley Jr, R.D., Beckmann, T., Schmidt, G.L., Dwyer, J.L., Hughes, M.J., Laue, B., 2017. Cloud detection algorithm comparison and

validation for operational Landsat data products. Remote Sens. Environ. 194, 379–390.

[Dataset] Francis, A., Mrziglod, J., Sidiropoulos, P., Muller, J.-P., 2020. Sentinel-2 Cloud Mask Catalogue (Version 1.0). Zenodo. https://doi.org/10.5281/zenodo.4172871

Frantz, D., 2019. FORCE—Landsat + Sentinel-2 Analysis Ready Data and Beyond. Remote Sens. 11, 1124.

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.

Frantz, D., Röder, A., Stellmes, M., Hill, J., 2016. An Operational Radiometric Landsat Preprocessing Framework for Large-Area Time Series Applications. IEEE Trans. Geosc. Remote Sens. 54 (7), 3928–3943.

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 Geoscie. Remote Sens. Lett. 12(6), 1242–1246.

Gascon, F., Bouzinac, C., Thépaut, O., Jung, M., Francesconi, B., Louis, J., Lonjou, V., Lafrance, B., Massera, S., Gaudel-Vacaresse, A., Languille, F., 2017. Copernicus Sentinel-2A calibration and products validation status. Remote Sens. 9(6), 584.

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(8), 1747–1755.

[Dataset] Hagolle, O., Huc, M., Desjardins, C., Auer, S., Richter, R., 2017. MAJA Algorithm Theoretical Basis Document (Version 1.0). Zenodo. http://doi.org/10.5281/zenodo.1209633

Holben, B.N., Eck, T.F., Slutsker, I.A., Tanre, D., Buis, J.P., Setzer, A., Vermote, E., et al. 1998. AERONET—A federated instrument network and data archive for aerosol characterization. Remote Sens. Environ. 66(1), 1–16.

Hollingsworth, B.V., Chen, L., Reichenbach, S.E., Irish, R.R., 1996. November). Automated cloud cover assessment for Landsat TM images. In: Imaging Spectrometry II, Vol. 2819. International Society for Optics and Photonics, pp. 170–179.

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. Remote Sens. 8(8), 666.

Hughes, M.J., Hayes, D.J., 2014. Automated detection of cloud and cloud shadow in single-date Landsat imagery using neural networks and spatial post-processing. Remote Sens. 6 (6), 4907–4926.

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 (10), 1179–1188.

Jeppesen, J.H., Jacobsen, R.H., Inceoglu, F., Toftegaard, T.S., 2019. A cloud detection algorithm for satellite imagery based on deep learning. Remote Sens. Environ. 229, 247–259.

Kotchenova, S.Y., Vermote, E.F., Matarrese, R., Klemm Jr, F.J., 2006. Validation of a vector version of the 6S radiative transfer code for atmospheric correction of satellite data. Part I: Path radiance. Appl. Opt. 45(26), 6762–6774.

Lamarche, C., Santoro, M., Bontemps, S., d'Andrimont, R., Radoux, J., Giustarini, L., Brockmann, C., Wevers, J., Defourny, P., Arino, O., 2017. Compilation and validation of SAR and optical data products for a complete and global map of

1090 inland/ocean water tailored to the climate modeling community. Remote Sens. 9(1),
1091 36.
1092 López-Puigdollers, D., Mateo-García, G., Gómez-Chova, L., 2021. Benchmarking Deep
1093 Learning Models for Cloud Detection in Landsat-8 and Sentinel-2 Images. Remote
1094 Sens. 13(5), 992.
1095 Louis, J., Debaecker, V., Pflug, B., Main-Knorn, M., Bieniarz, J., Mueller-Wilm, U., Cadau,
1096 E., Gascon, F., 2016. Sentinel-2 sen2cor: L2a processor for users, in: Proceedings
1097 Living Planet Symposium 2016. Spacebooks Online, pp. 1–8.
1098 Louis, J., 2021. Sentinel-2 Level-2A Algorithm Theoretical Basis Document.
1099 https://sentinels.copernicus.eu/documents/247904/446933/Sentinel-2-Level-2A-
1100 Algorithm-Theoretical-Basis-Document-ATBD.pdf (accessed 9 July 2021)
1101 Mateo-García, G., Laparra, V., López-Puigdollers, D., Gómez-Chova, L., 2020. Transferring
1102 deep learning models for cloud detection between Landsat-8 and Proba-V. ISPRS J.
1103 Photogramm. Remote Sens. 160, 1–17.
1104 Mejia, F.A., Kurtz, B., Murray, K., Hinkelman, L.M., Sengupta, M., Xie, Y., Kleissl, J.,
1105 2016. Coupling sky images with radiative transfer models: a new method to estimate
1106 cloud optical depth. Atm. Meas. Techn. 9(8), 4151–4165.
1107 Pahlevan, N., Mangin, A., Balasubramanian, S.V., Smith, B., Alikas, K., Arai, K., Barbosa,
1108 C., et al. 2021. ACIX-Aqua: A global assessment of atmospheric correction methods
1109 for Landsat-8 and Sentinel-2 over lakes, rivers, and coastal waters. Remote Sens.
1110 Environ. 258, 112366.
1111 [Dataset] Paperin, M., Wevers, J., Stelzer, K., Brockmann, C., 2021a. PixBox Sentinel-2
1112 pixel collection for CMIX (Version 1.0). Zenodo.
1113 https://doi.org/10.5281/zenodo.5036991

