Cloud Mask Inter-comparison eXercise
Final Report

Author(s):

Jan Wevers, BC
Carsten Brockmann, BC
Sergii Skakun, UMD

Reviewer(s):

Georgia Doxani, ESA

Approver(s):

Georgia Doxani, ESA

NASA LOGO CEOS LOGO ESA LOGO

Amendment Record

Releasing a new edition of the document in its entirety shall amend this document. The Amendment Record Sheet below records the history and issue status of this document.

Amendment Record Sheet

Issue	Date	Reason for change
V0.1	23-June-2020	Initial Version of the report
V0.2	08-December-2020	Additional inputs from UoM
V0.3	21-January-2021	Multiple updates and additions
V0.4	12-March-2021	Multiple updates and additions
V1.0	30-April-2021	Final edits
V1.1	21-June_2021	Additional inputs from the participants

Table of contents

List of figures

List of tables

Executive summary

Cloud cover is a limiting factor in exploiting data acquired by optical spaceborne remote sensing sensors. Multiple methods have been developed to address the problem of cloud and cloud shadow detection in satellite imagery, but very few studies were carried out to quantitatively inter-compare state-of-the-art methods in this domain. This report summarizes results of the first Cloud Masking Inter-comparison eXercise (CMIX) conducted within the Committee Earth Observation Satellites (CEOS) Working Group on Calibration & Validation (WGCV). CMIX is an international collaborative effort aimed at inter-comparing cloud detection algorithms for medium-spatial resolution (10-30 m) spaceborne optical sensors. The focus of this effort was on open, free and repetitive imagery acquired by Landsat 8 (NASA/USGS) and Sentinel-2 (ESA) missions. Ten algorithms developed by ten organizations representing universities and industry, as well as space agencies (CNES, ESA, DLR, and NASA), were evaluated within the CMIX. Those algorithms varied in principles and concepts utilized and were based on spectral properties, spatial and temporal features, as well as machine learning methods. Algorithms outputs were evaluated against existing publicly available reference ("ground truth") cloud mask datasets. Those datasets varied in the way they were sampled and geographically distributed, sample unit used (points, polygons, full image labels), and generated (experts, machine learning, sky images). Within CMIX, a qualitative definition of "cloud" was adopted, which provides an absolute (spectrally non-dependent) indication of cloudiness in the satellite image. Though different cloud masking algorithms produce different output cloud layers, ultimately each cloud mask was converted to "cloud" and "non-cloud" values, so the best matching inter-comparison could be performed. Algorithms were compared using the same set of data under the same conditions. The CMIX experiment did not address the issue of cloud shadow detection. Nevertheless, cloud shadow is planned to be included in a second round of CMIX.

Overall, the performance of algorithms varied depending on the reference dataset, which can be attributed to differences in cloud definitions used in producing reference datasets. More consensus among algorithms was achieved for thick clouds (which were opaque and had less uncertainties in cloud definitions) than thin/transparent clouds, detection of which relied on various definitions and intended applications. Not only CMIX allowed identification of strengths and weaknesses of existing algorithms and potential areas of improvements, but also the problems with existing reference datasets. The report concludes with recommendations on generating new reference datasets, metrics and analysis framework to be further exploited and inclusion of additional datasets to be considered within future CMIX activities.

1 Introduction

1.1 Objective

The Cloud Masking Inter-comparison eXercise (CMIX) is an international collaborative initiative in the frame of Committee Earth Observation Satellites (CEOS) Working Group on Calibration & Validation (WGCV) to inter-compare a set of cloud detection algorithms for space-borne medium-spatial resolution (10-30 m) optical sensors. The exercise focuses on Landsat 8 and Sentinel-2 imagery acquired over various locations and under a range of cloud conditions. The validation of cloud screening outputs is based on existing reference ("ground truth") cloud mask datasets. The inter-comparison of cloud masks is expected to contribute to the better understanding of strengths, limitations and applicability of different algorithms.

1.2 Scope

This document provides a quantitative inter-comparison between various cloud masking algorithms for Landsat 8 and Sentinel-2 against existing reference cloud datasets. Within CMIX, a qualitative definition of "cloud" was adopted, which provides an absolute (spectrally non-dependent) indication of cloudiness in the satellite image. Though different cloud masking algorithms produce different output cloud layers, ultimately each cloud mask was converted to "cloud" and "non-cloud" values, so the best matching inter-comparison could be performed. Algorithms were compared using the same set of data under the same conditions. It was communicated to algorithm providers that if the algorithm was developed/calibrated using the reference cloud dataset, that algorithm was excluded in the inter-comparison using that dataset.

2 Submitted Algorithms and participants

Nine participants, representing space agencies, universities, and the private sector, have submitted 10 algorithms to the CMIX. A summary of cloud-masking algorithms is presented in Table 1, while the detailed description is given in the subsequent subsections.

It is important to mention again that a binary masking of cloud/non-cloud was a requirement made to all participants prior to the exercise. Nearly all algorithms provide more detailed information on clouds, with some providing even cloud probability information on pixel level, which had to be converted into this binary mask. Very detailed information can be complicated to be handled by the users but give a high level of freedom for adaptions to users' needs/preferences.

Table 1: Algorithm characteristics (L8: Landsat 8, S2: Sentinel-2)

Processor	Organization	Methodology	Provided resolution, m	Temporality	Cloud mask dilation (buffer)
ATCOR	DLR	Spectral tests	L8: 30 S2: 20	Mono	No
CD-FCNN	University of Valencia	Machine learning	L8: 30 S2: 10/20/60	Mono	No

Fmask 4.0 CCA algorithm	USGS	Spectral tests	L8: 30 S2:20	Mono	Medium
FORCE	Humboldt-Universität zu Berlin	Spectral test +parallax (S2 only)	L8: 30 S2: 10	Mono	Medium/Large
IdePix	Brockmann Consult	Spectral tests	S2: 20	Mono	Small
InterSSIM	Sinergise	Machine learning + spatio-temporal context	S2: 10	Multi	Medium
LaSRC	NASA / University of Maryland	Spectral tests	L8: 30 S2: 10	Mono	Small
MAJA	CNES/CESBIO	Spectral tests	S2: 240	Multi	Large
s2cloudless	Sinergise	Machine learning	S2: 10	Mono	Medium
sen2cor	ESA / Telespazio France / DLR	Spectral test + auxiliary data	S2: 20	Mono	No

2.1 ATCOR

The first version of the ATCOR model was developed in 1990 (Richter, 1990). Over the years the model was continually improved and now it supports the mono-temporal atmospheric correction processing of multispectral and hyperspectral imagery in the reflective spectrum (400 – 2500 nm), as well as the thermal spectrum (7 – 14 µm).

ATCOR software code is written in IDL. ATCOR version 9.3.0 was used for the CMIX processing.

The ATCOR processor contains a scene pre-classification module. It is based on simple spectral criteria with the top-of-atmosphere (TOA) reflectance. CMIX processing of ATCOR did not use a Digital Elevation Model (DEM) and no other auxiliary data.

Pre-classification module classifies a satellite image into the following classes (Richter and Schläpfer, 2016):

- Land
- Water
- Non-cirrus cloud
- Cirrus cloud
- Snow/ice
- Cloud/shadow & topographic shadow.

Pre-classification results are provided both at 10 m and 20 m spatial resolution for Sentinel-2 and at 30 m spatial resolution for Landsat 8. ATCOR uses a 100 m buffer for clouds, 220 m buffer for cloud shadow and a 20 m buffer for snow.

The processing time (Intel 3.5 GHz PC, Ubuntu 14.04) for complete atmospheric correction including pre-classification is as follows:

- Landsat 8 OLI (7711 × 7861 pixels) — flat terrain: 2 min; with DEM: 8 min
- Sentinel-2: import of jp2 files (conversion into layer-stacked radiance cubes: 13 bands at 20 m, 4 bands at 10 m): 4 min
- S2 surface reflectance:
  - 13 band cube (5490 × 5490 pixels) — flat terrain: 3 min; with DEM: 6 min
  - 4 band cube (10,980 × 10,980 pixels) — flat terrain: 2 min; with DEM: 4 min

ATCOR provided pre-classification results for CMIX at 20 m spatial resolution for Sentinel-2 data and at 30 m spatial resolution for Landsat 8 data.

The classes in the output for CMIX are aggregated and coded as follows:

0: geocoded background
1: clear
2: semi-transparent cloud
3: cloud
4: cloud shadow

Values 1 and 4 were used for “non-cloud” class, and 2 and 3 for “cloud” class.

Note, that not all provided data products could be successful processed with ATCOR because of:

- solar zenith angle being larger than 75 deg. ATCOR processing is limited to solar zenith angles less than 75 degree, or
- the provided Landsat data (L8Biome data set) did not have the UTM map-projection requested by ATCOR (those scenes are in the polar projection).

2.2 CD-FCNN (University of Valencia)

The University of Valencia proposal for CMIX presents a cloud detection approach based on deep learning methods. It is intended to be applied to a single multispectral image from medium spatial resolution satellites, such as Landsat 8 and Sentinel-2. Cloud detection constitutes a complex classification problem due to the high variability of cloud characteristics, surface reflectance, and atmospheric conditions. Hence, statistical deep learning methods require a huge amount of manually annotated data in order to train accurate cloud detection models. This is a major difficulty, since quality labeled datasets usually do not exist, or are not publicly available, for most satellite sensors. Therefore, it is proposed to train fully convolutional neural networks using well established Landsat 8 datasets that could be also transferred to solve cloud detection in Sentinel-2 images. After a minimum adaptation of Sentinel-2 data, in terms of band selection and spatial resolution, the trained Landsat 8 models are directly applied to Sentinel-2. In particular, the method is based on a modified (simpler) version of the U-Net architecture proposed by the authors in (Mateo-García et al., 2020). Training and testing details of the final cloud detection model can be found in (López-Puigdollers et al., 2021), and a Python implementation of the proposed cloud detection algorithm is provided in a public repository (https://github.com/IPL-UV/DL-L8S2-UV) in order to allow the community to compare the proposed transfer models for both Landsat 8 and Sentinel-2 images.

Regarding the spatial resolution, it is worth nothing that all bands have been resampled to the native Landsat 8 resolution of 30 m. Therefore, since Sentinel-2 images have bands at several spatial resolutions of 10, 20 and 60 m, the resulting cloud mask is spatially resampled from 30 m to the corresponding resolution. In all this process, the work is done at a pixel level and no spatial dilation of the cloud mask is considered at any stage. The developed models are benchmarked against operational state-of-the-art cloud detection algorithms of both Landsat 8 and Sentinel-2. Experimental results show that the proposed transfer learning approach provides competitive accuracy on both Landsat 8 and Sentinel-2 datasets. However, a strong dependency on the particular labeled dataset used for training and validating the models appears on the results, which comes from the labeling process, and is usually neglected in most studies (López-Puigdollers et al., 2021).

The CD-FCNN cloud masks were provided in a binary mode: 1 – "cloud" and 0 – "non-cloud".

For this exercise it is important to note that CD-FCNN was trained based on the L8 Biome and the L8 SPARCS datasets (80% and 20%, respectively). L8Biome dataset is also part of the CMIX reference dataset (§ 3.1.4), so the results provided for the L8Biome dataset have to be neglected.

2.3 Fmask 4.0 CCA algorithm (USGS EROS)

Function of mask (Fmask) 4.0 is an algorithm for automated cloud and cloud shadow detection in Landsat 4-8 and Sentinel-2 images. Fmask works on a single scene basis, uses spectral tests, and creates a 3-pixel buffer around clouds and cloud shadows. It is resolution independent. Fmask is an original algorithm developed at Boston University (Zhu et al., 2015; Qui et al., 2017, 2019). For Fmask 4.0 three major innovative improvements were made as follows: (1) integration of auxiliary data, where Global Surface Water Occurrence (GSWO) data was used to improve the separation of land and water, and a global Digital Elevation Model (DEM) was used to normalize thermal and cirrus bands; (2) development of new cloud probabilities, in which a Haze Optimized Transformation (HOT)-based cloud probability was designed to replace temperature probability for Sentinel-2 images, and cloud probabilities were combined and re-calibrated for different sensors against a global reference dataset; and (3) utilization of spectral-contextual features, where a Spectral Contextual Snow Index (SCSI) was created for better distinguishing snow/ice from clouds in polar regions, and a morphology-based approach was applied to reduce the commission error in bright land surfaces (e.g., urban/built-up and mountain snow/ice).

In Fmask, pixel values 4 (Cloud) or 6 (Clear with dilated cloud) were used as "cloud" class, whereas all others were used as "non-cloud" class.

2.4 FORCE (Humboldt-Universität zu Berlin)

The cloud masking implemented in FORCE (Framework for Operational Radiometric Correction for Environmental monitoring, (Frantz, 2019), freely available from https://github.com/davidfrantz/force) is part of a mono-temporal Level 2 Processing System capable of generating Analysis Ready Data (ARD) for the Landsat and Sentinel-2 sensors. The cloud masking has branched from Fmask version 1.6.3 (Zhu and Woodcock, 2012), and has been developed in parallel since then. Updates from the original developers were partially incorporated (Zhu et al., 2015). Major modifications included the dropping of the termination criteria for shadow matching, i.e., shadow detection is more aggressive compared to the original code; it is expected that shadow commission errors are somewhat larger, however it was decided to favor commission over omission (Frantz et al., 2015). The similarity metric for matching shadows was modified, i.e. clouds were entirely excluded from the match (Frantz et al., 2016), and the shadow probability was incorporated into the match similarity to somewhat counterbalance commission errors that result from switching off the termination criterion. In addition, cloud shadows are not matched over water to reduce frequent commission errors over lakes and rivers. Furthermore, a darkness filter was included to mitigate false cloud detections in bifidly structured dryland areas,

where the scene-based temperature distribution tests for Landsat could result in commission errors of cold image parts, e.g. riverine vegetation in a desert image (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) – it should be noted that a modified version of this parallax-based procedure was adopted in the original developer's version of Fmask 4.0 (Qiu et al., 2019). The FORCE cloud masking is designed to aggressively detect clouds and cloud shadows to increase producer's accuracy at the deliberate expense of cloud commission for its safe usage in time series applications. Circular buffers are used to reduce false negatives (300 m for opaque clouds; 3 and 1 pixels at provided resolution for cloud shadows and snow, respectively).[1] FORCE provides quality bits, wherein 12 quality indicators with respect to atmospheric conditions are present (Frantz, 2019). Multiple of these 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 images, respectively.

In FORCE, pixel values with bits 1-2 set to '01' (cloud buffer), '10' (opaque cloud) or '11' (cirrus cloud) were used as "cloud" class, whereas all other pixels were used as "non-cloud" class.[2] In CMIX, the 30m SRTM DEM, filled with the 30m ASTER DEM was used for cirrus detection. All images were processed with FORCE v. 3.0-dev.

2.5 IdePix (Brockmann Consult)

IdePix (Identification of Pixels) is a multi-sensor pixel identification tool. It classifies pixels to a series of categories for further processing using a mono-temporal approach. The uniqueness consists of a certain set of features, which are calculated for each instrument, complemented by instrument specific features, a decision tree as well as probabilistic combination of these features in order to calculate a set of pixel classification attributes. Many of the instrument implementations include a neural network trained on manually classified pixels, but not the one for S2. The final classification is non-exclusive and therefore allows multiple classes to be set for a single pixel. The implementation of how the features are calculated is instrument specific. IdePix derives cloud and cirrus on multiple confidence levels, as well as cloud shadow, mountain shadow, snow and water. IdePix can be used as a stand-alone tool.

Important note for this exercise: A set of Sentinel-2 L1C products acquired during 2016 from multiple locations of the globe had been used to define adequate thresholds for the single tests. These products are not part of any validation dataset.

Cloud and snow detection

The following features (Table 2), including how they are calculated for Sentinel 2 instrument, are used in the decision tree. The decision tree is illustrated in Figure 1.

[1] After CMIX the algorithm was adapted based on the findings, to now allow the user to define custom buffers.
[2] https://force-eo.readthedocs.io/en/latest/components/lower-level/level2/format.html

Table 2: Sentinel-2 IdePix – features

Feature	Explanation/ Calculation
NDSI	NDSI = (B3-B11)/(B3+B11)
NDCI	NDCI = (B8A-B11)/(B8A+B11)
B3B11	B3B11 = B3/B11
VISBRIGHT	VISBRIGHT = (B2+B3+B4)/3
TC1	TC1 = 0.3029 B2 + 0.2786 B3 + 0.4733 B4 + 0.5599 B8A + 0.508 B11 + 0.1872 B12
TC4	TC4 = -0.8239 B2 + 0.0849 B3+ 0.4396 B4- 0.058 B8A + 0.2013 B11 – 0.2773 B12
TC4CIRRUS	$TC4_{CIRRUS}$ = -0.8239 B2+ 0.0849 B3+ 0.4396 B4- 0.058 B8A + 0.2013 B11 – 0.2773 B12 – B10

The pixel identification (IdePix) for Sentinel-2 is only working in single resolution (10m, 20m, 60m, or any specified/input resolution). The flags identified by IdePix are listed in the table (Table 3) and shown in the decision tree below (Figure 1).

Cloud edge pixels (CLOUD_BUFFER) are in principle regarded as neighbor pixels of a 'cloud' (CLOUD_SURE and CLOUD_AMBIGUOUS) as identified before in the pixel classification. The width of this edge (in number of pixels) can be set by the user. The computation is done using a moving window filter approach.

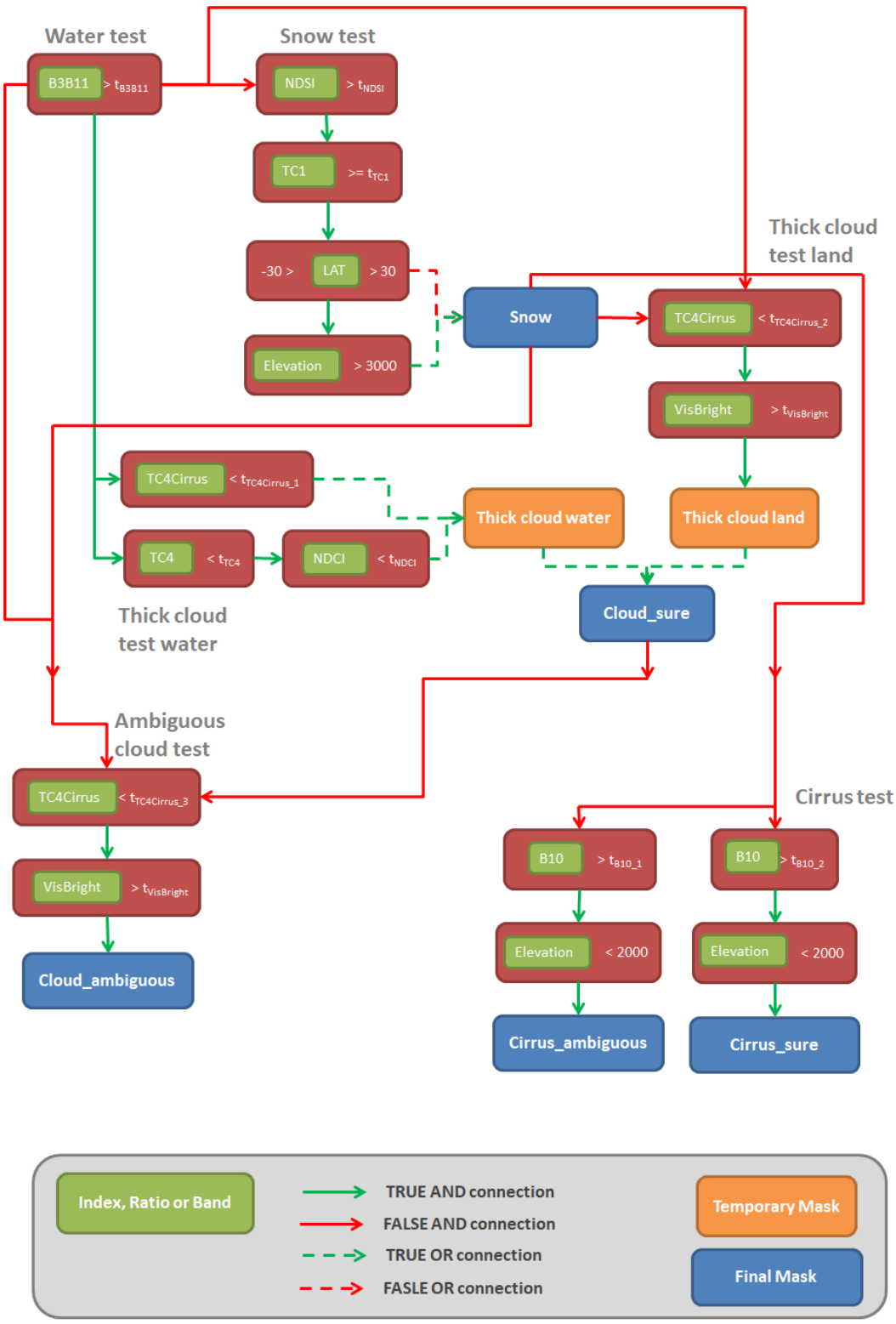


Figure 1: IdePix – Decision tree for Sentinel 2

Table 3: Sentinel-2 IdePix flagging

Bit	Integer value	Flag name	Description
1	1	IDEPIX_INVALID	Invalid pixels
2	2	IDEPIX_CLOUD	Pixels which are either cloud_sure
3	4	IDEPIX_CLOUD_AMBIGUOUS	Semi transparent clouds, or clouds where the detection level is uncertain
4	8	IDEPIX_CLOUD_SURE	Fully opaque clouds with full confidence of their detection
5	16	IDEPIX_CLOUD_BUFFER	A buffer of n pixels around a cloud. N is a user supplied parameter. Applied to pixels masked as 'cloud'
6	32	IDEPIX_CLOUD_SHADOW	Pixel is affected by a cloud shadow (combination of shifted cloud mask in cloud gaps and dark clusters coinciding with a corrected shifted cloud mask)
7	64	IDEPIX_SNOW_ICE	Clear snow/ice pixels
8	128	IDEPIX_BRIGHT	Bright pixels
9	256	IDEPIX_WHITE	White pixels
10	512	IDEPIX_COASTLINE	Pixels at a coastline
11	1024	IDEPIX_LAND	Clear land pixels
12	2048	IDEPIX_CIRRUS_SURE	Cirrus clouds with full confidence of their detection
13	4096	IDEPIX_CIRRUS_AMBIGUOUS	Cirrus clouds, or clouds where the detection level is uncertain
14	8192	IDEPIX_CLEAR_LAND	Clear land pixels
15	16384	IDEPIX_CLEAR_WATER	Clear water pixels
16	32768	IDEPIX_WATER	Water pixels
17	65536	IDEPIX_BRIGHTWHITE	'Brightwhite' pixels
18	131072	IDEPIX_VEG_RISK	Pixels with vegetation risk
19	262144	IDEPIX_MOUNTAIN_SHADOW	Pixel is affected by mountain shadow
20	524288	IDEPIX_POTENTIAL_SHADOW	Potentially a cloud shadow pixel
21	1048576	IDEPIX_CLUSTERED_CLOUD_SHADOW	Cloud shadow identified by clustering algorithm

In IdePix, bits corresponding to values 2, 4, 2048 were considered "cloud", whereas all others were considered "non-cloud".

2.6 InterSSIM (Sinergise)

The `InterSSIM` cloud detection algorithm is a multi-temporal extension of the `s2cloudless` algorithm[3] based on a gradient boosting algorithm (LightGBM)[4], which was as well trained on a training dataset with global coverage. Unlike `s2cloudless`, the `InterSSIM` algorithm takes temporal and spatial context into account. The input features are: Sentinel-2 reflectance values of the same 10 bands from the target time frame; spatially averaged reflectance values for the target frame using a Gaussian filter; minimum and mean reflectance values of all available time frames at a given spatial coordinate; maximum and mean differences of reflectance values between the target frame and any other time frame; and maximum, mean, and standard deviation of structural similarity indices computed between the target and every other time frame. Additionally, if not explicitly provided, `s2cloudless` probabilities for each timeframe are computed from reflectance values and used as inputs as well. The algorithm has been integrated into [eo-learn][5] Python library published under the MIT License on [GitHub][6]. 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 `s2cloudless` algorithm.

The ` InterSSIM` cloud masks were provided in a binary mode: 1 – "cloud" and 0 – "non-cloud".

2.7 LaSRC (NASA / University of Maryland)

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). LaSRC has been used extensively for multiple space-borne remote sensing instruments, such as MODIS (Vermote & Kotchenova, 2008), Visible Infrared Imaging Radiometer Suite (VIIRS) (Vermote et al., 2014), OLI (Vermote et al., 2016) and MSI (Doxani et al., 2018). LaSRC is the main algorithm for atmospheric correction of Landsat 8/OLI and Senitnel-2/MSI data for NASA's HLS product (Claverie et al., 2018), and has been extensively validated within the Atmospheric Correction Inter-comparison eXercise (ACIX) (Doxani et al., 2018).

Within the atmospheric correction process, LaSRC generates several quality assurance (QA) layers, including a cloud mask (Skakun et al., 2019; Vermote et al., 2016). The main metric for deriving a cloud mask is a per-pixel inversion residual error, which shows the goodness of aerosol optical thickness (AOT) estimation process:

$$r_{vis\_swir} = \sqrt{\frac{1}{3}\left(\left(\rho_S^{cb} - r_{cb,r}\rho_S^{r}\right)^2 + \left(\rho_S^{b} - r_{b,r}\rho_S^{r}\right)^2 + \left(\rho_S^{sw} - r_{sw,r}\rho_S^{r}\right)^2\right)},, \quad (1)$$

where ρ_S^{cb}, ρ_S^{b}, ρ_S^{r} and ρ_S^{sw} are the surface reflectance values in coastal blue (~0.440 µm), blue (~0.480 µm), red (~0.660 µm) and shortwave infrared bands (~2.1 µm), respectively (Landsat 8 bands 1, 2, 4 and 7 or Sentinel-2 bands 1, 2, 4 and 12), derived using AOT inverted from the red and blue; and $r_{cb,r}$, $r_{b,r}$ and $r_{sw,r}$ are ratios between red and blue and SWIR bands derived from MODIS and Multi-angle Imaging Spectroradiometer (MISR) and downscaled at Landsat 8 spatial resolution (Vermote et al., 2016). This residual metric (Eq. 1) is the main criterion for detecting thick clouds, since the latter will either prevent the AOT inversion process from convergence, or will drive the residual metric to high values. For both Landsat 8 and Sentinel-2, we used a threshold of 0.05 for the residual to identify

[3] https://medium.com/sentinel-hub/on-cloud-detection-with-multi-temporal-data-f64f9b8d59e5
[4] https://lightgbm.readthedocs.io/en/latest/
[5] https://eo-learn.readthedocs.io/en/latest/
[6] https://github.com/sentinel-hub/eo-learn

cloudy pixels. Pixels adjacent to clouds within 5 pixels are separately masked as being “adjacent to clouds”. For S2, a conservative threshold of 0.003 (reflectance units) was also 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.

2.8 MAJA (CNES/CESBIO)

MAJA is a comprehensive L2A processor applicable to optical Earth observation satellites, which perform repetitive observations at similar viewing angles. It is developed by CNES with CS-SI as a contractor since 2006; its methods were designed by the CESBIO laboratory, and it includes a few modules contributed by DLR. Its particularity is to use multi-temporal methods for several aspects of its processing. MAJA software is freely accessible to anyone and accessible as an open source software. MAJA’s cloud masks are produced operationally on some zones of the world and can be downloaded freely from https://theia.cnes.fr .An on-demand processing is also accessible using the maja-peps service : https://github.com/olivierhagolle/maja_peps.

MAJA's cloud and shadow detection methods include several tests, which use the multi-spectral and multi-temporal properties of surfaces, clouds, and shadows to classify the different types of pixels. They are described in detail in (Hagolle et al., 2010) and (Hagolle et al., 2017). The Sentinel-2 cloud masks obtained with MAJA are dilated using a buffer of 240 m, firstly to account for the parallax effects due to differences in observation angles between spectral bands, but also to account for the adjacency effects of clouds and for their fuzzy borders. MAJA aims at a good reliability for surface reflectance monitoring, its tests and threshold are therefore optimized to minimize cloud or cloud shadow omission, but still with a low commission error.

In CMIX, the cloud masks for Sentinel-2 were computed at 240 m resolution, to optimize the computation time, but this can prevent MAJA from detecting very small clouds. Since then, thanks to computing performances optimization, it is now possible to compute the clouds and shadows masks at 120 m with the same duration as for CMIX experiment, which should further improve MAJA's performances.

Due to the use of multi-temporal criteria, MAJA processes time series and for each single reference images, at least 10 images have to be processed by our team. Moreover, some reference images were acquired in the early phase of Sentinel-2A, with a revisit of 10 to 20 days, compared to the nominal revisit of 5 days for the Sentinel System. In order to limit the number of images to process specifically for CMIX and provide results in the nominal condition of the mission, we did not process the images acquired before July 2017.

MAJA has been intensively validated and some of its validation data sets (Baetens et al., 2019) were used in the CMIX experiment.

MAJA’s values “all clouds” (corresponding to bit 1) and “thinnest clouds” (corresponding to bit 4) were used as “cloud” class, whereas all others were used as a “non-cloud” class. Cloud shadows have therefore been included in the non-cloud class.

2.9 S2cloudless (Sinergise)

The `s2cloudless` is an automated cloud-detection algorithm[7] for Sentinel-2 imagery based on a gradient boosting algorithm LightGBM (https://lightgbm.readthedocs.io/en/latest/). The model was trained on a large training dataset with a global coverage. The algorithm is mono-temporal, does not take into account any spatial context, and 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 reflectance values of the following 10 bands: B01, B02, B04, B05, B08, B8A, B09, B10, B11, B12. The output of the algorithm is a cloud probability map. Users of the algorithm can convert the cloud probability map to a cloud mask by setting a threshold on the cloud probability map. The recommended value for the threshold is 0.4. 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 with a disk 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. The algorithm is published under the MIT License on [GitHub][8]. [Sentinel Hub][9] and [Google Earth Engine][10] provide precomputed `s2cloudless` cloud probability maps and masks to their users.

The `s2cloudless` cloud masks were provided in a binary mode: 1 – cloud and 0 – non-cloud.

2.10 Sen2cor (ESA / Telespazio France / DLR)

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 the two main modules: an atmospheric correction module (originally based on ATCOR from DLR) and a scene classification module (designed by Telespazio France) that provides a "scene classification map", which is used internally in Sen2Cor's atmospheric correction module to distinguish between cloudy, clear and water pixels.

Sen2Cor v.2.8 algorithm is used by the European Space Agency to generate the official Sentinel-2 Level-2A products within the Sentinel-2 ground segment.

Sen2Cor v2.8 software can be found on this page: https://step.esa.int/main/third-party-plugins-2/sen2cor/ , the code written in Python is open source.

Description of the algorithm[11]

Sen2Cor v2.8 cloud screening algorithm uses the reflective properties of scene features (L1C TOA reflectances). Potential cloudy pixels undergo a sequence of filtering based on spectral bands thresholds, ratios, and indexes computations (NDSI, NDVI). The result of each pixel test is a cloud probability (ranging from 0 for high confidence clear sky to 1 for high confidence cloudy). After each step, the cloud probability of a potentially cloudy pixel is updated by multiplying the current pixel cloud probability by the result of the test. Finally, the cloud probability of a pixel is the product of all the

[7] https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13?source=collection_category---4------3----------

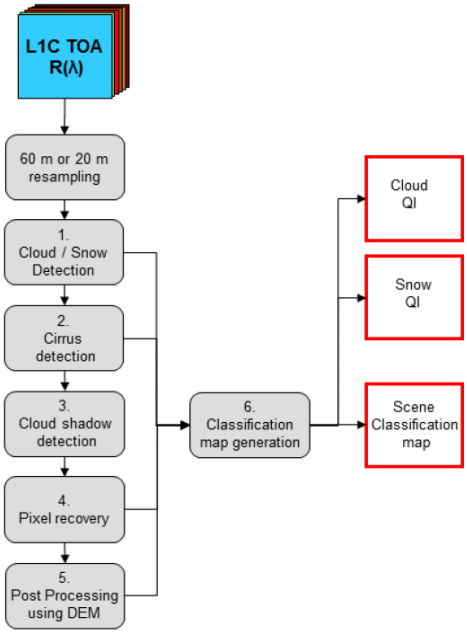
[8] https://github.com/sentinel-hub/sentinel2-cloud-detector

[9] https://medium.com/sentinel-hub/cloud-masks-at-your-service-6e5b2cb2ce8a

[10] https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_CLOUD_PROBABILITY

[11] https://sentinels.copernicus.eu/documents/247904/446933/Sentinel-2-Level-2A-Algorithm-Theoretical-Basis-Document-ATBD.pdf

individual tests. The sequential filtering stops when a test result set the pixel cloud probability to zero. The pixel is then considered to be high confidence clear sky in the cloud probability map and the pixel is classified to its corresponding class in the classification map (Figure 2). Each part of the different processing steps shown in Figure 2 are detailed in Sen2Cor ATBD available online.



Label	Classification
0	NO_DATA
1	SATURATED_OR_DEFECTIVE
2	DARK_AREA_PIXELS
3	CLOUD_SHADOWS
4	VEGETATION
5	NOT_VEGETATED
6	WATER
7	UNCLASSIFIED
8	CLOUD_MEDIUM_PROBABILITY
9	CLOUD_HIGH_PROBABILITY
10	THIN_CIRRUS
11	SNOW

Figure 2: Sen2Cor classification

The cloud shadow algorithm method is a combination of:

a) Radiometric approach: Dark areas and potential cloud shadows are identified by their spectral signature using S2 spectral bands B2, B3, B4, B8, B11 and B12.
b) Geometric approach: A mask of probable cloud shadows is derived using the final cloud mask, sun position and an a-priori distribution of top-cloud height
c) Final step: The final cloud shadow mask is obtained by multiplying the result of the radiometric branch by the result of the geometric branch. The result is a probabilistic cloud shadow mask.

SCL mask usage for CMIX

Sen2Cor v.2.8 has not been modified or particularly trained to fit the CMIX datasets definition for clouds. Non-cloud or shadow dilation is applied neither proposed, despite a prototype Sen2Cor version foresees this dilation for users interested in "clean" pixels.

L2A products directories have been cleaned to reduce the L2A product size and only the SCL mask is kept in the "../IMG_DATA/R20m" directory.

We provide hereafter some basic recommendations for SCL mask usage within CMIX:

- For **clear pixels land application** (e.g. vegetation/land monitoring) we usually suggest to use classes **"vegetation" and "not_vegetated"**. Water class can be self-understood, it includes marine and inland waters.
- For the cloud percentage computed as a unique figure for L2A product, the combination of **"cloud medium probability", "cloud high probability" and "thin_cirrus"** is used to derive the

L2A **global cloud percentage**. L2A surface reflectance and subsequent bio-geophysical parameters are expected to be impacted by cloud presence.

- Considering the distinction of **"ambiguous case",** the combination of **"cloud medium probability" and "thin_cirrus"** maybe more appropriate to classify those **semi-transparent** clouds. With some expected commission error as cloud medium probability could also catch some opaque clouds (e.g. cloud borders).

The following values are used for "cloud" class: Cloud medium probability (8), Cloud high probability (9), Thin cirrus (10). The following values are used for "non-cloud" class: Invalid (0), Saturated or defective (1), Dark area pixel (2), Cloud shadow (3), Vegetation (4), Not vegetated (5), Water (6), Unclassified (7), and Snow (11).

Auxiliary Data

Sen2Cor uses:

- a Digital Elevation Model (DEM): SRTM v4 or PlanetDEM
- ESA CCI Maps a priori information
    - CCI-LC Water Bodies Map (5 years) at 150 m
    - Land Cover Map v.2.0.7 (2015) at 300 m
        - urban class = 190
        - bare classes = 200, 201, 202
    - CCI-LC Snow Condition (7-day) at 500 m
- a snow climatology included in the installation package.

3 Validation

This section includes a detailed description of the validation datasets and the validation methods, and the results will be presented. During the exercise it became obvious that the validation results not only depend on the performance of the algorithms, but the performance of an algorithm also varies with the validation datasets. Therefore, a good insight into the validation datasets and their characteristics is extremely valuable for the interpretation of the results.

3.1 Validation datasets

The validation performed in the CMIX was based on existing Sentinel-2 and Landsat 8 cloud reference datasets. These datasets have been collected/generated for different purposes using different methodologies and cloud class nomenclatures. Some of the datasets are single-pixel collections (i.e. where a minimum mapping unit is a pixel), while others are the collections of connected pixel areas (polygons). In most of the datasets, pixels were classified manually; in others, the labelling process was semi-automatic with extensive manual checking afterwards.

The following Table 4 gives a brief overview of the used datasets and the different methodologies.

Table 4: Validation datasets overview

Dataset	Spatial domain	Level of automatization	Purpose	Thematic depth	Satellites	Spatial resolution	# scenes	Data Availability
Hollstein	Pixels collected by polygons (hundreds to thousands of pixels per polygon)	Manually selected and classified by an expert	Training and Validation	Shallow (6 classes)	S2	Polygons (20 m)	59	https://gitext.gfz-potsdam.de/EnMAP/sentinel2_manual_classification_clouds
Pixbox	Single pixels	Manually selected and classified by an expert	Validation	Very high (10 and more categories with multiple classes)	S2, L8	S2: 10 m L8: 30 m	S2: 29 L8: 11	https://zenodo.org/record/5036991#.YNrhAOhMGUk https://zenodo.org/record/5040271#.YNrvluhMGUl
L8 Biome	Classification of full Landsat 8 scenes	Manually classified by an expert	Training and Validation	Shallow (4 classes)	L8	30 m	L8: 96	https://landsat.usgs.gov/landsat-8-cloud-cover-assessment-validation-data
CESBIO	Pixel classification of a	Classification using an iterative and supervised	Validation	Shallow (6 classes)	S2	60 m	S2: 30	https://zenodo.org/record/146096

	complete scene	active learning method						1#.YFn3Ui2z00o
GSFC	Pixels collected by polygons (hundreds to thousands of pixels per polygon)	Manually selected and classified by an expert assisted by ground-based images of the sky	Validation	Shallow (4 classes)	L8, S2	Polygons (in vector format)	L8: 6 S2: 28	https://data.mendeley.com/datasets/r7tnvx7d9g/1

3.1.1 S2 Hollstein dataset (Hollstein et al. 2016)

The ”S2 Hollstein dataset” is a database of manually labeled Sentinel-2A spectra, which were used in the paper by (Hollstein et al., 2016). The database is currently hosted by the Environmental Mapping and Analysis Program (EnMAP) of Deutsches GeoForschungsZentrum GFZ. The data is available from the GFZ Git[12]. The collection was done based on the early Sentinel-2 products. These products consisted of a 290 km image divided into 100 km granules in UTM/WGS84 projection.

By means of different spectral tools, granule pixels were selected and classified into one of the following six classes:

Class	Coverage
cloud	opaque clouds
cirrus	cirrus and vapor trails
snow	snow and ice
shadow	shadows from clouds, cirrus, mountains, buildings, etc
water	lakes, rivers, seas
clear-sky	remaining: crops, mountains, urban, etc

Spectral tools include *false-color composites*, *image enhancements* and *graphical visualization of spectra*. The aim is to create highly heterogeneous classes with a balanced number of pixels.

Figure 3 shows the coast of Fiji in two different composites: (a) bands 4/3/2 and (b) bands 8a/3/2. Colored polygons represent four different classes. Cyan, yellow, dark blue and green colors stand for water, shadow, cloud and clear-sky pixels

[12] https://gitext.gfz-potsdam.de/EnMAP/sentinel2_manual_classification_clouds

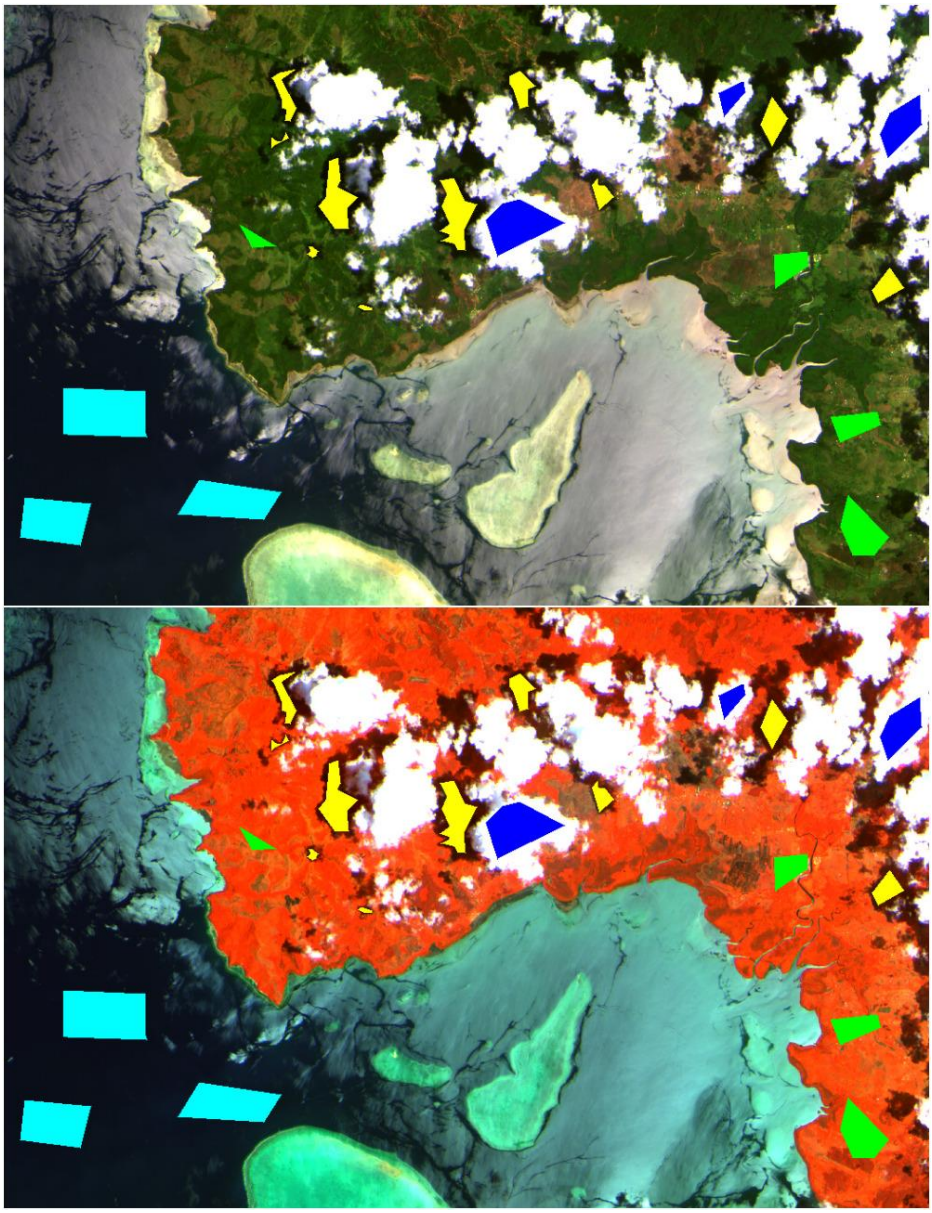


Figure 3: Classification example from S2 Hollstein dataset.