[Dataset] Paperin, M., Stelzer, K., Lebreton, C., Brockmann, C., Wevers, J., 2021b. PixBox Landsat 8 pixel collection for CMIX (Version 1.0). Zenodo. https://doi.org/10.5281/zenodo.5040271

Puc, J., Žust, L., 2019. On cloud detection with multi-temporal data. https://medium.com/sentinel-hub/on-cloud-detection-with-multi-temporal-data-f64f9b8d59e5 (accessed 09 June 2021)

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., Schläpfer, D., 2019a. ATCOR-3 User Guide. Version 9.3.0. https://www.rese-apps.com/pdf/atcor3_manual.pdf (accessed 09 June 2021)

Richter, R., Schläpfer, D., 2019b. ATCOR Theoretical Background Document. Version 1.0. https://www.rese-apps.com/pdf/atcor_ATBD.pdf (accessed 03 July 2021).

Scaramuzza, P.L., Bouchard, M.A., Dwyer, J.L., 2012. Development of the Landsat data continuity mission cloud-cover assessment algorithms. IEEE Trans. Geosci, Remote Sens. 50(4), 1140–1154.

Segal-Rozenhaimer, M., Li, A., Das, K., Chirayath, V., 2020. Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN). Remote Sens. Environ. 237, 111446.

Skakun, S., Vermote, E.F., Artigas, A.E.S., Rountree, W.H., Roger, J.-C., 2021. An experimental sky-image-derived cloud validation dataset for Sentinel-2 and Landsat 8 satellites over NASA GSFC. Int. J. Appl. Earth Observ. Geoinform. 95, 102253.

Skakun, S., Vermote, E., Roger, J.-C., Justice, C., 2017. Multispectral misregistration of Sentinel-2A images: Analysis and implications for potential applications. IEEE Geosci. Remote Sens. Lett. 14(12), 2408–2412.

1138 Skakun, S., Vermote, E.F., Roger, J.-C., Justice, C.O., Masek, J.G., 2019. Validation of the
1139 LaSRC cloud detection algorithm for Landsat 8 images. IEEE J. Sel. Topics Appl.
1140 Earth Observ. Remote Sens. 12(7), 2439–2446.
1141 Stubenrauch, C.J., Rossow, W.B., Kinne, S., Ackerman, S., Cesana, G., Chepfer, H., Di
1142 Girolamo, L., Getzewich, B., Guignard, A., Heidinger, A., Maddux, B.C., *et al.*, 2013.
1143 Assessment of global cloud datasets from satellites: Project and database initiated by
1144 the GEWEX radiation panel. Bull. American Meteorol. Soc. 94(7), 1031–1049.
1145 Tarrio, K., Tang, X., Masek, J.G., Claverie, M., Ju, J., Qiu, S., Zhu, Z., Woodcock, C.E.,
1146 2020. Comparison of cloud detection algorithms for Sentinel-2 imagery. Sci. Remote
1147 Sens. 2, 100010.
1148 [Dataset] U.S. Geological Survey, 2016. L8 SPARCS Cloud Validation Masks. U.S.
1149 Geological Survey data release. USGS. doi:10.5066/F7FB5146
1150 Vermote, E., Justice, C., Claverie, M., Franch, B., 2016. Preliminary analysis of the
1151 performance of the Landsat 8/OLI land surface reflectance product. Remote Sens.
1152 Environ. 185, 46–56.
1153 Vermote, E.F., Tanré, D., Deuze, J.L., Herman, M., Morcette, J.J., 1997. Second simulation
1154 of the satellite signal in the solar spectrum, 6S: An overview. IEEE Trans Geosci.
1155 Remote Sens. 35(3), 675–686.
1156 Wevers, J., Müller, D., Scholze, J., Kirches, G., Quast, R., Brockmann. C., 2021. IdePix for
1157 Sentinel-2 MSI Algorithm Theoretical Basis Document (Version 1.0). Zenodo.
1158 https://doi.org/10.5281/zenodo.5788067
1159 Wieland, M., Li, Y., Martinis, S., 2019. Multi-sensor cloud and cloud shadow segmentation
1160 with a convolutional neural network. Remote Sens. Environ. 230, 111203.