The dataset consists of a total of N=5647725 pixels. Pixel information is saved into different tables in the HDF5 file.

Relative to Sentinel-2 spatial and spectral resolutions:

- **band** associates a band position with its label
- further band descriptions can be found in
    - **bandwidth_nm**,
    - **central_wavelength_nm** and
    - **spatial_sampling_m**

Relative to the classes:

- **classes** (1xN table) includes the class id to which each pixel in the dataset is associated
- **class_ids** describes the id associated to each class that appears in **class_names**

Relative to the spectra:

- **spectra** (13xN table) collects the spectral values of each pixel. Sentinel-2 instrument samples 13 spectral bands.

Relative to the image metadata:

- **latitude** and **longitude** gather pixel coordinates
- each pixel is located in a **granule_id**, where several granules correspond to an image associated with a **product_id**
- the same product will share the sensing date -**date**-, four different sampling angles -**sun_azimuth_angle**, **sun_zenith_angle**, **viewing_azimuth_angle**, **viewing_zenith_angle**- and the geographical location -**continent** and **country**.

Preparation of the dataset for the CMIX

The collection of the S2 Hollstein dataset was done when only the old (multi-granule) products were distributed by ESA. The old product name includes sensing and creation date, as well as the relative orbit number of the image are stored inside the HDF5 database. Additionally, the granule IDs are stored. By using the stored information, we tried to identify the correct newly formatted (granule based) Sentinel-2 products. 60 products have been identified of which 59 were available from the Copernicus Open Access Hub[13]. These products have been provided to the participants.

Somehow, there is a varying number of products given for the collection of the dataset. On multiple grey resources throughout the internet a number of 108 products is listed, where in the GFZ Git repository[14] a list with 98 products is given. As stated above, only 60 individual products have been identified within the database.

3.1.2 S2/L8 Pixbox dataset

The overarching idea of Pixbox is a quantitative assessment of the quality of a pixel classification which is the result of 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. Such pixel classification attributes are typically used to further guide higher level processing.

The Pixbox method comprises 2 elements:

- A **Reference Data Set**: trained experienced expert(s) manually classify pixels of an image sensor into a pre-defined detailed set of classes. These are typically different cloud transparencies, cloud shadow, condition of underlying surface ("semi-transparent clouds over snow", "clouds over bright scattering water"). The collected dataset includes several 10-thousands of pixels because it has to be representative for all classes, and for various observation and environmental conditions, such as climate zones, sun illumination etc. Quality control of the collected pixels is important in order to detect misclassifications and systematic errors. An auto-associative neural network is trained for this purpose.
- **Analysis**: the analysis compares the reference dataset with one (or more) automated classification procedures. This includes at least a confusion matrix of the directly comparable classes. However, the reference dataset includes a very fine granularity of classes, which is usually not provided by the automated procedure. This allows to detect systematic weaknesses of the automated procedure and to formulate recommendations for improvements.

Figure 4 shows the Pixbox user interface. The expert sets the classes for multiple pixel characteristic categories (surface properties) and starts collecting matching pixels, then classes are changed, and

[13] https://scihub.copernicus.eu/

[14] https://gitext.gfz-potsdam.de/EnMAP/sentinel2_manual_classification_clouds/-/blob/master/code/list_scenes.py

new pixels are collected, and so on. Based on the multiple categories, and classes per category, a very high thematic detail is reached.

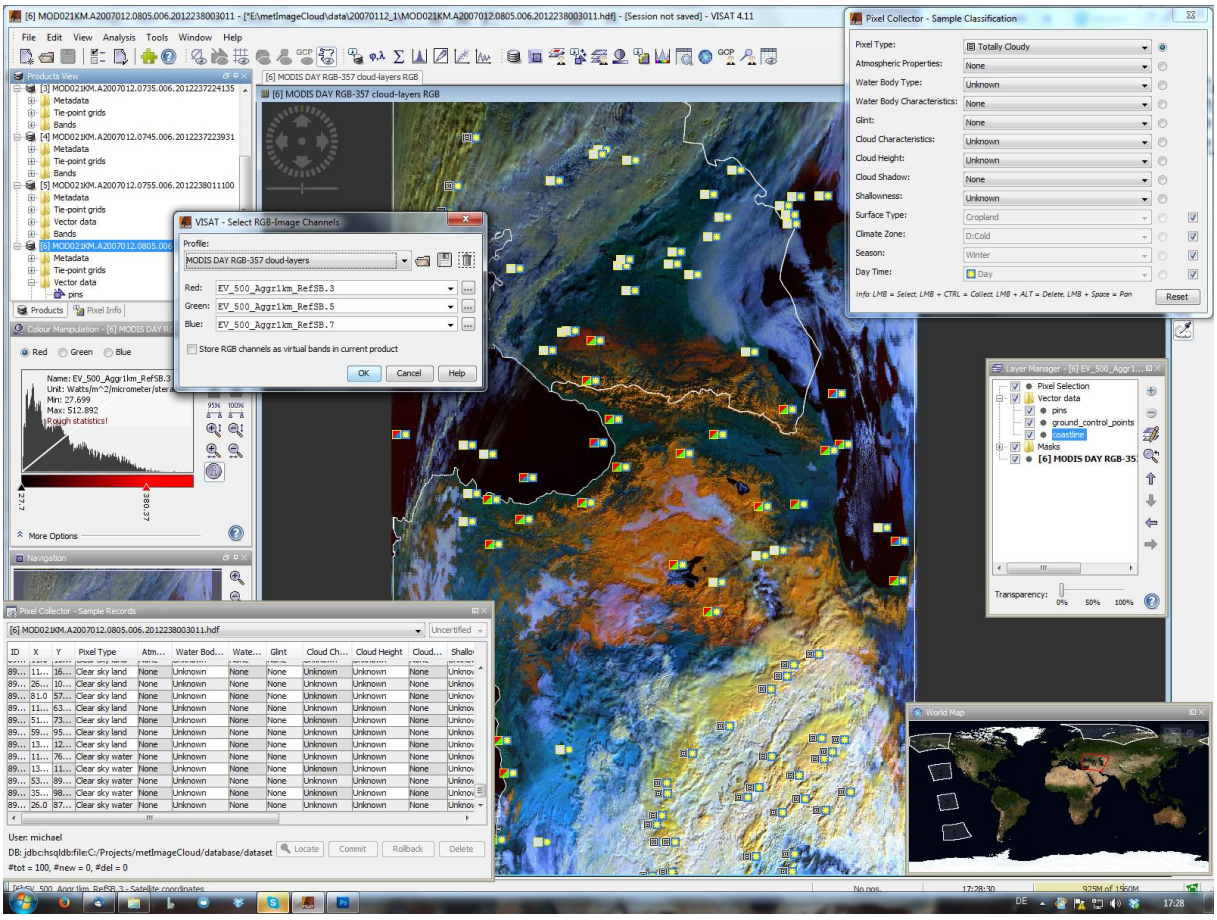


Figure 4: Pixbox user interface

For CMIX a subset of two pre-existing Pixbox pixel collections were used. Details are given in the following sections.

Preparation

No data preparation was needed as all products are stored with the database during collection and are archived. These archived products have been provided to the participants.

3.1.2.1 S2 Pixbox dataset

The Sentinel-2 pixel collection contains 17,351 pixels manually collected from 29 Sentinel-2 A & B Level 1C products. This collection is a subset of a collection made in 2018 containing 54,000 pixels from 87 products. The used dataset is spatially, temporally, and thematically evenly distributed. Figure 5 and Figure 6 give a small example of the thematic and spatial distribution of the dataset. Figure 7 and Figure 8 show in detail the collected categories and classes of the dataset.

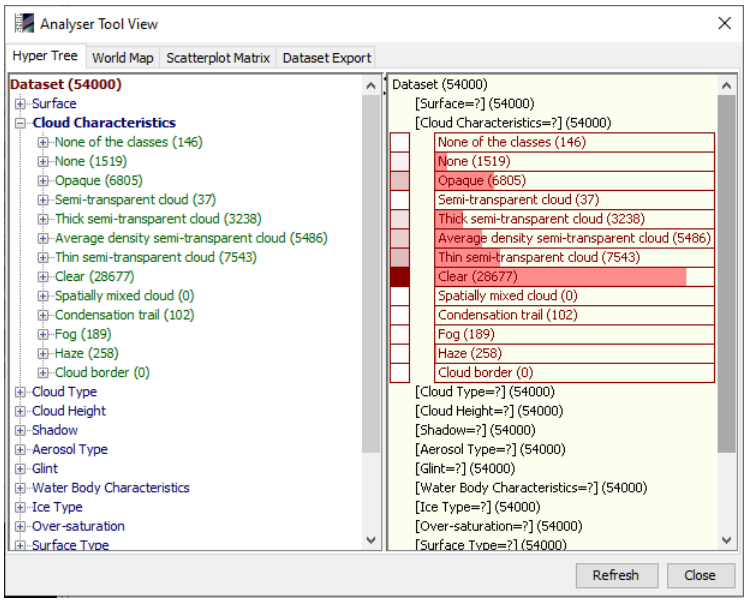


Figure 5: Example of thematic categories and classes of S2 Pixbox collection

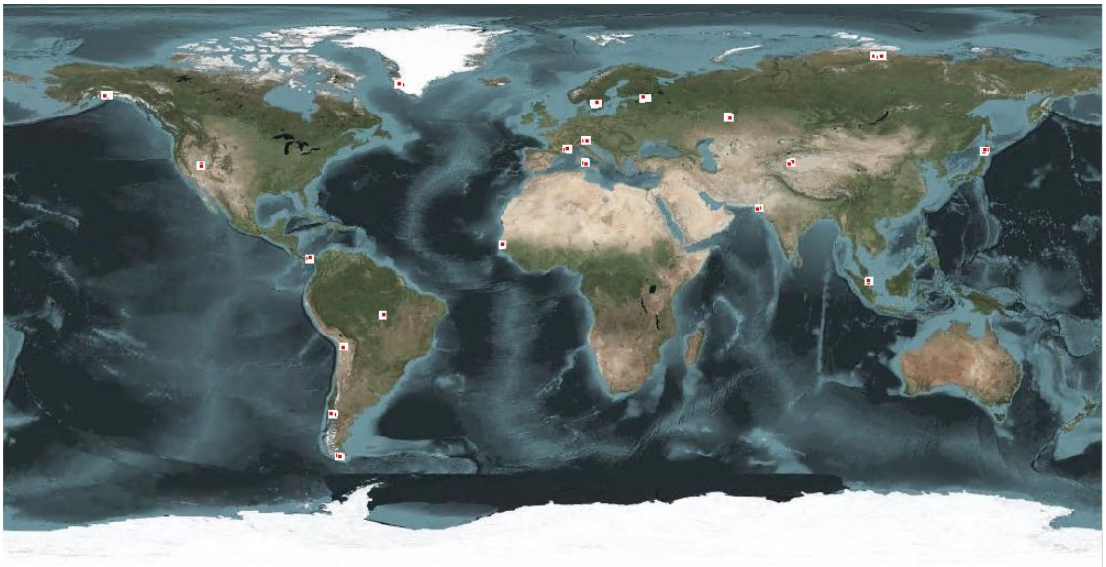


Figure 6: Spatial distribution of S2 products used for Pixbox collection

CLOUD_CHARACTERISTICS_ID	
ID	Name
0	None
1	None of the classes
2	Clear
3	Opaque
4	Semi-transparent cloud
5	Thick semi-transparent cloud
6	Average density semi-transparent cloud
7	Thin semi-transparent cloud
8	Spatially mixed cloud

CLOUD_TYPE_ID	
ID	Name
0	None
1	None of the classe
2	Stratus
3	Cumulus
4	Convective Cloud
5	Cirrus

CLOUD_HEIGHT_ID	
ID	Name
0	None
1	None of the clas
2	low (<3km)
3	middle (3-6km)
4	high (>6km)

SHADOW_ID	
ID	Name
0	None
1	None of the classes
2	Topographic shadow (incl tree & buildings)
3	Cloud shadow
4	Shadow above cloud

GLINT_ID	
ID	Name
0	None
1	None of the classes
2	Glint

WATER_BODY_CHARACTERISTICS_ID	
ID	Name
0	None
1	None of the classes
2	Bright turbid water (blue or brown)
3	Cocolithophorides
4	Floating Cyanobacteria bloom
5	Floating vegetation
6	Dark water
7	Wave breaking
8	Algae

Figure 7: Categories and classes of the S2 Pixbox collection - part 1

SURFACE_ID	
ID	Name
0	None
1	None of the classes
2	Land
3	Mountain
4	Urban
5	Agriculture
6	Desert
7	Dry/Salt Lake
8	Intertidal Area
9	Spatially mixed land/water
10	Water
11	Lake
12	River
13	Costal
14	Open ocean
15	Spatially mixed water/snow-ice
16	Snow-ice
17	Spatially mixed land/snow-ice

ICE_TYPE_ID	
ID	Name
0	None
1	None of the classes
2	Snow
3	Ice
4	Floating sea ice
5	Brash sea ice
6	Sea ice edge
7	Fractured sea ice
8	Flooded sea ice
9	Fast sea ice
10	Pack sea ice

AEROSOL_TYPE_ID	
ID	Name
0	None
1	None of the classes
2	Desert dust
3	Smoke
4	Volcanic eruption
5	Dust storm
6	Smog

OVERSATURATION_ID	
ID	Name
0	None
1	None of the classes
2	Oversaturation

SHALLOWNESS_ID	
ID	Name
0	None
1	None of the classes
2	Benthic sediment
3	Benthic vegetation
4	Optically deep

CLIMATE_ZONE_ID	
ID	Name
0	None
1	None of the classes
2	A:Tropical
3	B:Dry
4	C:Temperate
5	D:Cold
6	E:Polar

SURFACE_TYPE_ID	
ID	Name
0	Evergreen Needleleaf Forest
1	Barren/Desert
2	Permanent Snow/Ice
3	Crop/Natural Veg, Mosaic
4	Urban
5	Cropland
6	Permenant Wetland
7	Grassland
8	Water Bodies
9	Savanna
10	Open Shrubland
11	Closed Shrubland
12	Mixed Deciduous Forest
13	Deciduous Broadleaf Forest
14	Deciduous Needleleaf Forest
15	Evergreen Broadleaf Forest
16	Woody Savanna
17	Tundra

SEASON_ID	
ID	Name
0	Spring
1	Summer
2	Autumn
3	Winter

DAY_TIME_ID	
ID	Name
0	Day
1	Night
2	Twilight

Figure 8: Categories and classes of the S2 Pixbox collection - part 2

3.1.2.2 L8 PixBox dataset

The Landsat 8 pixel collection contains 18,830 pixels manually collected from 11 Landsat-8 Level 1 products. This collection is a subset of a collection made in 2015 containing 37,000 pixels from 21 products. The used dataset is temporally and thematically evenly distributed. Spatially it is focused on coastal areas, mainly in Europe. Figure 9 and Figure 10 give a small example of the thematic and spatial distribution of the dataset. Figure 11 and Figure 12 show in detail the collected categories and classes of the dataset.

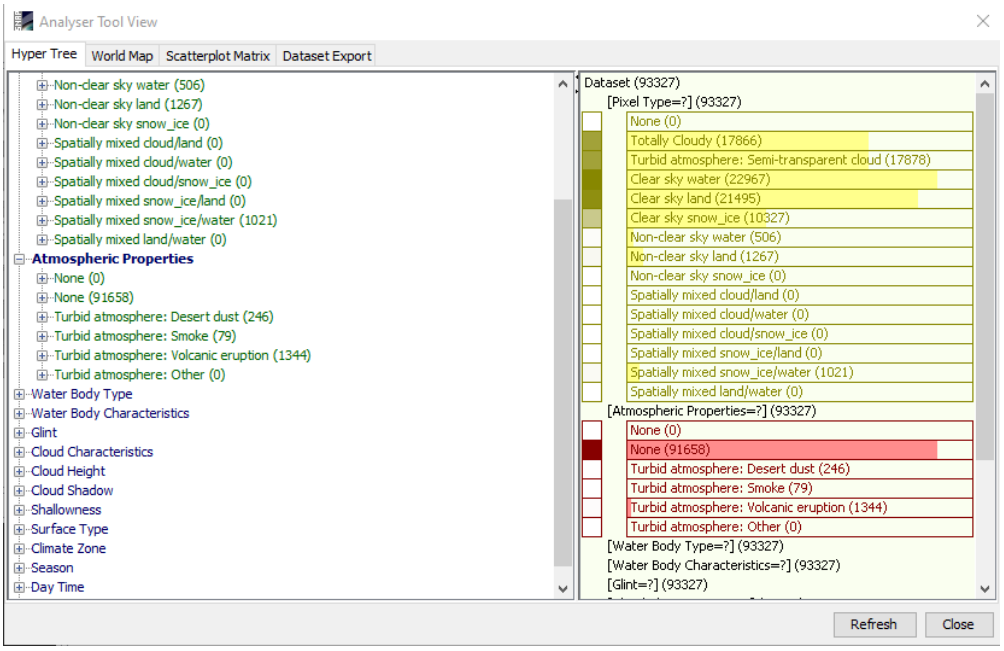


Figure 9: Example of thematic categories and classes of L8 Pixbox collection

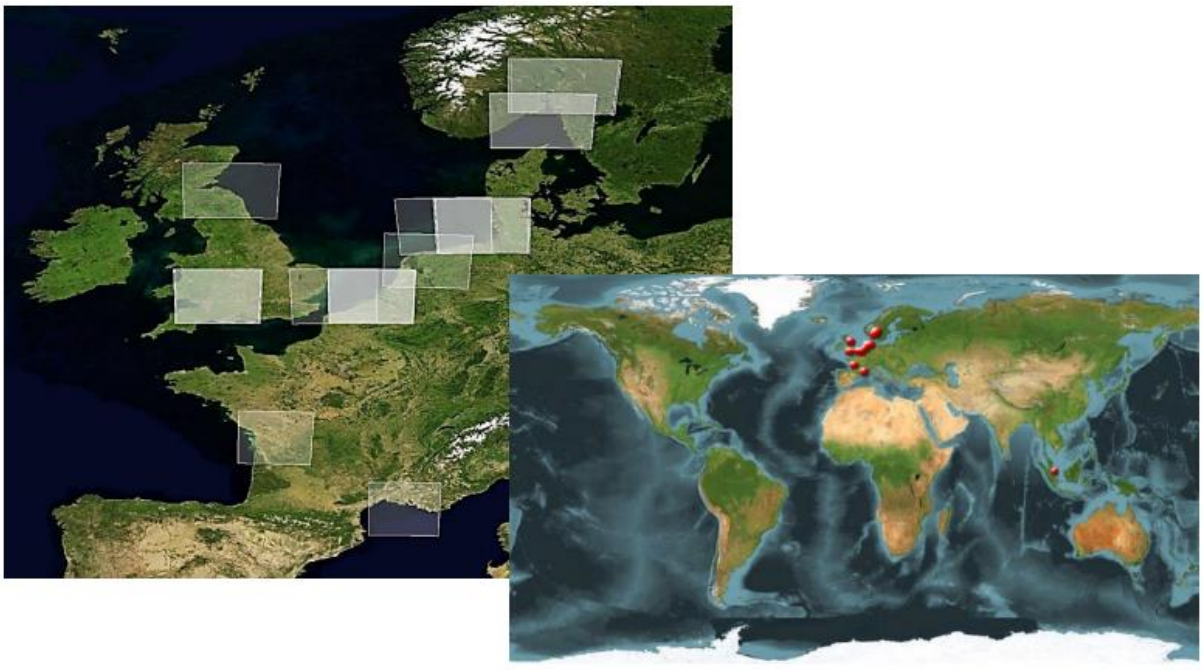


Figure 10: Spatial distribution of L8 products used for Pixbox collection

CLOUD_CHARACTERISTICS_ID	
ID	Name
0	None
1	None of the classes
2	Stratus
3	Cumulus
4	Convective Cloud
5	Cirrus

CLOUD_HEIGHT_ID	
ID	Name
0	None
1	None of the clas
2	low (<3km)
3	middle (3-6km)
4	high (>6km)

SHADOW_ID	
ID	Name
0	None
1	None of the classes
2	Cloud shadow

GLINT_ID	
ID	Name
0	None
1	None of the classes
2	Glint

WATER_BODY_CHARACTERISTICS_ID	
ID	Name
0	None
1	None of the classes
2	Bright turbid water (blue or brown)
3	Cocolithophorides
4	Floating Cyanobacteria bloom
5	Floating vegetation
6	Dark water
7	Wave breaking
8	Algae

WATER_BODY_TYPE_ID	
ID	
0	None
1	None of the classes
2	Snow
3	Ice
4	Bright turbid water (blue or brown)
5	Cocolithophorides
6	Floating Cyanobacteria bloom
7	Floating vegetation
8	Dark water

Figure 11: Categories and classes of the L8 Pixbox collection - part 1

PIXEL_SURFACE_TYPE_ID	
ID	Name
0	None
1	None of the classes
2	Totally Cloudy
3	Turbid atmosphere: Semi-transparent cloud
4	Clear sky water
5	Clear sky land
6	Clear sky snow_ice
7	Non-clear sky water
8	Non-clear sky land
9	Non-clear sky snow_ice
10	Spatially mixed cloud/land
11	Spatially mixed cloud/water
12	Spatially mixed cloud/snow_ice
13	Spatially mixed snow_ice/land
14	Spatially mixed snow_ice/water
15	Spatially mixed land/water

ATMOSPHERIC_PROPERTIES_ID	
ID	Name
0	None
1	None of the classes
2	Turbid atmosphere: Desert dust
3	Turbid atmosphere: Smoke
4	Turbid atmosphere: Volcanic eruption
5	Turbid atmosphere: Other

SHALLOWNESS_ID	
ID	Name
0	None
1	None of the classes
2	Benthic sediment
3	Benthic vegetation
4	Optically deep

CLIMATE_ZONE_ID	
ID	Name
0	None
1	None of the classes
2	A:Tropical
3	B:Dry
4	C:Temperate
5	D:Cold
6	E:Polar

SURFACE_TYPE_ID	
ID	Name
0	Evergreen Needleleaf Forest
1	Barren/Desert
2	Permanent Snow/Ice
3	Crop/Natural Veg, Mosaic
4	Urban
5	Cropland
6	Permenant Wetland
7	Grassland
8	Water Bodies
9	Savanna
10	Open Shrubland
11	Closed Shrubland
12	Mixed Deciduous Forest
13	Deciduous Broadleaf Forest
14	Deciduous Needleleaf Forest
15	Evergreen Broadleaf Forest
16	Woody Savanna
17	Tundra

SEASON_ID	
ID	Name
0	Spring
1	Summer
2	Autumn
3	Winter

DAY_TIME_ID	
ID	Name
0	Day
1	Night
2	Twilight

Figure 12: Categories and classes of the L8 Pixbox collection - part 2

3.1.3 S2/L8 GSFC

Cloud reference data were collected over the NASA GSFC (Figure 13). The area is quite heterogenous with major land cover classes being forest (~52%) and impervious surfaces (31%) with patches of natural vegetation and cultivated areas (totaling 17%). NASA GSFC also has an AERONET station (Holben et al., 1998), which provides aerosol optical thickness (AOT) and water vapor, and one of the sites used in ACIX-I/ACIX-II. 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 (Figure 14) as well as seasonal cycles.

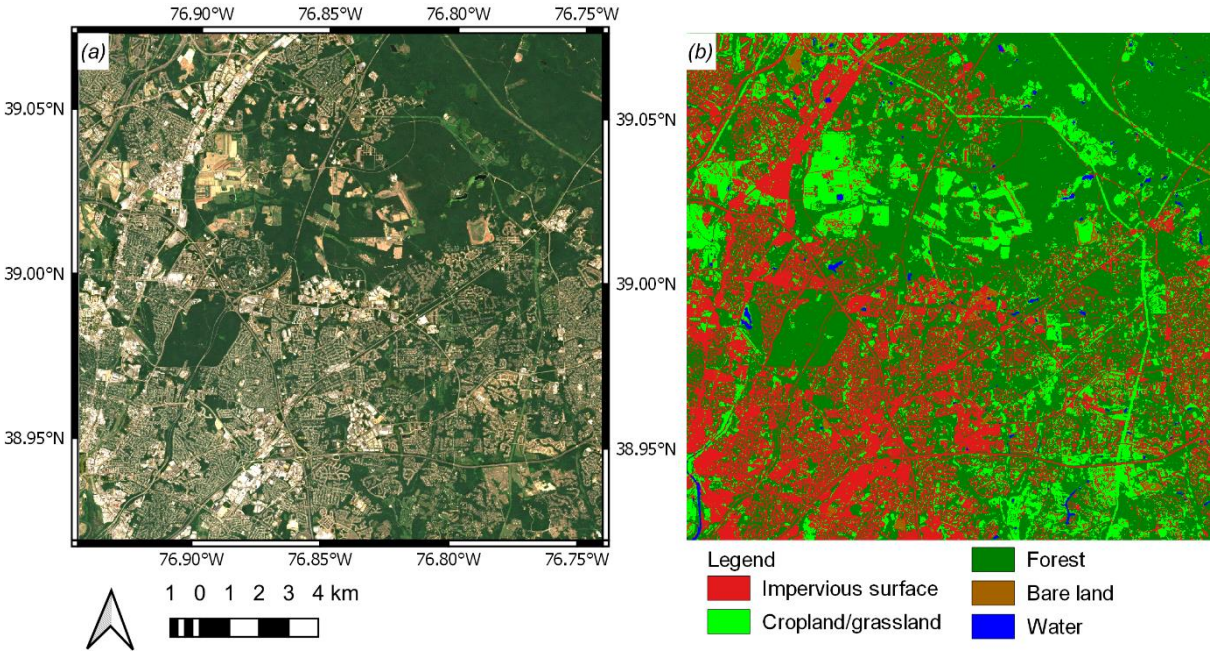


Figure 13: Overview of the area over the NASA Goddard Space Flight Center (GSFC) with the corresponding land cover map (b). Left panel (a) shows a Sentinel-2A image acquired on July 10, 2020. Shown is the true color combination of surface reflectance values in spectral bands B04 (red), B03 (green) and B02 (blue) derived from LaSRC (Vermote et al., 2016) and stretched from 0 to 0.15 (in reflectance units)

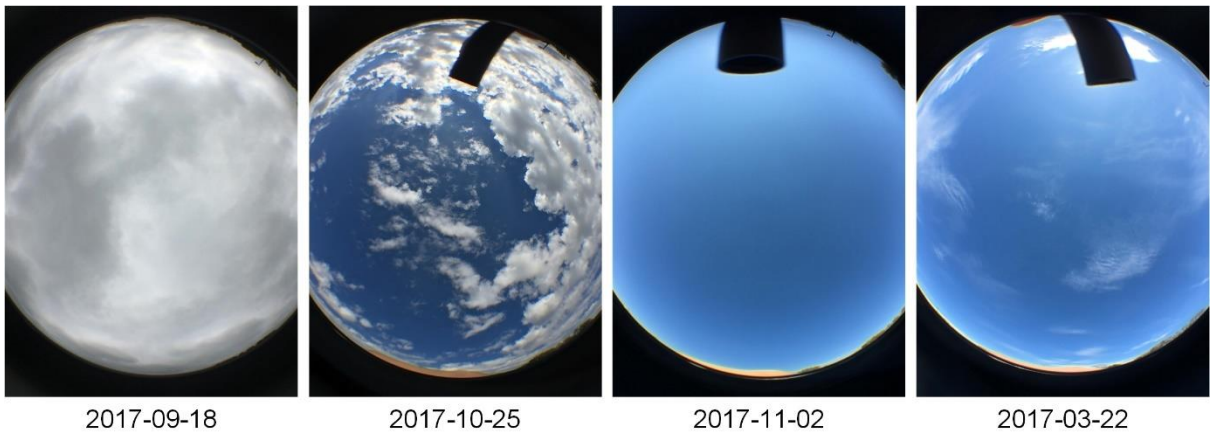


Figure 14: Ground-based images of the sky under various cloud conditions. The dark object at the top of the images was used to mask the Sun and reduce sun glare on the camera lens

Ground-based images were used to assist with manual labelling of clouds in Landsat 8 and Sentinel-2 imagery. In order to facilitate the labeling process, some of the ground-based images were manually geo-referenced to satellite imagery by manually selecting control points (CPs) using the shape of clouds and referencing a ground-based image to the satellite one. On average, 20-30 CPs were needed for a scene, and a second order polynomial function was used to transform the ground-based photo to the satellite one. For completely clear or overcast scenes, geo-referencing of ground-based photos was not necessary. For clear days or days with minimal cloud cover, we also checked estimated AOT values from the Aeronet station in order to not mislabel potential thin clouds.

Labeling of satellite imagery was manually performed into cloud, thin cloud, shadows and clear classes. To facilitate the labelling process, Sentinel-2 and Landsat 8 images were presented in various spectral combination including true color (red-green-blue) and false color (NIR-red-green, SWIR1-NIR-red), and using a cirrus band (at 1.38 μm). When defining cloud polygons, the boundary between clouds and clear classes was omitted (~1-2 pixels). There are several reasons for that. First, it is practically impossible to accurately define the boundary, especially at medium spatial resolution (10-30 m), which would contain a lot of "mixed" pixels. Second, uncertainty of cloud boundaries substantially increases for Sentinel-2 images, because of a parallax (Skakun et al., 2018), which introduces a shift of clouds in different spectral bands. MSI is designed in such a way that detectors, responsible for different spectral bands, are shifted against each other (Gascon et al., 2017) and therefore acquire images at slightly different angles, which introduces a parallax. That parallax is being corrected during image pre-processing, so images acquired at different wavelength will be aligned; however, these pre-processing routines do not correct shifts for moving and/or high-altitude objects.

Figure 15 shows examples of labeling Sentinel-2A scenes into reference classes. The resulting reference data are provided in the form of vector polygons that can be rasterized for the target spatial resolution of the cloud masking algorithm. The detailed description of the GSFC dataset is given in Skakun et al. (2021).

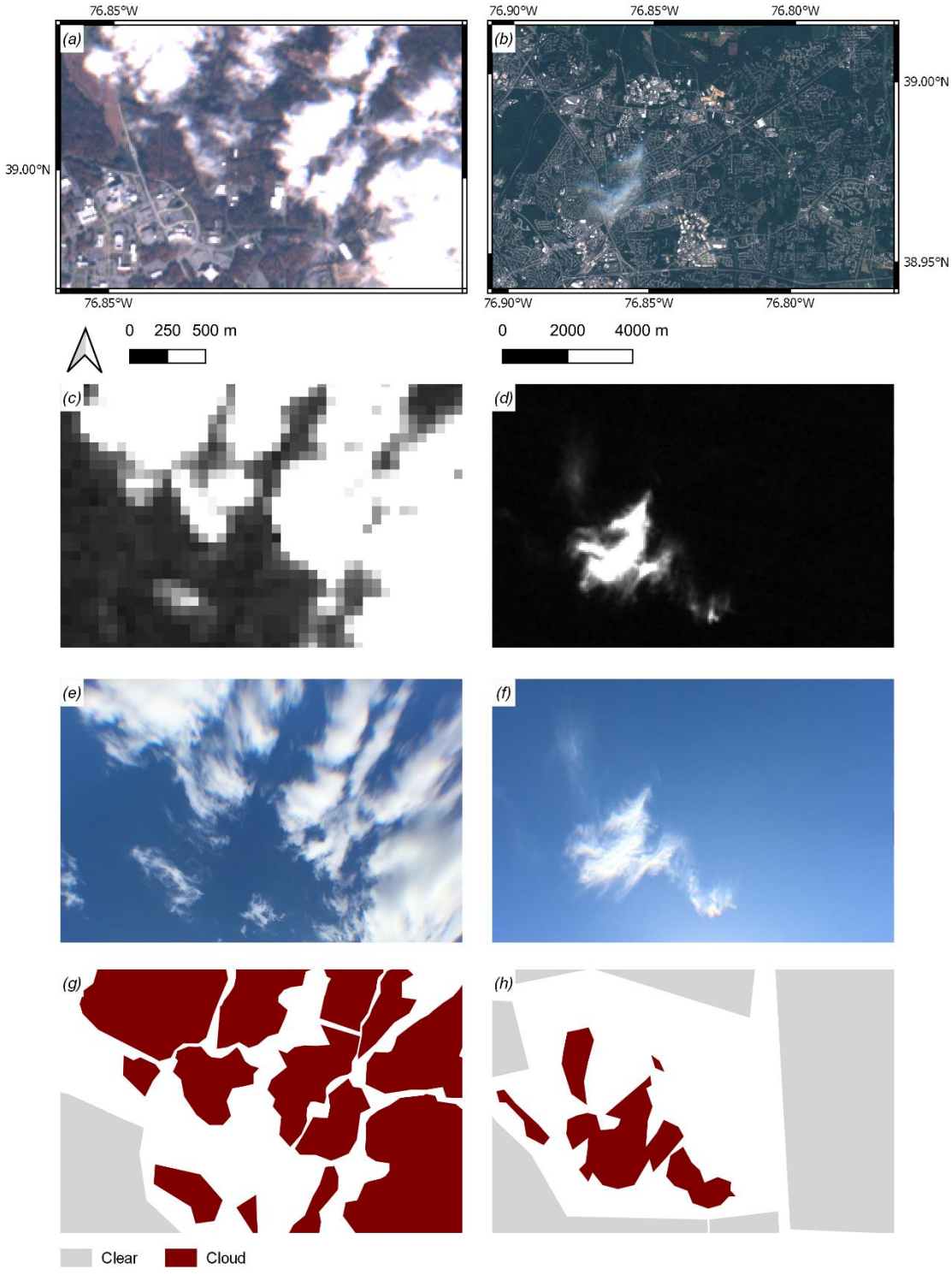


Figure 15: True color combination of TOA reflectance (B04-B03-B02 stretched from 0 to 0.25) of Sentinel-2A image acquired on August 9, 2018 (a) and September 23, 2017 (b). Corresponding cirrus bands (B10) stretched from 0.005 to 0.020 (c) and (d). Geo-referenced ground-based image of the sky during Sentinel-2A overpass (e) and (f). Reference clouds masks (g) and (h). From Skakun et al. (2021)

3.1.4 L8 Biome

The "L8 Biome" cloud validation dataset[15] consists of 96 Landsat 8 scenes (Figure 16), which were selected using a semi-random sampling by biomes (Foga et al., 2017). These biomes 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 into the following classes: clear, thin cloud, cloud, and cloud shadow. It should be noted that no specific threshold was used to detect thin 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 given in Foga et al. (2017).

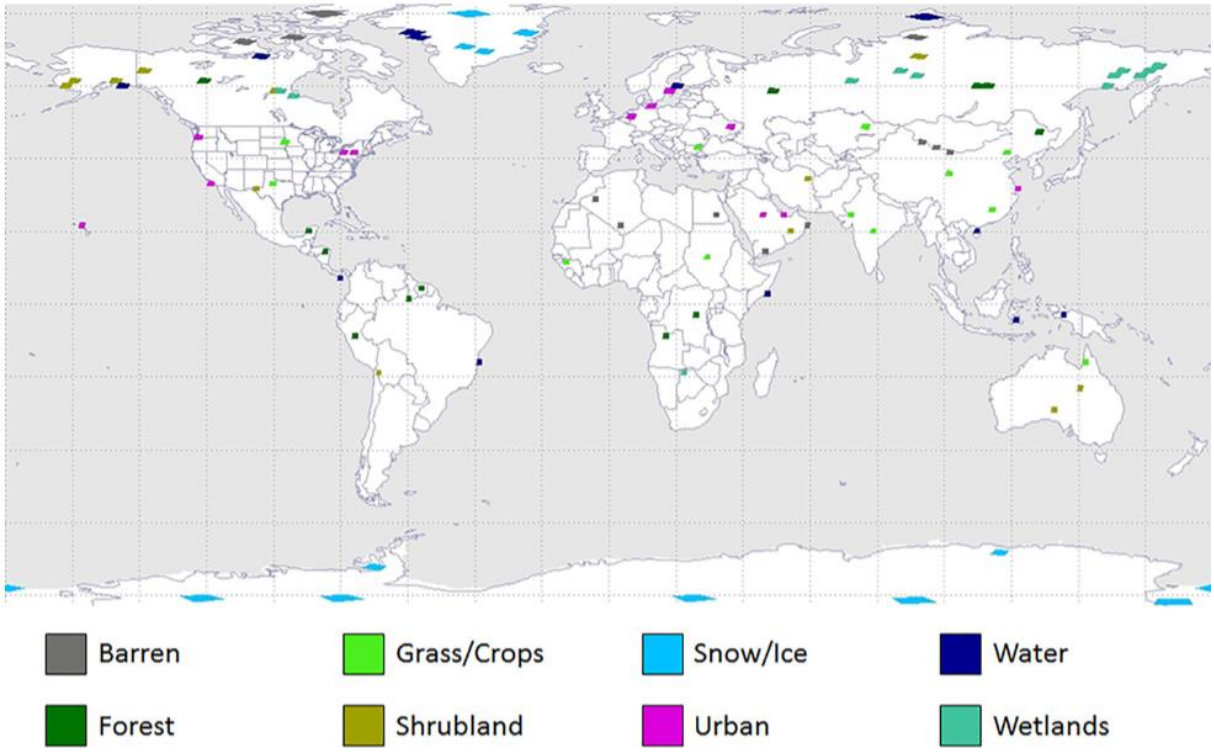


Figure 16: Global distribution of the 96 unique Landsat 8 Cloud Cover Assessment (CCA) scenes, sorted by International Geosphere-Biosphere Programme (IGBP) biome. Twelve scenes were selected for each of the eight biomes. From Foga et al. (2017)

[15]L8 Biome Cloud Validation Masks. (2016). U.S. Geological Survey, data release. [Online]. Available: http://doi.org/10.5066/F7251GDH.

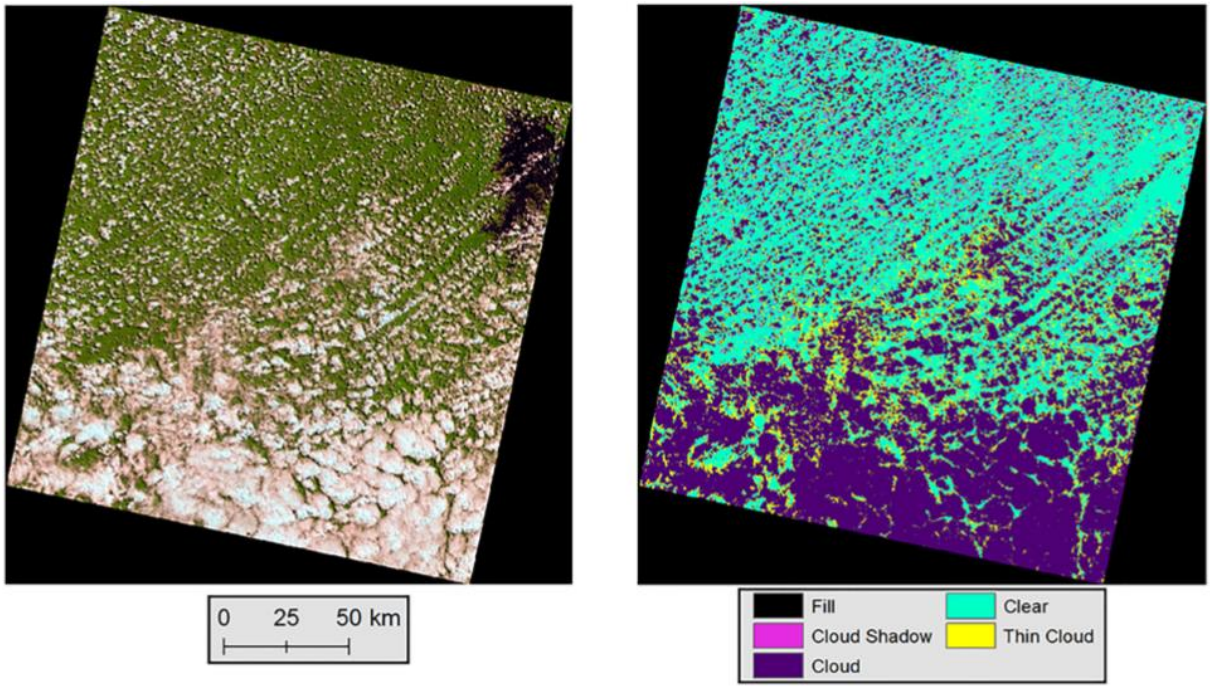


Figure 17: Left: Landsat 8 Operational Land Imager (OLI) scene used for cloud and cloud shadow mask digitization, acquired over WRS-2 Path 229, Row 57 on 21 May 2014, displayed as a false color composite (bands 6, 5, and 4, respectively). Right: The final "L8 Biome" cloud mask product. From Foga et al. (2017)

Several experimental scenarios were run with the L8Biome dataset, including all validation images (with some processors such as ATCOR and LaSRC not processing all of the images because of polar coordinates and snow cover); removing snow scenes from the validation datasets, so the algorithms will be compared against the same set of scenes/pixels; and, finally, removing thin clouds from the reference dataset, so the effect of thin clouds (which are more subjective in identification compared to thick clouds in L8Biome) can be explored.