1161 Wulder, M.A., Loveland, T.R., Roy, D.P., Crawford, C.J., Masek, J.G., Woodcock, C.E.,
1162 Allen, R.G., et al., 2019. Current status of Landsat program, science, and applications.
1163 Remote Sens. Environ. 225, 127–147.
1164 Xie, F., Shi, M., Shi, Z., Yin, J., Zhao, D., 2017. Multilevel cloud detection in remote sensing
1165 images based on deep learning. IEEE J. Sel. Topics Appl. Earth Observ. Remote
1166 Sens. 10(8), 3631–3640.
1167 Zhu, Z., Wang, S., Woodcock, C.E., 2015. Improvement and expansion of the Fmask
1168 algorithm: Cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel
1169 2 images. Remote Sens. Environ. 159, 269–277.
1170 Zhu, Z., Woodcock, C.E., 2012. Object-based cloud and cloud shadow detection in Landsat
1171 imagery. Remote Sens. Environ. 118, 83–94.
1172 Zhu, Z., Woodcock, C.E., 2014. Automated cloud, cloud shadow, and snow detection in
1173 multitemporal Landsat data: An algorithm designed specifically for monitoring land
1174 cover change. Remote Sens. Environ. 152, 217–234.
1175 Zupanc, A., 2017. Improving Cloud Detection with Machine Learning.
1176 https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-
1177 c09dc5d7cf13 (accessed 09 June 2021)
1178

List of Figure Captions

Figure 1. Geographical distribution of the Landsat 8 and Sentinel-2 scenes in the reference datasets used in CMIX.

Figure 2. Distribution of labeled pixels in the CESBIO dataset.

Figure 3. Distribution of labeled pixels in the GSFC S2 dataset (left) and land cover classes (right).

Figure 4. Distribution of labeled pixels in the Hollstein dataset.

Figure 5. Distribution of labeled pixels in the L8Biome dataset.

Figure 6. Distribution of labeled pixels and land cover classes in the PixBox dataset.

Figure 7. Part of the L8Biome scene (LC81570452014213LGN00) with some thin clouds not labelled. Thin clouds are shown in orange, and thick clouds in maroon.

Figure 8. Examples of labeled data in the three datasets: CESBIO (fully labeled images); GSFC (polygons avoiding uncertain areas, such cloud boundaries); PixBox (sample-based approach).

Figure 9. Comparison of BOA values and distribution of PA/UA for the CESBIO reference dataset.

1204

1205 Figure 10. Comparison of BOA values and distribution of PA/UA (for all clouds) for the
1206 GSFC S2 reference dataset.

1207

1208 Figure 11. Comparison of BOA values and distribution of PA/UA (for all clouds) for the
1209 Hollstein reference dataset.

1210

1211 Figure 12. Comparison of BOA values and distribution of PA/UA (for all clouds) for the
1212 PixBox S2 reference dataset.

1213

1214 Figure 13. Examples of cloud masking by various algorithms over the Sentinel-2 scene
1215 S2A_MSIL1C_20170629T103021_N0205_R108_T31TFJ_20170629T103020.

1216

1217 Figure 14. Performance of algorithms in terms of clear producer's accuracy over the non-
1218 cloudy regions depending on the land cover types in the PixBox S2 dataset.

1219

1220 Figure 15. Performance of the Landsat 8 cloud detection algorithms for the L8Biome dataset
1221 depending on the biomes. The same set of 80 Landsat 8 scenes was used to calculate PA and
1222 UA accuracy values.

1223

1224 Figure 16. PA values for various types of classes in the PixBox L8 dataset.

1225

1226 Figure 17. Distribution of cloud PA and UA over all Sentinel-2 cloud masking algorithms
1227 and reference datasets (left) and algorithms' average values along with the standard deviation

over four reference datasets (right). Averaging was performed using PA and UA values from Table 6, Table 7, Table 8 and Table 9 for all cloud types.

Figure 18. Distribution of cloud PA and UA over all Landsat 8 clouds masking algorithms and reference datasets.

Figure 19. Average performance of algorithms for Sentinel-2 for four cloud reference datasets.

Figure 20. Change in performance of Sentinel-2 cloud masking algorithms, when thin/semi-transparent clouds removed from the reference datasets.

Figure 21. Example of thin/semi-transparent cloud in various band combinations (true color and false color in top-of-atmosphere reflectance) along with the shadow from that cloud (Sentinel-2 scene, L1C_T18SUJ_A011777_20170923T160124).

Figure 22. Average performance of algorithms for Landsat 8 for three cloud reference datasets.