3.1.5 S2 CESBIO dataset

The "Sentinel-2 reference cloud masks generated by an active learning method" (Baetens & Hagolle, 2018), from herein after called "S2 CESBIO dataset", provides a reference cloud mask data set for 38 Sentinel-2 scenes.

These reference masks were created with the ALCD tool, developed by Louis Baetens, under the direction of Olivier Hagolle at CESBIO/CNES (Baetens et al., 2019), to validate the cloud masks generated by the MAJA software (Hagolle et al., 2010).

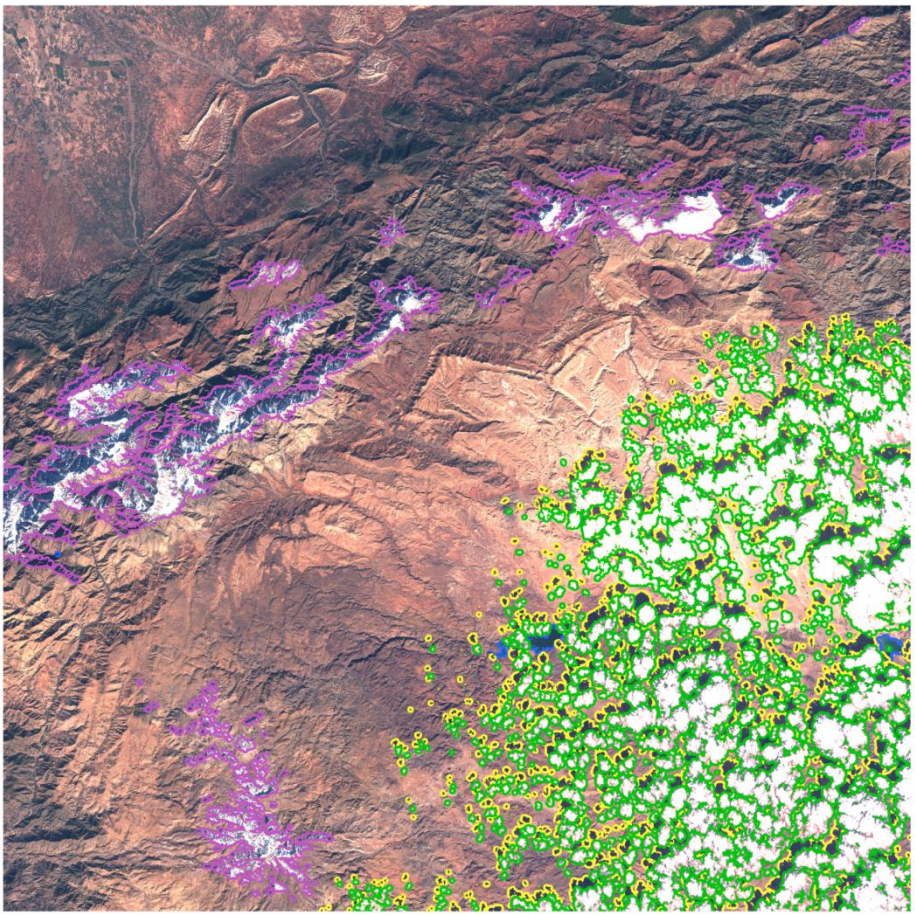


Figure 18: one of the CESBIO reference images with the contours of the masks overlaid (green: clouds, yellow: cloud shadows, pink: snow, and blue: water).

The active learning method consists in a heavily manually supervised classification method. The cloud expert selects and labels a large number of pixels, uses them to generate a first classification using a random forest classifier. This classification is manually controlled by the expert, and new pixels are added to the learning data base either to correct errors in the initial classification, or to provide additional information where the algorithm confidence is low. A new classification is then performed, and this sequence is iterated usually four to five times until no error is spotted by the expert. At each step a proportion of the learning data base is not used for the learning, but to obtain a confusion matrix to control the results.

The input features for the active learning classification, as well as for the the labelling by the expert, are all the bands of the image to classify, and a reference cloud free image acquired a few days apart. Thanks to this, the expert can control that the suspected clouds are real clouds, and the active learning classification is more efficient, as the problem is much simpler with multi-temporal detection than with only spectral tests. While MAJA uses reference cloud free data acquired before the image to classify, in ALCD, only reference images acquired after the image were used to classify the image in order to maximize the independence of results.

The CESBIO reference masks have been validated using a 10-fold cross validation, which led to an overall accuracy of 98.8%, and were compared to the Hollstein reference dataset for a set of 7 scenes, and the intersection of both datasets agreed to 98.9%. The differences were analyzed visually and traced back to errors in the Hollstein data base (such as those shown in section 3.2.1), or to a different definition of shadows, which in Hollstein may correspond to terrain or cloud shadows. After exclusion

of the terrain shadows, the agreement was 99.7%. For details on the validation, see (Baetens et al., 2019).

The data directories contain the following S2 data:

- The `Reference_dataset` directory contains 31 scenes selected in 2017 or 2018.
- The `Hollstein` directory contains 7 scenes that were used to validate the ALCD tool by comparison to manually generated reference images kindly provided by Hollstein et al. (2016).

For each product listed in the "reference_data" directory a set of files exists. The following files have been used for this exercise:

- `classification_map.tif` --- the main product, which is the classified scene. 7 classes are available. Each one is represented with a different integer.
    - 0: no_data.
    - 1: not used.
    - 2: low clouds.
    - 3: high clouds.
    - 4: cloud shadows.
    - 5: land.
    - 6: water.
    - 7: snow.
- `confidence_enhanced.tif` --- enhanced confidence map of the classification. The values are between 0 and 255 .
- `Samples/` --- this directory contains all the shapefiles of the training and validation database, one per class.

Preparation

For CMIX only 31 scenes from the "reference_dataset" have been used, as the data from the "Hollstein" directory is already part of the S2 Hollstein dataset described in section 3.1.1. The collection of the CESBIO dataset was done mostly on newly formatted (granule based) products. Nevertheless, as one scene is shared with the Hollstein dataset and thus is still in the old format, this specific product was used in the old format. The old product name includes sensing and creation date, as well as the relative orbit number of the image which are stored inside database. Additionally, the granule IDs are stored. By using the stored information, the correct newly formatted (granule based) Sentinel-2 products have been identified. The 31 products from the reference dataset have been acquired through the Copernicus Open Access Hub[16]. These products have then been provided to the participants.

[16] https://scihub.copernicus.eu/

3.2 Validation datasets strength and weakness analysis

Each validation dataset has its strength and weaknesses. Especially as all datasets have not specifically been designed for the CMIX. In this section we will discuss the strength and weaknesses of the single datasets based on the requirement definitions for pixel-based validation given in section 3.3.1. Table 5 gives a brief overview of the strength and weaknesses of all used validation datasets.

Table 5: Validation dataset strength and weakness overview table

Dataset	Strength	Weakness
Hollstein	Manual classification of polygons.	Slight lack of sample quality (some false classified samples) Low level of detail
PixBox	High level of detail High level of classification precision Global coverage with stratified sampling (S2)	Single pixel, thus a comparably small dataset Based on expert knowledge (could be biased)
GSFC	Assisted with ground-based imagery Over the same territory – can be used for consistency analysis	Limited field of view and coverage Surfaces classes limited to location of sky camera Tendency to classify very thin cirrus as cloudy that may have only little effect on surface reflectance
L8Biome	Global coverage with stratified sampling All pixels in the images are manually classified	Subjectivity of cloud definition
CESBIO	All pixels in the images are classified using a Machine Learning approach	Dependence on the labeled data Classification error due to automatic classified data. Based on expert knowledge (could be biased)

3.2.1 Hollstein dataset

Before analyzing strengths and weaknesses of the Hollstein dataset, its purpose needs to be considered. The main purpose of the dataset was to be a training dataset for a classifier. This means, it was not meant to be used for validation. As the authors themselves have used the data for validation during training, by splitting the database randomly. But they point out, that this kind of validation is not comparable with any standard EO validation approach, where the “truth” is either real ground truth or simply a source with lower uncertainties.

Keeping the initial purpose of the dataset in mind, following strengths and weaknesses are based on the objective to use the dataset as a validation source.

The strength of the Hollstein dataset is the selection of products, which has a good distribution of solar zenith angles, seasons, land cover and cloud conditions (not evaluating the selected samples, but the products) but missing very high latitudes (arctic and antarctica).

The weaknesses of the dataset are multiple quality issues of the dataset concerning, purpose of the dataset, resolution, maintenance, format, product processing baseline, data correlation, and sample quality.

Resolution: The data was collected on 20 m resolution. This fact leads to fuzzy information for 10 m bands.

Maintenance: The dataset is not maintained, meaning there are no adjustments made since the underlying L1C data have been reprocessed.

Format: The samples are stored on pixel level in a HDF5 database. Unfortunately, the polygons used to produce the database have not been archived.

Processing baseline of used products: The collection is based on very early products. The products used for the collection have still been in the old multi-granule format. The products are not distributed with the database, only the pixel information (spectra, geometries, etc.) for the collected samples. Therefore, any change in the L1C data, especially geometric changes will make this dataset unusable for processors not capable of operating on mono-temporal single pixel basis.

Correlated data: The samples are collected as polygons. This means that each polygon consists of tenths to hundreds of pixels. Figure 19 shows an example[17] of a clear sample (samples have been extracted from 20 m resolution, here shown on 10 m RGB). The sample spacing is 20 meter which leads to a certain amount of correlation between the single samples. In the case of the example the correlation is obvious. This kind of correlations must be avoided when creating validation datasets. If all samples of the Hollstein dataset had been collected using the same shape and be evenly and stratified distributed over one scene, it would have been a better validation source. The only solution to this weakness would be to use the underlying polygons for a validation approach. This would mean to validate on vector level based on percentage of class agreement for each polygon.

Figure 19: Example of spatially correlated samples

Besides these issues, there are some quality constraints concerning the samples themselves.

[17] S2A_MSIL1C_20151204T170702_N0204_R069_T15TWJ_20151204T170659

Hollstein et al. 2016 had been aware of this fact, as they state: *"It is evident that there is a large degree of freedom on how polygons are placed, and which objects are marked. Also, the extent of objects with diffuse boundaries poses a particular burden on the consistency of the manual classification. This merely indicates that a certain degree of subjectivity is inherent in this approach."*

In the following section some examples of quality issues are shown, starting with the most important samples for this exercise, the cloud samples. Cloud samples are only taken at the center of opaque clouds. Semitransparent clouds are completely missing in the cloud class, as shown in Figure 20, Figure 21 and Figure 22.[18]

[18] S2A_MSIL1C_20151203T105422_N0204_R051_T30SXH_20151203T105811

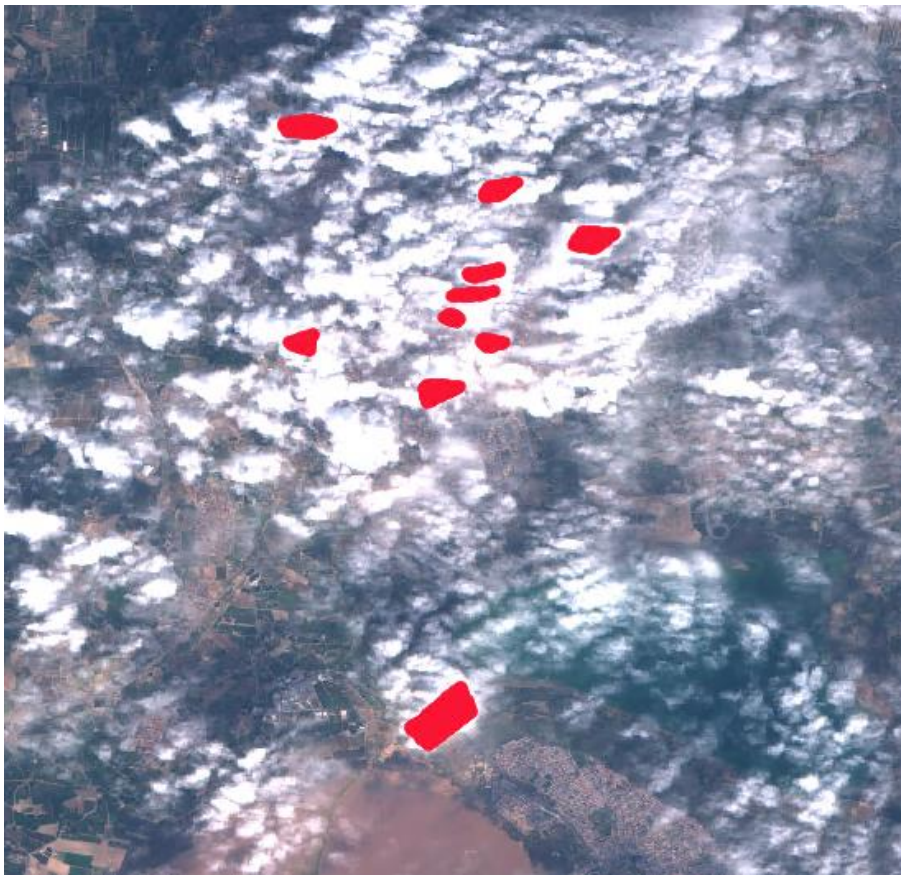


Figure 20: Example of only opaque cloud samples in the Hollstein dataset

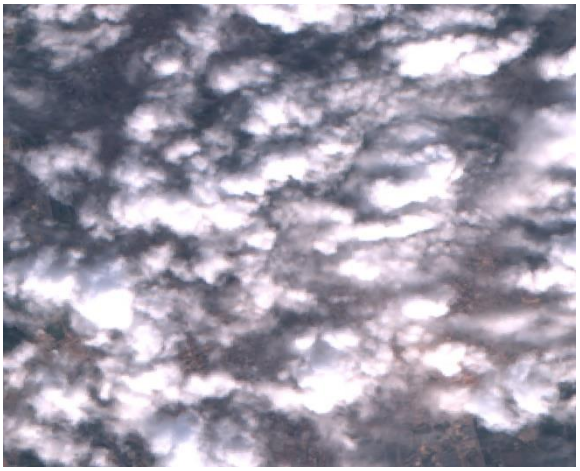


Figure 21: Detailed view (RGB only)

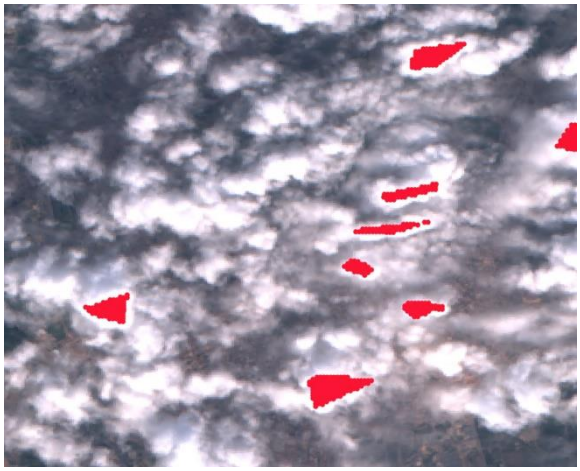


Figure 22: Detailed view with cloud samples

Nevertheless, semi-transparent clouds are collected within the cirrus class. Problematic with this collection is the mixture of semi-transparent clouds, not being cirrus clouds (e.g. thin, warm, low altitude, stratus or strato-cumulus clouds) and actual cirrus clouds. This mixture makes this class nearly unusable, especially in the context of CMIX this seems problematic.

Clear samples are sometimes taken under arguable circumstances, especially clear samples over. While checking the data, the example in Figure 23 was found. The example is from Lago Maggiore and shows fog on the water surface. It is simple to prove that this is fog and no sun glint, since the time of the day of the image is too early for sun glint and there are trails cutting through the fog, that can be clearly

traced to boats, when zooming in. Figure 24 shows that clear water samples had been extracted for the database. This is quite worrying, as Hollstein et al. (2016) had stated in their publication that each sample site has been revisited to ensure the best quality.

Figure 23: Clear water sample (only RGB)

Figure 24: Clear water sample with sample points

3.2.2 S2/L8 PixBox dataset

The strength of the PixBox validation dataset is its depth in detail. It captures for each pixel a great variety of information, i.e. opacity of cloud coverage, surface type, snow coverage, shadow type, etc. Furthermore, the dataset fulfills the requirement of stratified sampling, as the samples are equally distributed thematically and spatially per scene. Additionally, the single products of the collection are spatially well distributed.

The weakness of the dataset is the bias introduced by the expert collecting the information. On the one hand, a single expert is better trained to identify and classify pixel properties compared to a single layman, through his expert knowledge. On the other hand, multiple laymen could achieve better results by lowering the bias compared to a single expert. Using multiple experts to conduct a PixBox collection would be very costly. Nevertheless, the error introduced by this bias can be considered low compared to semi-automatically produced validation datasets. Another weakness of the dataset is the inability to detect systematic errors, a weakness that prevails all validation datasets.

3.2.3 S2/L8 GSFC dataset

The GSFC is the only dataset (among those used in CMIX) that uses ancillary information to help the expert to identify clouds in the satellite imagery – ground-based images of the sky and AOT measurements from the Aeronet station. This method was designed to reduce subjectivity in labelling clouds, but all the masking of the sky images was made manually, as well as the correspondence with the images. As seen on the figures, some thin cloud pixels in the sky images are not included, so the subjectivity remains. Almost all pixels in the imagery were labelled except of those on the boundaries between cloudy and clear areas. Another strength of this dataset is that it's acquired over the same territory, consistent errors in cloud detection can be analyzed.

The limitations of the dataset include the following: it's a single site and ideally multiple sites (like in the Aeronet case) are needed, works are in progress to expand the sites; the field of view is limited to ~30 km in diameter and cannot cover the whole extent of a Landsat 8 (~185 km) or Sentinel-2 (290 km)

scene. The dataset also has the tendency to classify very thin cirrus as cloudy, that may have only little effect on the surface reflectance.

3.2.4 L8 Biome dataset

The strengths of the L8Biome dataset include stratified random sampling of the scenes that represent various cloud and land cover conditions. All pixels of the Landsat scenes are classified. Limitations of this dataset is that labelling is fully driven by an expert and based on the subjective definition of what the cloud is. Nevertheless, the error introduced by this bias can be considered low compared to semi-automatically produced validation datasets.

3.2.5 S2 CESBIO dataset

The strengths of the S2 CESBIO dataset is that it provides fully labelled scenes derived from the machine learning algorithm interactively supervised by an expert, along with the confidence of that classification. The main weakness is that machine learning classification is not error-free and providing a confidence of classification is still a major challenge in the machine learning domain. Nevertheless, the scenes were fully checked and iteratively improved by the experts to achieve best possible results.

3.3 Validation methods

Since reliable cloud masking and flagging of unreliable pixels of Level 1 data is most important for many downstream algorithms and processors (for CMIX the purpose is AC over land & water), the quality assessment and validation of cloud masks and their algorithms have become more and more important.

Typically, there are five main approaches to assess the quality of cloud masking algorithms. It is important to note that all different methodologies are also bound to a specific definition of input data. The four main approaches are:

1. Visual inspection of images
   - typical cases
   - critical cases (know issues for S2 spectral bands: cloud vs snow, semi-transparent clouds, small patchy cloud fields, coastlines, bright beaches, salt lakes, urban areas, etc.)
2. Statistical assessment on global, representative scale
   - Expert pixel collections
   - Manual classifications
3. Self consistency in reflectance time series
   - Undetected clouds add noise of surface reflectance time series. Comparing the noise on this surface reflectance time series allows to compare the performances of different cloud masking methods
4. Level 3 composites
   - Temporal aggregation of surface reflectance or L2 parameters
   - criteria
     - „colouring“ of L3s, artefacts (i.e. whiter pixels showing cloud residuals)
     - number of valid observations.
5. Object-oriented: errors related to
   - oversegmentation,
   - undersegmentation,
   - edge-location,
   - fragmentation and shape

During the CMIX the first two approaches were used for validation with statistical assessment being the main validation tool. The third and fourth approaches were not listed at the beginning of the exercise as production of a complete time series over multiple sites would have increased the processing requirements for all participants a lot. Nevertheless, this approach might be used during CMIX II. The fifth methodology was listed in the first workshop as well as in the protocol. However, the effort of creating a needed reference dataset exceeded the resources for this exercise. In addition, the approach will only work for small opaque clouds, evenly distributed throughout a product with very sharp cloud edges. Therefore, the method was not used during CMIX.

3.3.1 Pixel based validation (confusion matrices)

Thematic accuracy describes the relationship of a mapped class of a remotely sensed pixel to a defined "truth" reference for the respective pixel. To ensure a meaningful result, all reference data (in this case pixels) must be correct or at least knowledge about the accuracy of the reference data is needed to allow a fair comparison with meaningful results (Congalton, 2007). Congalton (1991) points out that, "Although no reference data set may be completely accurate, it is important that the reference data have high accuracy or else it is not a fair assessment. Therefore, it is critical that the ground or reference data collection be carefully considered in any accuracy assessment."

Not all used datasets described in section 3.1 fulfill this requirement. The CESBIO dataset for example is based on a supervised classification. Even though, it is validated against reference data and shows a high accuracy, the validation results are only a proxy for the quality and only overall accuracies are provided with the dataset. Additionally classification errors for example in the Hollstein dataset have been found, which has been the validation source. Therefore, the high accuracy for all pixels is not given for these two datasets. Nevertheless, the other datasets have not provide accuracy numbers either. This needs to be considered when evaluating the validation results.

Pixel based validation was done based on the validation datasets described in section 3.1.

The basis for the accuracy assessment (validation) is a so-called confusion or error matrix, a squared array of numbers. In this array columns define the classes of the reference data, while rows show the algorithm results. Agreement between the two data can be found on the diagonal of the matrix. A confusion matrix is an efficient way of quantitatively assessing two classifications (reference vs. algorithm results). There are multiple measures of accuracy that can be calculated from the matrix. They are defined as follows:

- Overall accuracy (OA): Sum of all the entries along the diagonal divided by the total number of samples in the matrix (in percent).
- Balanced overall accuracy (BOA): Calculated as the average of the proportion corrects of each class individually (in percent).
- User accuracy (UA): Number of correctly classified samples of a class, divided by the sum of all classified samples of the same class (in percent).
- Producer accuracy (PA): Number of correctly classified samples of a class, divided by the sum of all reference samples of the same class (in percent).
- F score: Harmonic mean of UA and PA.

From accuracies also the error can be calculated. Errors are divided into commission and omission errors, samples from a specific reference class that have been classified wrong by the algorithm:

- Commission error: Percent of samples classified by the algorithm as a certain class that are classified differently in the reference dataset. The same as 100% minus user accuracy.
- Omission error: Percent of samples from a specific reference class that have been classified wrong by the algorithm. The same as 100% minus producer accuracy.

High number in User's Accuracy for the CLOUD flag means that the pixel under the CLOUD flag is most probably a cloud. If the Producer's Accuracy is high for the reference cloud it means that a reference cloud is most probably classified as CLOUD. A low Producers' Accuracy for clouds indicates that not all clouds are classified as CLOUD (error of omission), while a low User's Accuracy for the CLOUD flag accuracy indicates that the CLOUD flag has classified also clear surfaces (error of commission).

Figure 25 shows an example of a confusion matrix including all accuracies and errors.

		Truth (validation dataset)					
		A	**B**	**C**	**Total**	**User's accuracy [%]**	**Commission error [%]**
Classification (algorithm output)	**A**	**195**	10	15	220	89%	11%
	B	5	**120**	15	140	86%	14%
	C	0	70	**170**	240	71%	29%
	Total	200	200	200	600		
	Producer's accuracy [%]	98%	60%	85%			
	Omission error [%]	3%	40%	15%			
	Overall accuary	81%					
	Overall error	19%					

Figure 25: Example of a confusion matrix for a three classed (A,B,C) classification

Since each omission from the correct class is a commission into a wrong class, it is important considering user's and producer's accuracy. Reporting only one measure could be misleading (Congalton, 2007).

Congalton (1991) stated that the following factors must be considered to generate a valid confusion/error matrix:

1. Reference data collection.
2. Classification scheme.
3. Sampling scheme
4. Spatial autocorrelation
5. Sample size and sample unit

Not considering one of these factors could already lead to significant shortcomings in the accuracy assessment. The five points in relation to this exercise will be addressed briefly.

Reference data collection:

Reference data collection is the most important factor and first step in any assessment procedure. The significance of the assessment is depending on the correctness of the reference. There are multiple ways of creating a reference data collection. Each collection method adds a certain bias to the reference. This bias needs to be known and kept as low as possible.

Classification scheme:

The classification scheme of the classification (algorithm output) and the reference data must be identical. Not in all cases the reference has the same classification scheme and a reclassification is needed before validation. By reclassifying reference data or algorithm outputs a certain bias is introduced into the validation results, especially if the definition of the classes is not 100% identical. In case of validating a reference against multiple algorithm outputs, all algorithm outputs must have identical classes which are identically defined, to ensure a precise comparability.

Sampling scheme:

This factor is not of a big importance for CMIX, as only two classes (cloud vs. non-cloud) are examined. Nevertheless, it should be noted that when validating multiple classes, knowledge of the distribution of the classes in the algorithm output should be considered. A stratified random sampling approach is normally considered to be the most useful approach. For CMIX this approach is less feasible, as these sampling approaches start from the classification (algorithm output). In CMIX there is not only one classification output but multiple outputs that are compared. Therefore, the sampling needs to be done on the original input data used for creating the reference dataset. For some datasets this methodology cannot even be applied, like the GSFC dataset, which is based on sky-camera observation. Nevertheless, the distribution of the classes and samples in relation to the validated product need to be considered when analyzing the assessment result.

Another complication lies in the transition of a classification, if the classes are not simply “binary” like transparent clouds/cloud borders. Several data sets tend not to sample the limits of clouds, where the difficulty often lies. It is the case for GSFC and Hollstein.

Spatial Autocorrelation

Congalton (2007) states: “Because of sensor resolution, landscape variability, and other factors, remotely sensed data are often spatially autocorrelated. Spatial autocorrelation involves a dependency between neighboring pixels such that a certain quality or characteristic at one location has an effect on that same quality or characteristic at neighboring locations (Cliff and Ord 1973, Congalton 1988a). Spatial autocorrelation can affect the result of an accuracy assessment if an error in a certain location can be found to positively or negatively influence errors in surrounding locations. The best way to minimize spatial autocorrelation is to impose some minimum distance between sample units.”

This factor mostly needs to be considered when analyzing the Hollstein, CESBIO, GSFC and L8Biome dataset, as these datasets consist of classified pixel areas. In case of L8Biome and CESBIO dataset even of a completely classified product.

Sample Size and Sample Unit:

For the size of the samples a general rule of thumb is to use a minimum of 50 to 100 samples per class. Using this rule each class can be assessed individually (Congalton & Green, 1999). Each of the validation datasets fulfills this requirement clearly.

The sample unit factor can be neglected for CMIX as the reference data is based on the algorithm input products. This means the input for creating the validation dataset are also Sentinel-2 or Landsat 8 data and thus the sample unit is equal to Sentinel-2 or Landsat 8 resolution.

Note:

Sample unit is the size of the sample. If you are dealing with in-situ data collected in the field, this can be relevant. For example, if you are validating a LC map with a pixel resolution of 30m. Assuming you did not use any minimum mapping unit above this, your in-situ samples should at least cover 900m², to represent your classification properly.

Confusion matrices used in the CMIX are mostly based on binary classifications, in this case cloud and non-cloud. Figure 26 shows an example of a confusion matrix used in CMIX.

	Reference				
	CLOUD	**NO CLOUD**	**Total**	**User's accuracy [%]**	**Commission error [%]**
CLOUD	R_cloud_as_M_cloud	R_ncloud_as_M_cloud	Sum(M_cloud)	R_cloud_as_M_cloud / Sum(M_cloud)	1-(R_cloud_as_M_cloud / Sum(M_cloud))
NO CLOUD	R_cloud_as_M_ncloud	R_ncloud_as_M_ncloud	Sum(M_ncloud)	R_ncloud_as_M_ncloud / Sum(M_ncloud)	1-(R_ncloud_as_M_ncloud / Sum(M_ncloud))
Total	Sum(R_cloud)	Sum(R_ncloud)	Sum(R_cloud) + Sum(R_ncloud) or Sum(M_cloud) + Sum(M_ncloud)		
Producer's accuracy [%]	R_cloud_as_M_cloud / Sum(R_cloud)	R_ncloud_as_M_ncloud / Sum(R_ncloud)			
Omission error [%]	1-(R_cloud_as_M_cloud / Sum(R_cloud))	1-(R_ncloud_as_M_ncloud / Sum(R_ncloud))			
Overall accuary	(R_cloud_as_M_cloud + R_ncloud_as_M_ncloud) / Sum(R_cloud) + Sum(R_ncloud)				
Overall error	1 - (R_cloud_as_M_cloud + R_ncloud_as_M_ncloud) / Sum(R_cloud) + Sum(R_ncloud)				

Figure 26: Example of a confusion matrix used for CMIX incl. definitions

3.3.2 Visual analysis

Statistical assessment of a classification, as described in section 3.3.1, is commonly based on samples. Therefore, it can only investigate a subset of a classified satellite image. Even a very carefully created reference will in most cases not be able to identify the critical cases for each cloud masking algorithm. This is especially true for systematic errors which can only be identified by monitoring time series of single pixels of potentially critical cases. In case of Sentinel-2 these normally are bright urban targets or other bright bare surfaces like salt lakes or beaches. While these systematic errors have not been studied during CMIX, as no time series had been produced and no reference dataset allowed this type of analyses, critical cases for any cloud masking algorithm still exist, like cloud borders, spatially continuous thinning of clouds, as well as unsystematic detection of other bright surfaces. Besides analyzing the behavior of all algorithms regarding these critical cases, visual inspection also allows a simple way of comparing the performance of a cloud mask in the spatial domain.

3.4 Intercomparison results

3.4.1 Pixel based validation (confusion matrices)

3.4.1.1 S2 CESBIO dataset

Table 6 shows performance metrics, when applying cloud masking algorithms on the Sentinel-2 CESBIO datasets. Several remarks shall be made, when analyzing these results. The number of reference pixels varied, since 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), and 240 m (MAJA). Cloud and non-cloud classes were imbalanced in the reference dataset, therefore it would lead to the OA being more biased towards non-cloud classes. Therefore, the balanced OA (BOA) would be a more appropriate metric. BOA varied from 79.5% to 90.5%, with average being 85.9±3.9%. Except MAJA (whose developers generated the CESBIO dataset) with 92.9%, the highest Cloud-PA was 85.6%, with average being 75.9±8.7%, meaning that some of the algorithms (ATCOR, Sen2Cor and CD-FCNN) missed almost 25% of clouds identified in the CESBIO dataset.

Table 6: Performance metrics of algorithm using the CESBIO data

					Cloud			Non-cloud		
Processor	**% cloud**	**Total num valid pixels**	**OA**	**BOA**	**PA**	**UA**	**F**	**PA**	**UA**	**F**
ATCOR	24.3	772,086,084	88.6	80.4	64.4	84.9	73.3	96.3	89.4	92.8
Idepix	24.3	772,086,084	91.7	86.9	77.5	86.9	81.9	96.2	93.0	94.6
MAJA	25.6	697,056,918	89.2	90.5	92.9	72.7	81.5	88.0	97.3	92.4
Fmask 4.0 CCA	24.3	772,086,084	93.3	88.9	80.4	90.8	85.3	97.4	93.9	95.6
FORCE	24.3	3,088,386,349	91.1	88.9	84.7	79.9	82.2	93.2	95.0	94.1
InterSSIM	24.3	3,088,386,349	93.2	88.0	77.8	93.1	84.8	98.2	93.3	95.6
LaSRC	24.3	3,088,386,349	81.2	82.7	85.6	57.6	68.9	79.8	94.6	86.6
S2cloudless	24.3	3,088,386,349	93.1	88.8	80.4	90.2	85.0	97.2	93.9	95.5
sen2cor	24.3	772,086,084	91.0	84.7	72.3	88.7	79.6	97.0	91.6	94.3
CD-FCNN	24.3	85,782,723	89.5	79.5	60.3	94.1	73.5	98.8	88.6	93.4

3.4.1.2 S2 GSFC

Table 7 shows the results of comparing algorithm outcomes against the S2 GSFC dataset. As with CESBIO dataset, the different amount of reference pixels is the result of algorithms producing maps at various spatial resolutions. Also, 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 the difference between OA and BOA are minimal. BOA varied from 80.7% to 96.8% with LaSRC being the outlier (developers of LaSRC produced GSFC data), with average being 85.9±2.9% (not considering LaSRC). Average cloud-PA and cloud-UA was 73.7±5.6% and 98.2±2.7%, respectively, meaning large omission errors.

The reason for all algorithms producing lower accuracies compared to LaSRC is that they were unable to correctly classify thin (transparent/cirrus) clouds, which, in turn, LaSRC is masking using a conservative threshold (0.003 in reflectance units; for LaSRCv3.5.5) applied for the cirrus band (B10). Those clouds were labelled as thin, since they were clearly visible in the ground-based images. If thin clouds are not considered in the analysis (Table 8), all algorithms show much better performance: average BOA is 94.4±2.9% (an average gain +7.4±2.6%) and cloud-PA is 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 determining and identifying thin (transparent/cirrus) clouds, at the same time mostly agreeing on thick clouds.

Table 7: Performance metrics of algorithms using the GSFC data

					Cloud			Non-cloud		
Processor	**% cloud**	**Total num valid pixels**	**OA**	**BOA**	**PA**	**UA**	**F**	**PA**	**UA**	**F**
ATCOR	60.6	11,566,166	77.9	81.7	63.5	100.0	77.7	100.0	64.0	78.1
Idepix	60.6	11,566,166	84.8	86.1	80.1	93.9	86.5	92.0	75.1	82.7
MAJA	49.2	15,609,378	80.9	80.7	66.2	93.0	77.4	95.2	74.4	83.5
Fmask 4.0 CCA	60.6	11,566,166	86.0	88.4	77.1	99.7	86.9	99.6	73.9	84.8
FORCE	60.6	46,266,297	86.1	88.2	78.2	98.6	87.2	98.3	74.6	84.8
InterSSIM	60.6	46,266,297	85.0	87.6	75.4	99.7	85.9	99.7	72.5	84.0
LaSRC	60.6	46,257,284	96.7	96.8	96.3	98.2	97.3	97.3	94.5	95.9
S2cloudless	60.6	46,266,297	85.2	87.7	76.1	99.3	86.2	99.2	73.0	84.1
sen2cor	60.6	11,566,166	85.2	87.8	75.8	99.7	86.1	99.7	72.8	84.2
CD-FCNN	60.6	46,266,297	82.4	85.4	71.0	99.9	83.0	99.8	69.1	81.7

Table 8: Performance metrics of algorithms using the GSFC data and removing thin (transparent) clouds from the reference

					Cloud			Non-cloud		
Processor	**% cloud**	**Total num valid pixels**	**OA**	**BOA**	**PA**	**UA**	**F**	**PA**	**UA**	**F**
ATCOR	55.5	10,236,222	86.9	88.2	76.4	100.0	86.6	100.0	77.3	87.2
Idepix	55.5	10,236,222	92.5	92.5	92.9	93.6	93.2	92.0	91.3	91.6
MAJA	40.8	13,380,304	92.7	92.2	89.1	92.7	90.9	95.2	92.7	93.9
Fmask 4.0 CCA	55.5	10,236,222	96.1	96.5	93.3	99.7	96.4	99.6	92.3	95.8
FORCE	55.5	40,946,551	95.9	96.1	94.0	98.5	96.2	98.3	93.0	95.5
InterSSIM	55.5	40,946,551	95.6	96.0	92.4	99.7	95.9	99.7	91.3	95.3
LaSRC	55.5	40,937,538	98.0	97.9	98.5	97.8	98.2	97.3	98.1	97.7
S2cloudless	55.5	40,946,551	95.7	96.1	93.0	99.3	96.0	99.2	91.9	95.4
sen2cor	55.5	10,236,222	95.0	95.4	91.2	99.7	95.3	99.7	90.1	94.6
CD-FCNN	55.5	40,946,551	92.9	93.6	87.3	99.9	93.1	99.8	86.3	92.6

3.4.1.3 S2 Pixbox dataset

Not all algorithms could process all 29 Products of the PixBox S2 dataset. The reasons for this were limitations of allowed geometries (ATCOR) or too sparse time series around the acquisition (MAJA).

Table 9: Overview of submitted products and formats for the PixBox dataset

Algorithm	Number of products	File format	Number of classes
ATCOR	27	BSQ	5
CD-FCNN	29	HDF5	3
Fmask 4.0 CCA	29	TIFF	8
FORCE	29	TIFF	15
Idepix	29	NetCDF	21
InterSSIM	29	TIFF	2
LaSRC	29	TIFF	7
MAJA	14	TIFF	8
S2cloudless	29	TIFF	2
sen2cor	29	JP2	12

To account for the difference of available products for validation, two different comparisons were made, one using all available products for each algorithm and a second one using only the products that all algorithms have been produced. We call the second dataset the least common denominator (LCD) subset, while the first is referred to as the “complete dataset”. The whole comparison could have been made only on the LCD subset, but this reduces the complete dataset by half, which reduces its significance. Therefore, the complete dataset also was used for comparison. In this comparison using the complete dataset, results for MAJA must be taken with caution, as they are only based on 14 out of 29 products.

As the PixBox dataset includes a great variety of collected pixel features (see section 3.1.2) a very detailed analysis could be made. This includes analyzing the performance of the algorithms including and excluding very thin clouds, a detailed comparison of cloud/clear of the algorithms compared to

different classes of the PixBox dataset, as well as the analysis of the performance separated over land and water.

Table 10 lists the combinations that have been analyzed.

Table 10: Analysis scenarios for the PixBox S2 dataset

Scenarios	Dataset	Incl. thin clouds	Incl. snow	Surface
1	Complete	No	Yes	All
2	Complete	Yes	Yes	All
3	Complete	Yes	No	All
4	LCD	No	Yes	All
5	LCD	Yes	Yes	All
6	LCD	Yes	No	All
7	Complete	Yes	No	Land
8	Complete	Yes	No	Water

In addition to these scenarios also detailed "confusion matrix like" figures have been created. These show the distribution of cloud and clear detected pixels of each algorithm compared to the collected pixel's features. These plots help to identify strength and weaknesses of each algorithm. The single confusion matrices per algorithm are listed in the annex.

Table 11 shows the results of all algorithms for the first scenario. The complete dataset of each algorithm was used over all surfaces excluding thin clouds.

Table 11: S2 PixBox results - complete dataset without thin clouds, over all surfaces

Algorithm	PA Non-cloud	PA Cloud	UA Non-cloud	UA Cloud	OA	Balanced OA	Number of pixels
ATCOR	89.9	70.8	83.1	81.4	82.54	80.35	13094
CD-FCNN	93.4	82.7	90.3	87.9	89.46	88.05	13981
Fmask 4.0 CCA	89	90.8	94.3	82.7	89.64	89.9	13981
FORCE	81.2	90.4	93.6	73.6	84.56	85.8	13986
Idepix	66.7	95.3	96.1	62.4	77.21	81	13986
InterSSIM	95.2	86.2	92.2	91.3	91.92	90.7	13986
LaSRC	48.2	93.8	93.1	51.3	64.95	71	13986
MAJA	82.3	94.3	96.4	74.3	86.52	88.3	6760
S2cloudless	91.6	91.6	94.9	86.4	91.6	91.6	13986
sen2cor	86.9	82.7	89.6	78.6	85.36	84.8	13986

Some algorithms show a fair amount of commissioning error of clear observations as clouds, like FORCE, IdePix, sen2cor and especially LaSRC. In the detailed figures given in this section, it will become clear that especially LaSRC and IdePix detected most clear snow pixels as cloud. The IdePix team had indicated that this was caused by a bug in the algorithm bypassing the snow test. This bug was fixed but the data used in CMIX include this error. For LaSRC, which has the main aim at atmospheric

correction, snow is defined as non-valid pixels, because aerosol retrievals over snow are problematic and highly unreliable.[19]

Figure 27 to Figure 36 give a detailed insight to more information stored in the PixBox dataset. The figures show for multiple classes of the PixBox dataset, how the algorithms handle these pixels to be either cloud or clear. This detailed information is not part of the scenario 1, but gives a detailed view to better understand the results of the following scenario (including thin clouds). Most classes, like the cloud classes are self-explaining, while some need a bit of detail. “Clear” for example are all clear pixels except snow- or ice-covered clear pixels. These snow- and ice-covered clear pixels are listed under “Clear Snow”. “Cloud Border” comprises pixels that are located directly at a cloud border and are characterized by being a mixed cloud/clear pixel.

Figure 27 to Figure 36 show the above-described shortcoming of some algorithms to properly identify snow/ice pixels as clear. It also shows that algorithms that are using a cloud buffer/dilation (FORCE, FMask 4.0 CCA, InterSSIM, MAJA, S2cloudless) are superior in identifying mixed cloud/clear pixels at cloud borders.

The figures also show that only three algorithms (IdePix, LaSRC, and MAJA) have a higher tendency to correctly detect thin semi-transparent clouds.

In addition to this, Figure 57 and Figure 58 in the annex, give a detailed overview of correct classification and misclassification of clear pixel over different surfaces for each algorithm.

[19] Note of advice to the algorithm developers: Create a separate flag for these invalid snow observations instead of including them in a cloud flag.

ATCOR

Class	Clear	Clear Snow	Thin Semi-transp. Cloud	Average Semi-transp. Cloud	Thick Semi-transp. Cloud	Opaque Cloud	Cloud Border	Fog
Clear	5462	1616	1367	855	306	310	202	117
Cloud	571	152	1175	1039	646	1880	127	0
Sum	6033	1768	2542	1894	952	2190	329	117

Figure 27: Validation results of cloud/clear vs. different in-situ cloud types, clear and snow - for ATCOR

CD-FCNN

Class	Clear	Clear Snow	Thin Semi-transp. Cloud	Average Semi-transp. Cloud	Thick Semi-transp. Cloud	Opaque Cloud	Cloud Border	Fog
Clear	5785	2216	1791	747	129	13	172	0
Cloud	453	132	959	1211	847	2191	157	117
Sum	6238	2348	2750	1958	976	2204	329	117

Figure 28: Validation results of cloud/clear vs. different in-situ cloud types, clear and snow - for CD-FCNN

Fmask 4.0 CCA

Class	Clear	Clear Snow	Thin Semi-transp. Cloud	Average Semi-transp. Cloud	Thick Semi-transp. Cloud	Opaque Cloud	Cloud Border	Fog
Clear	5765	1934	1153	362	99	14	8	0
Cloud	473	414	1597	1596	877	2190	321	117
Sum	6238	2348	2750	1958	976	2204	329	117

Figure 29: Validation results of cloud/clear vs. different in-situ cloud types, clear and snow - for Fmask 4.0 CCA

FORCE

Class	Clear	Clear Snow	Thin Semi-transp. Cloud	Average Semi-transp. Cloud	Thick Semi-transp. Cloud	Opaque Cloud	Cloud Border	Fog
Clear	5371	1693	1164	386	84	23	5	0
Cloud	868	656	1587	1573	892	2183	324	117
Sum	6239	2349	2751	1959	976	2206	329	117

Figure 30: Validation results of cloud/clear vs. different in-situ cloud types, clear and snow - for FORCE

IdePix

Class	Clear	Clear Snow	Thin Semi-transp. Cloud	Average Semi-transp. Cloud	Thick Semi-transp. Cloud	Opaque Cloud	Cloud Border	Fog
Clear	5289	473	875	201	36	3	173	0
Cloud	950	1876	1876	1758	940	2203	156	117
Sum	6239	2349	2751	1959	976	2206	329	117

Figure 31: Validation results of cloud/clear vs. different in-situ cloud types, clear and snow - for IdePix

InterSSIM

Class	Clear	Clear Snow	Thin Semi-transp. Cloud	Average Semi-transp. Cloud	Thick Semi-transp. Cloud	Opaque Cloud	Cloud Border	Fog
Clear	5937	2232	1444	582	114	13	41	0
Cloud	302	117	1307	1377	862	2193	288	117
Sum	6239	2349	2751	1959	976	2206	329	117

Figure 32: Validation results of cloud/clear vs. different in-situ cloud types, clear and snow - for InterSSIM

LaSRC

Class	Clear	Clear Snow	Thin Semi-transp. Cloud	Average Semi-transp. Cloud	Thick Semi-transp. Cloud	Opaque Cloud	Cloud Border	Fog
Clear	4089	153	728	263	53	1	169	0
Cloud	2150	2196	2023	1696	923	2205	160	117
Sum	6239	2349	2751	1959	976	2206	329	117

Figure 33: Validation results of cloud/clear vs. different in-situ cloud types, clear and snow - for LaSRC

MAJA

Class	Clear	Clear Snow	Thin Semi-transp. Cloud	Average Semi-transp. Cloud	Thick Semi-transp. Cloud	Opaque Cloud	Cloud Border	Fog
Clear	2643	960	270	89	44	2	4	0
Cloud	554	222	892	794	507	945	157	117
Sum	3197	1182	1162	883	551	947	161	117

Figure 34: Validation results of cloud/clear vs. different in-situ cloud types, clear and snow - for MAJA

S2cloudless

Class	Clear	Clear Snow	Thin Semi-transp. Cloud	Average Semi-transp. Cloud	Thick Semi-transp. Cloud	Opaque Cloud	Cloud Border	Fog
Clear	5726	2125	1127	349	74	9	21	0
Cloud	513	224	1624	1610	902	2197	308	117
Sum	6239	2349	2751	1959	976	2206	329	117

Figure 35: Validation results of cloud/clear vs. different in-situ cloud types, clear and snow - for S2cloudless

Sen2cor

Class	Clear	Clear Snow	Thin Semi-transp. Cloud	Average Semi-transp. Cloud	Thick Semi-transp. Cloud	Opaque Cloud	Cloud Border	Fog
Clear	5523	2072	1107	600	191	98	285	0
Cloud	716	277	1644	1359	785	2108	44	117
Sum	6239	2349	2751	1959	976	2206	329	117

Figure 36: Validation results of cloud/clear vs. different in-situ cloud types, clear and snow - for Sen2cor

Table 12 shows now the results of the second scenario, including thin semi-transparent clouds. It becomes obvious that thin semitransparent clouds are not well detected by any algorithm. Even these performing around 90% balanced overall accuracy (BOA) on the dataset without thin clouds, like S2cloudless, InterSSIM and Fmask 4.0 CCA, are now performing below 86% BOA. Only LaSRC and MAJA show decrease in BOA below 4%, even though the performance of LaSRC is very low in general due to the snow pixels, as described above.

When taking a closer look at the UA for non-cloud pixels, you will recognize a good decrease after additionally considering the thin semi-transparent clouds for most of the algorithms. In other words, there is a good amount of commission error of non-cloud flagged pixels, which are cloud.

Table 12: S2 PixBox results - complete dataset including thin clouds, over all surfaces

Algorithm	PA Non-cloud	PA Cloud	UA Non-cloud	UA Cloud	OA	Balanced OA	Number of pixels
ATCOR	89.9	62.5	71.8	85.3	76.64	76.2	15636
CD-FCNN	93.4	66	75.5	89.9	80.49	79.7	16731
Fmask 4.0 CCA	89	79.4	82.9	86.5	84.45	84.2	16731
FORCE	81.2	79	81.2	78.9	80.15	80.1	16737
Idepix	66.7	85.9	84.1	69.7	75.73	76.3	16737
InterSSIM	95.2	72.7	79.6	93.2	84.62	83.95	16737
LaSRC	48.2	86.8	80.3	59.9	66.36	67.5	16737
MAJA	82.3	88.6	89.9	80.2	85.09	85.45	7922
S2cloudless	91.6	80.2	83.9	89.5	86.25	85.9	16737
sen2cor	86.9	74.7	79.4	83.6	81.16	80.8	16737

Nevertheless, algorithms having a high commission error of non-cloud pixels often have a low commission error for cloud pixels (The usage of a buffer/dilation might have quite an influence here). This shows that there are three major classes of algorithms, cloud conservative and non-cloud conservative and balanced. Non-cloud conservative approaches are mostly needed for applications that do not allow cloud contamination (e.g. the remote sensing of land, sea-, and ice surface temperatures, of vegetation biophysical variables, of total column water vapor or of aerosol optical properties), while cloud conservative approaches are mostly needed for cloud remote sensing applications. Depending on the application and sensor, a commission error can be rated as less problematic, for example due to the high repetition cycle of Sentinel-2. For these cases it might be a big issue, if the error is systematic.

It might be important to note that these categorizations are based on the cloud definitions each participant had provided, as the requirement for CMIX was to provide a binary cloud mask. The provided binary masks represent what each participants thought was best to represent the provided algorithm. Nevertheless, most of the algorithms provide more detailed cloud masks or additional probabilities to give the used more control on the cloud masking behavior of the algorithm.

As two of the algorithms had strong limitations or even failure with snow pixels, we also excluded the snow from the dataset. The third scenario is therefore used to see the influence snow has on the cloud detection, especially for these two algorithms having an issue with this (IdePix and LaSRC). When comparing the BOA, it becomes clear that removing snow does not have a big difference on the results of the other algorithms having a proper snow detection implemented. All results (except IdePix and LaSRC) get only a bit better without snow, with a maximum of 2% for FMask 4.0 CCA. But excluding snow from the analysis, shows that now IdePix and LaSRC show comparable results to the other algorithms. It is also worth mentioning that the most stable algorithm over the three shown scenarios is MAJA, with a change of BOA between the scenarios below 3%, but MAJA produced only half of the products.

Table 13: S2 PixBox results - complete dataset including thin clouds, over all surfaces except snow

Algorithm	PA Non-cloud	PA Cloud	UA Non-cloud	UA Cloud	OA	Balanced OA	Number of pixels
ATCOR	89.5	64.8	69.8	87.1	76.55	77.15	13187
CD-FCNN	93	67.8	72.9	91.3	79.96	80.4	13485
Fmask 4.0 CCA	91.4	81.1	81.8	91	86.04	86.25	13485
FORCE	84.5	79.7	79.5	84.7	82.01	82.1	13490
Idepix	83.5	84.5	83.3	84.7	84.03	84	13490
InterSSIM	95.3	74.4	77.6	94.5	84.46	84.85	13490
LaSRC	63.2	85.1	79.8	71.4	74.57	74.15	13490
MAJA	82.7	89.5	89.2	83.2	86.03	86.1	6255
S2cloudless	92	81.4	82.1	91.6	86.49	86.7	13490
sen2cor	86.4	77.7	78.3	86.1	81.91	82.05	13490

Scenarios 4 to 6 have been created to have a direct comparison between the algorithms based on the minimum number of products produced by all of them. This was mostly driven by MAJA, as this algorithm had produced only half of the complete dataset, due to processing requirements.

Table 14 shows overall and balanced overall accuracies for scenarios 4 to 6, which are basically the same as the previous three (scenarios 1 to 3) but using a reduced dataset.

The results are a bit different compared to the complete dataset. But they show the same tendencies for all algorithms. Depending on the tendency of being cloud conservative, non-cloud conservative or balanced the results for the single algorithms increase or decrease a few precent.

Table 14: S2 PixBox results – comparison of algorithms using the LCD dataset (scenarios 4 to 6)

	Excl. thin clouds		Incl. thin clouds		Incl. thin clouds, excl. snow	
Algorithm	**OA**	**BOA**	**OA**	**BOA**	**OA**	**BOA**
ATCOR	85.46	81.6	80	78.3	79.58	79.25
CD-FCNN	88.22	85.95	80.18	78.6	79.8	79.5
Fmask 4.0 CCA	90.85	89.65	86.03	85.1	87.09	86.9
FORCE	87.09	88.15	83.15	83	85.21	85.15
Idepix	74.14	78.8	72.67	73.8	83.04	83
InterSSIM	93.09	91.1	85.66	84.2	85.47	85.2
LaSRC	66.67	73.35	68.53	70.65	77.73	78
MAJA	86.52	88.3	85.09	85.45	86.03	86.1
S2cloudless	93.82	93.05	88.15	87.25	87.96	87.75
sen2cor	87.47	85.25	83.36	82.25	85.58	85.4

For convenience Table 15 shows a comparison of balanced overall accuracies of scenarios 1 & 4, 2 & 5, as well as 3 & 6. The differences between the complete dataset and the LCD dataset are mostly below 2%, with a few exceptions.

Table 15: S2 PixBox results – comparison of BOA of scenarios 1 & 4, 2 & 5, as well as 3 & 6.

	Excl. thin clouds		Incl. thin clouds		Incl. thin clouds, excl. snow	
Algorithm	**BOA comp.**	**BOA LCD**	**BOA comp.**	**BOA LCD**	**BOA comp.**	**BOA LCD**
ATCOR	80.35	81.6	76.2	78.3	77.15	79.25
CD-FCNN	88.05	85.95	79.7	78.6	80.4	79.5
Fmask 4.0 CCA	89.9	89.65	84.2	85.1	86.25	86.9
FORCE	85.8	88.15	80.1	83	82.1	85.15
Idepix	81	78.8	76.3	73.8	84	83
InterSSIM	90.7	91.1	83.95	84.2	84.85	85.2
LaSRC	71	73.35	67.5	70.65	74.15	78
MAJA	88.3	88.3	85.45	85.45	86.1	86.1
S2cloudless	91.6	93.05	85.9	87.25	86.7	87.75
sen2cor	84.8	85.25	80.8	82.25	82.05	85.4

As we now have gotten some insight into the behavior of the algorithms in relation to different cloud opacities, snow, and a reduced dataset, we like to take a closer look at the performance over the two major earth surfaces, land and water, separately.

Table 16 shows the results for all algorithms only over land surfaces, excluding snow, while Table 17 shows the results over water surfaces, also excluding snow/ice. The results show a small superiority of the algorithms to detect clouds over land, compared to water. Nevertheless, some algorithms perform very good over land, while performance drops 8% to 15% over water, especially CD-FCNN, InterSSIM,

ATCOR, and S2cloudless (all over 10% decrease). For others, the performance decreases less. For the four algorithms showing the biggest decrease in performance over water it might be important to note that three of these rely on machine learning (CD-FCNN, InterSSIM, and S2cloudless). Therefore, the under-performance might be cause by a training dataset having a bias for land surfaces.

Table 16: S2 PixBox results - complete dataset including thin clouds, over land surfaces except snow.

Algorithm	PA Non-cloud	PA Cloud	UA Non-cloud	UA Cloud	OA	Balanced OA	Number of pixels
ATCOR	88.9	73.9	74.7	88.5	80.86	81.4	7190
CD-FCNN	89.2	82.7	82.1	89.6	85.77	85.95	7275
Fmask 4.0 CCA	91.6	88.7	87.8	92.2	90.07	90.15	7275
FORCE	83.8	87.6	85.7	85.9	85.79	85.7	7277
Idepix	82.6	90.1	88.1	85.4	86.57	86.35	7277
InterSSIM	95.5	85.5	85.4	95.6	90.2	90.5	7277
LaSRC	57.8	89.8	83.5	70.6	74.77	73.8	7277
MAJA	80.4	98	98.1	78.9	87.92	89.2	3460
S2cloudless	90.6	91	90	91.6	90.84	90.8	7277
sen2cor	86.8	81.4	80.6	87.4	83.93	84.1	7277

Table 17: S2 PixBox results - complete dataset including thin clouds, over water surfaces except snow/ice.

Algorithm	PA Non-cloud	PA Cloud	UA Non-cloud	UA Cloud	OA	Balanced OA	Number of pixels
ATCOR	92.1	47.7	61.7	86.8	68.94	69.9	4505
CD-FCNN	99	41.7	60.6	97.8	68.93	70.35	4580
Fmask 4.0 CCA	92.3	68.2	72.4	90.7	79.63	80.25	4580
FORCE	86.7	67.5	70.7	84.9	76.63	77.1	4582
Idepix	86.1	75.5	76.1	85.7	80.55	80.8	4582
InterSSIM	96.3	56.2	66.6	94.4	75.27	76.25	4582
LaSRC	75.8	75.8	74	77.6	75.84	75.8	4582
MAJA	86.9	79.8	75.1	89.6	82.71	83.35	2418
S2cloudless	96	65	71.3	94.7	79.73	80.5	4582
sen2cor	86.9	71.8	73.6	85.8	78.94	79.35	4582

3.4.1.4 S2 Hollstein et al. 2016 dataset

As shown in section 3.2.1, the S2 Hollstein dataset has multiple weaknesses, one of these weaknesses being the cloud class only consisting of opaque clouds. Another issue was the mixture of actual cirrus clouds with other semi-transparent cloud types. Nevertheless, it was decided to use the dataset to validate two scenarios. In the first scenario, only opaque clouds are used. The results of this scenario have been presented at the CMIX workshop. In the second scenario, the cirrus class of the S2 Hollstein dataset set is used additionally.

Table 18: S2 Hollstein dataset results – only opaque clouds (classes == 50 used for cloud).

Algorithm	PA Non-cloud	PA Cloud	UA Non-cloud	UA Cloud	OA	Balanced OA	Number of pixels
ATCOR	95.1	81.8	86.4	93.2	89.07	88.45	1,063,605
CD-FCNN	97.3	98.3	98.6	96.7	97.76	97.8	1,084,232
Fmask 4.0 CCA	90.9	99.9	100	89.8	94.94	95.4	1,093,687
FORCE	90.5	97.4	97.7	89.1	93.56	93.95	1,094,539
Idepix	86.9	98.2	98.4	85.7	91.92	92.55	1,094,539
InterSSIM	98	96.8	97.5	97.5	97.48	97.4	1,094,539
LaSRC	75.6	96.7	96.6	76	84.99	86.15	1,094,539
S2cloudless	95.3	97.6	98	94.3	96.31	96.45	1,094,539
sen2cor	91.5	93	94.2	89.8	92.19	92.25	1,094,539

Table 19: S2 Hollstein dataset results – opaque clouds (classes == 50) and semi-transparent clouds/cirrus (classes == 40).

Algorithm	PA Non-cloud	PA Cloud	UA Non-cloud	UA Cloud	OA	Balanced OA	Number of pixels
ATCOR	95.1	84.6	79.4	96.5	88.61	89.85	1,517,387
CD-FCNN	97.3	71.1	67.1	97.7	80.98	84.2	1,583,604
Fmask 4.0 CCA	90.9	91.3	86.6	94.2	91.16	91.1	1,593,059
FORCE	90.5	88.2	82.6	93.8	89.11	89.35	1,593,911
Idepix	86.9	94.1	90	92.1	91.33	90.5	1,593,911
InterSSIM	98	85.7	80.9	98.6	90.4	91.85	1,593,911
LaSRC	75.6	97.7	95.3	86.7	89.28	86.65	1,593,911
S2cloudless	95.3	89.2	84.5	96.8	91.54	92.25	1,593,911
sen2cor	91.5	85.6	79.7	94.3	87.89	88.55	1,593,911

ATCOR

Class	Cloud	(Shadow)	(Cirrus)	Snow	Water	Clear	Sum
Clear	87543	134390	56605	87253	102536	230582	698909
Cloud	392617	9	397177	17102	0	11573	818478
Sum	480160	134399	453782	104355	102536	242155	1517387

Figure 37: Validation results of cloud/clear versus all S2 Hollstein dataset classes for ATCOR

CD-FCNN

Class	Cloud	(Shadow)	(Cirrus)	Snow	Water	Clear	Sum
Clear	8059	134745	276943	99821	102937	244044	866549
Cloud	478371	726	222429	4534	334	10661	717055
Sum	486430	135471	499372	104355	103271	254705	1583604

Figure 38: Validation results of cloud/clear versus all S2 Hollstein dataset classes for CD-FCNN

Fmask 4.0 CCA

Class	Cloud	(Shadow)	(Cirrus)	Snow	Water	Clear	Sum
Clear	269	136924	85459	55969	110923	248316	637860
Cloud	486161	285	413913	48386	130	6324	955199
Sum	486430	137209	499372	104355	111053	254640	1593059

Figure 39: Validation results of cloud/clear versus all S2 Hollstein dataset classes for Fmask 4.0 CCA

FORCE

Class	Cloud	(Shadow)	(Cirrus)	Snow	Water	Clear	Sum
Clear	12838	110099	103132	83470	111729	245126	666394
Cloud	473592	27196	396240	20885	25	9579	927517
Sum	486430	137295	499372	104355	111754	254705	1593911

Figure 40: Validation results of cloud/clear versus all S2 Hollstein dataset classes for FORCE

IdePix

Class	Cloud	(Shadow)	(Cirrus)	Snow	Water	Clear	Sum
Clear	8786	132025	49750	42912	110877	242621	586971
Cloud	477644	5270	449622	61443	877	12084	1006940
Sum	486430	137295	499372	104355	111754	254705	1593911

Figure 41: Validation results of cloud/clear versus all S2 Hollstein dataset classes for IdePix

InterSSIM

Class	Cloud	(Shadow)	(Cirrus)	Snow	Water	Clear	Sum
Clear	15581	136573	125492	93423	111754	254349	737172
Cloud	470849	722	373880	10932	0	356	856739
Sum	486430	137295	499372	104355	111754	254705	1593911

Figure 42: Validation results of cloud/clear versus all S2 Hollstein dataset classes for InterSSIM

LaSRC

Class	Cloud	(Shadow)	(Cirrus)	Snow	Water	Clear	Sum
Clear	16128	129070	6556	40	109922	220967	482683
Cloud	470302	8225	492816	104315	1832	33738	1111228
Sum	486430	137295	499372	104355	111754	254705	1593911

Figure 43: Validation results of cloud/clear versus all S2 Hollstein dataset classes for LaSRC

S2cloudless

Class	Cloud	(Shadow)	(Cirrus)	Snow	Water	Clear	Sum
Clear	11752	135916	94520	85134	110948	247469	685739
Cloud	474678	1379	404852	19221	806	7236	908172
Sum	486430	137295	499372	104355	111754	254705	1593911

Figure 44: Validation results of cloud/clear versus all S2 Hollstein dataset classes for S2cloudless

Sen2cor

Class	Cloud	(Shadow)	(Cirrus)	Snow	Water	Clear	Sum
Clear	34011	137261	107508	60851	111666	246859	698156
Cloud	452419	34	391864	43504	88	7846	895755
Sum	486430	137295	499372	104355	111754	254705	1593911

Figure 45 Validation results of cloud/clear versus all S2 Hollstein dataset classes for Sen2cor

3.4.1.5 L8 GSFC

Table 20 shows results of cloud detection algorithms for the GSFC L8 dataset. Overall, algorithms showed high performance for Landsat 8, though on 6 scenes.

Table 20: L8 GSFC results.

Algorithm	PA Non-cloud	PA Cloud	UA Non-cloud	UA Cloud	OA	Balanced OA	Number of pixels
ATCOR	99.77	94.84	95.18	99.75	97.33	97.3	864,419
CD-FCNN	99.99	94.58	94.97	99.99	97.32	97.29	864,419
Fmask 4.0 CCA	99.96	97.34	97.47	99.96	98.67	98.65	864,419
FORCE	99.75	96.53	96.71	99.73	98.16	98.14	864,419
LaSRC	98.15	94.75	95.03	98.04	96.47	96.45	864,419

3.4.1.6 L8 Pixbox dataset

Before looking at the numbers in details, it is important to note, that the analysis was not limited to the product area with thermal band coverage. This means these areas, where no thermal information is available, but all other bands deliver usable data, have been validated as well. This decision was taken, as not all users use the thermal bands in their applications. Some may only use the visible bands. For these users a cloud mask covering all pixels of non-thermal bands should be provided. For those algorithms limited to thermal data extent, it is advised to have an additional method in place to deal with these remaining pixels. Regarding the five compared algorithms, only ATCOR delivered a cloud mask covering all pixels, while the others have been limited to the thermal coverage. As validation samples had been collected in those non-thermal regions too, ATCOR has a slight head start when comparing numbers. This will be put into perspective in section 3.4.3.2. Additionally, it is important to notice, the dataset is quite imbalanced between cloud and non-cloud validation samples, with a 3 to 1 ratio of non-cloud to cloud pixels. In this case, looking at user accuracies, could be a bit misleading, but some conclusion can still be drawn from the numbers.

When analyzing BOA over all surfaces (see Table 21) and comparing it to land (Table 22) and water (Table 23), you can see, Fmask 4.0 CCA is performing best disregarding the surface type. But you can also see that all algorithms perform a good deal worse over water compared to land. Especially CD-FCNN and LaSRC having problems detecting clouds over water.

Additionally, the results reflect the size of a chosen buffer if any was used. It can be observed that Fmask 4.0 CCA, FORCE and LaSRC have higher UA non-cloud compared to cloud implying a commission error of non-cloud pixels in the cloud mask. For FORCE it is quite higher compared to Fmask 4.0 CCA, due to a bigger buffer size, while for LaSRC it is even higher. But this is mostly due to detection errors combined with a medium sized buffer.

Furthermore, the results show especially ATCOR and also CD-FCNN can be valuable for cloud conservative applications, delivering a high UA for cloud pixels. However, this is only achieved by omitting a good amount of cloud pixels.

Table 21: L8 PixBox results - complete dataset over all surfaces.

Algorithm	PA Non-cloud	PA Cloud	UA Non-cloud	UA Cloud	OA	Balanced OA	Number of pixels
ATCOR	99.2	73.3	90.8	97.2	92.08	86.25	20127
CD-FCNN	97.4	59	86.7	89.4	87.19	78.2	19496
Fmask 4.0 CCA	93.3	82.5	93.6	81.8	90.42	87.9	19496
FORCE	81.7	76.5	90.2	61.3	80.3	79.1	20128
LaSRC	87.7	47.8	81.6	59.5	76.75	67.75	20128

Table 22: L8 PixBox results - complete dataset over land surfaces only.

Algorithm	PA Non-cloud	PA Cloud	UA Non-cloud	UA Cloud	OA	Balanced OA	Number of pixels
ATCOR	98.5	85.5	93.3	96.5	94.24	92	11027
CD-FCNN	95	89.5	95.1	89.3	93.26	92.25	10536
Fmask 4.0 CCA	93.2	94.6	97.4	86.6	93.64	93.9	10536
FORCE	78.1	86.7	92.3	65.9	80.89	82.4	11028
LaSRC	78.8	71.8	85.1	62.3	76.48	75.3	11028

Table 23: L8 PixBox results - complete dataset over water surfaces only.

Algorithm	PA Non-cloud	PA Cloud	UA Non-cloud	UA Cloud	OA	Balanced OA	Number of pixels
ATCOR	99.9	49.9	88.3	99.4	89.46	74.9	9100
CD-FCNN	99.9	3.9	79.9	92.4	80.06	51.9	8960
Fmask 4.0 CCA	93.4	60.7	90.1	70.6	86.63	77.05	8960
FORCE	85.5	57.2	88.3	51.1	79.58	71.35	9100
LaSRC	96.9	2.1	78.9	15.2	77.08	49.5	9100

3.4.1.7 L8 Biome

Table 24 provides a summary of performance metrics for the L8Biome dataset. Results in this table should not be used directly for inter-comparing algorithms because of the following reasons: (i) ATCOR processed only 86 images out of 96 images, since images in polar regions were removed; (ii) LaSRC processed 80 images, since snow 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, which does not have valid values in all spectral bands.

Table 25 and Table 26 provide a correct inter-comparison between algorithms since the amount of reference scenes and pixels used was the same. Results in Table 26 do not include thin clouds in the L8Biome data. Not including CD-FCNN, the average BOA was 90.0±1.4% (Table 25) and 91.5±2.1% (Table 26). The reason not including CD-FCNN in the analysis is that deep learning algorithm partially used L8Biome in the training process. Like in the case of GSFC data, removing thin clouds from the reference increases BOA and Cloud-PA accuracies.

Table 24: Performance metrics of algorithms using the L8Biome data

					Cloud			Non-cloud		
Processor	**% cloud**	**Total num valid pixels**	**OA**	**BOA**	**PA**	**UA**	**F**	**PA**	**UA**	**F**
ATCOR	48.3	3,550,231,219	86.8	86.7	83.2	88.8	85.9	90.2	85.2	87.6
Fmask 4.0 CCA	47.9	3,963,655,082	90.0	90.2	93.6	86.6	90.0	86.7	93.6	90.0
FORCE	47.9	3,963,655,082	84.9	85.3	96.0	77.7	85.9	74.6	95.3	83.7
LaSRC	49.4	3,300,706,977	90.9	90.9	92.7	89.2	90.9	89.1	92.6	90.8
CD-FCNN	47.9	3,963,655,082	97.3	97.3	97.5	96.8	97.1	97.1	97.7	97.4

Table 25: Performance metrics of algorithms using the L8Biome data on the same set of Landsat 8 scenes

					Cloud			Non-cloud		
Processor	**% cloud**	**Total num valid pixels**	**OA**	**BOA**	**PA**	**UA**	**F**	**PA**	**UA**	**F**
ATCOR	49.4	3,300,706,977	88.2	88.2	84.6	90.9	87.6	91.7	85.9	88.7
Fmask 4.0 CCA	49.4	3,300,706,977	91.3	91.4	96.2	87.4	91.6	86.5	95.9	91.0
FORCE	49.4	3,300,706,977	89.4	89.5	96.8	84.2	90.0	82.2	96.3	88.7
LaSRC	49.4	3,300,706,977	90.9	90.9	92.7	89.2	90.9	89.1	92.6	90.8
CD-FCNN	49.4	3,300,706,977	97.4	97.4	97.8	96.9	97.4	97.0	97.8	97.4

Table 26: Performance metrics of algorithms using the L8Biome data on the same set of Landsat 8 scenes without considering thin clouds

					Cloud			Non-cloud		
Processor	**% cloud**	**Total num valid pixels**	**OA**	**BOA**	**PA**	**UA**	**F**	**PA**	**UA**	**F**
ATCOR	42.6	2,909,423,820	89.6	89.2	86.8	88.6	87.7	91.7	90.3	91.0
Fmask 4.0 CCA	42.6	2,909,423,820	92.1	93.1	99.7	84.6	91.5	86.5	99.7	92.7
FORCE	42.6	2,909,423,820	89.0	90.2	98.1	80.4	88.4	82.2	98.3	89.6
LaSRC	42.6	2,909,423,820	92.8	93.5	97.8	86.9	92.1	89.1	98.2	93.4
CD-FCNN	42.6	2,909,423,820	98.2	98.4	99.8	96.1	97.9	97.0	99.8	98.4

3.4.2 Pixel based inter-dataset comparison & validation dataset comparison

In this section the balanced overall accuracy (BOA) and user accuracy (UA) for cloud and non-cloud pixels of all algorithms are compared across all datasets. The reference datasets are also compared with each other based on the performance of the algorithms and characteristics of the reference datasets. This comparison has been made because the analysis had shown that the performance of each algorithm changes between the different reference datasets. For some algorithms, the performance measures varied from very good to reasonably poor. Since the reference datasets are used as the truth for the intercomparison, a comparable performance of the algorithms throughout the different reference datasets was expected. Therefore, this huge change in performance needs to be analyzed and understood, in order to draw more robust conclusions.

In addition, this very compressed form of presentations allows a good comparison of the single algorithms. The three indicators (BOA, UA cloud, and UA non-cloud) have been chosen as they are good indicators for the three main user needs: balanced, cloud conservative, non-cloud conservative.

It was decided to compare results excluding and including thin- transparent clouds, as the detection of these types of clouds seems to be challenging for some algorithms. Furthermore, the definition of cloud or non-cloud for these thin cloud pixels is disputed widely as it is a smooth transition between both states. Depending on the application very thin clouds can even be corrected and do not need to be detected. Since the application is not known, clear sky conservative applications have to be considered (like aerosol retrieval) and thus the detection of thin clouds has to be analyzed.

3.4.2.1 Sentinel-2 balanced overall accuracy

Results without thin clouds:

Comparing the average BOA (see Table 27) of all algorithms between the different reference datasets shows a ~10% difference. The algorithms perform a lot better on GSFC and Hollstein, compared to the PixBox dataset. The differences can be explained partially. While in the PixBox dataset the thin clouds consist only of very thin semi-transparent clouds leaving medium-transparent to opaque clouds in the dataset, the Hollstein dataset has only two categories, opaque and transparent (class: cirrus). This means, removing transparent clouds leads to only opaque clouds remaining, which are easily detected by all algorithms. The GSFC dataset is comparable with the Hollstein dataset in the sense that transparent clouds comprise all levels of transparency, not only very thin semi-transparent clouds, leading to mostly opaque clouds. The CESBIO dataset did not allow the separation of thin clouds and opaque clouds.

These findings give the following indications:

1. Removing thin clouds from the datasets and comparing these results to those incl. thin clouds give an indication for the ability of each algorithm to detect thin transparent clouds.
2. The differences in classifying thin clouds and the ability to remove these accordingly from the datasets, vary so much, that the datasets become quite incomparable after removal of thin clouds.

Based on the second finding, no inter-dataset comparison is made.

Table 27: Comparison of balanced overall accuracies of all algorithms across all (Sentinel-2) reference datasets excluding thin clouds. The three best performing algorithms per dataset are highlighted in green and the least performing are highlighted in orange.

Reference dataset comparison						
Algo	**GSFC**	**PixBox**	**Hollstein**	**Average**	**median**	**Std**
ATCOR	88.20	80.35	88.45	**84.35**	**84.30**	**3.98**
CD-FCNN	93.60	88.05	97.80	**89.74**	**90.83**	**6.85**
Fmask 4.0 CCA	96.50	89.90	95.40	**92.68**	**92.65**	**3.32**
FORCE	96.10	85.80	93.95	**91.19**	**91.43**	**4.06**
Idepix	92.50	81.00	92.55	**88.24**	**89.70**	**4.77**
InterSSIM	96.00	90.70	97.40	**93.03**	**93.35**	**3.83**
LaSRC	97.90	71.00	86.15	**84.44**	**84.43**	**9.59**
MAJA	92.20	88.30		**90.33**	**90.50**	**1.60**
S2cloudless	96.10	91.60	96.45	**93.24**	**93.85**	**3.20**
sen2cor	95.40	84.80	92.25	**89.29**	**88.53**	**4.67**
Average	**94.45**	**85.15**	**93.38**			
median	**95.70**	**86.93**	**93.95**			
Std	**2.71**	**5.95**	**3.77**			
No. of pixels	~ 40.95 Mio	13,986	1.09 Mio			
No. of products	28	29	59			
Average No. Of pixels per product	~ 1.46 Mio	482	18,552			

Results incl. thin clouds:

When including thin clouds (see Table 28) in the analysis the mean performance of the algorithms across all datasets decreases around three to seven percent. It also becomes obvious that including semitransparent clouds makes the datasets more comparable, as the difference of mean performance between the datasets is now only three to seven percent instead of the prior ten percent, and the standard deviation of all algorithms across the datasets decreases.

Including thin clouds also shows that some algorithms seem to have issue with detecting semi-transparent clouds. The biggest decreases can be found for CD-FCNN, InterSSIM, and S2cloudless. While these algorithms performed good without semitransparent cloud, the mean performance across the datasets with thin clouds decreases around five to six percent, but still showing an overall good performance.

When evaluating the performance of each algorithm across all reference datasets mean and median performance give a good first indication about the general performance. In addition, the standard deviation must be considered showing the stability of an algorithm across the reference datasets. While for example Fmask 4.0 CCA shows a good (88.66%) mean performance, the performance varies only little between the different datasets. On the other side of the spectrum, there are algorithms like ATCOR performing not as good (82.28% mean BOA) and additionally vary quite a bit more. The standard deviation also helps identifying those algorithms performing disproportionally good on a single dataset, like LaSRC on the GSFC dataset or MAJA on the CESBIO dataset.

Considering BOA as a good first indicator for users with a balanced need on cloud mask performance, Fmask 4.0 CCA, FORCE, InterSSIM and S2cloudless should be favored algorithms. But BOA is only one indicator to evaluate the usability for balanced performance demands, as a high BOA still can be achieved by disproportionally high UA for one class. This will be evaluated in the upcoming sections.

Table 28: Comparison of balanced overall accuracies of all algorithms across all (Sentinel-2) reference datasets (incl. thin clouds for GSFC, PixBox, and Hollstein and excluding snow for PixBox). The three best performing algorithms per dataset are highlighted in green and the least performing are highlighted in orange.

Reference dataset comparison							
Algo	**CESBIO**	**GSFC**	**PixBox**	**Hollstein**	**Average**	**median**	**Std**
ATCOR	80.40	81.70	77.15	89.85	**82.28**	**81.05**	**4.68**
CD-FCNN	79.50	85.40	80.40	84.20	**82.38**	**82.30**	**2.48**
Fmask 4.0 CCA	88.90	88.40	86.25	91.10	**88.66**	**88.65**	**1.72**
FORCE	88.90	88.20	82.10	89.35	**87.14**	**88.55**	**2.94**
Idepix	86.90	86.10	84.00	90.50	**86.88**	**86.50**	**2.35**
InterSSIM	88.00	87.60	84.85	91.85	**88.08**	**87.80**	**2.49**
LaSRC	82.70	96.80	74.15	86.65	**85.08**	**84.68**	**8.14**
MAJA	90.50	80.70	86.10		**85.77**	**86.10**	**4.01**
S2cloudless	88.80	87.70	86.70	92.25	**88.86**	**88.25**	**2.09**
sen2cor	84.70	87.80	82.05	88.55	**85.78**	**86.25**	**2.59**
Average	**85.93**	**87.04**	**82.38**	**89.37**			
median	**87.45**	**87.65**	**83.05**	**89.85**			
Std	**3.69**	**4.15**	**3.94**	**2.44**			
No. of pixels	~ 3.08 Bil	~ 46.26 Mio	13,490	1,517,387			
No. of products	30.00	28.00	29.00	59.00			
Average No. Of pixels per product	~ 102.94 Mio	~ 1.65 Mio	465	25,718			

3.4.2.2 Sentinel-2 user accuracy non-cloud

Results without thin clouds:

The results excluding thin clouds in Table 29 are only shown for completeness. As explained already in section 3.4.2.1, excluding thin clouds leads to quite incomparable datasets. The results without thin clouds can still be used to compare with the results including thin clouds in Table 30.

Table 29: Comparison of user accuracies for non-cloud classified pixels of all algorithms across all (Sentinel-2) reference datasets excluding thin clouds. The three best performing algorithms per dataset are highlighted in green and the least performing are highlighted in orange.

Reference dataset comparison						
Algo	**GSFC**	**PixBox**	**Hollstein**	**Average**	**median**	**Std**
ATCOR	77.30	83.10	86.40	**84.05**	**84.75**	**4.49**
CD-FCNN	86.30	90.30	98.60	**90.95**	**89.45**	**4.64**
Fmask 4.0 CCA	92.30	94.30	100.00	**95.13**	**94.10**	**2.91**
FORCE	93.00	93.60	97.70	**94.83**	**94.30**	**1.81**
Idepix	91.30	96.10	98.40	**94.70**	**94.55**	**2.74**
InterSSIM	91.30	92.20	97.50	**93.58**	**92.75**	**2.37**
LaSRC	98.10	93.10	96.60	**95.60**	**95.60**	**1.90**
MAJA	92.70	96.40		**95.47**	**96.40**	**1.99**
S2cloudless	91.90	94.90	98.00	**94.68**	**94.40**	**2.20**
sen2cor	90.10	89.60	94.20	**91.38**	**90.85**	**1.79**
Average	**90.43**	**92.36**	**96.38**			
median	**91.60**	**93.35**	**97.70**			
Std	**5.17**	**3.74**	**3.83**			
No. of pixels	~ 40.95 Mio	13,986	1.09 Mio			
No. of products	28	29	59			
Average No. Of pixels per product	~ 1.46 Mio	482	18,552			

Results incl. thin clouds:

When comparing the results for UA non-cloud (Table 30) with the results for BOA (Table 28), it quickly becomes obvious, that an overall good performance does not necessarily imply a good performance for a user having non-cloud conservative needs. It also shows that algorithms overall not achieving the best performance, can still be good at delivering non-cloud observations, with only low levels of cloud contamination. Needless to say, algorithms that do not achieve a high overall performance but good performance in UA of one class are imbalanced. This imbalance leads to commissioning errors in the opposed class. While this behavior should be minimized in general, for a certain set of users (non-cloud conservative) this could be favored.

While InterSSIM for example had an overall good performance, it seems to struggle to provide non-cloud observations not contaminated by clouds. On the other hand, IdePix, and LaSRC, performing average to low in terms of BOA, are quite capable of providing non-cloud observation with low levels of cloud contaminations. Fmask 4.0 CCA, and FORCE performing very good in terms of BOA still performs quite good for UA non-cloud.

Thus, for a non-cloud conservative user, considering UA non-cloud as a good indicator, Fmask 4.0 CCA, IdePix, LaSRC, MAJA, and FORCE should be the favorable algorithms over all datasets.

Table 30: Comparison of user accuracies for non-cloud classified pixels of all algorithms across all (Sentinel-2) reference datasets (incl. thin clouds for GSFC, PixBox, and Hollstein and excluding snow for PixBox). The three best performing algorithms per dataset are highlighted in green and the least performing are highlighted in orange.

Reference dataset comparison							
Algo	**CESBIO**	**GSFC**	**PixBox**	**Hollstein**	**Average**	**median**	**Std**
ATCOR	89.40	64.00	69.80	79.40	**75.65**	**74.60**	**9.66**
CD-FCNN	88.60	69.10	72.90	67.10	**74.43**	**71.00**	**8.44**
Fmask 4.0 CCA	93.90	73.90	81.80	86.60	**84.05**	**84.20**	**7.27**
FORCE	95.00	74.60	79.50	82.60	**82.93**	**81.05**	**7.53**
Idepix	93.00	75.10	83.30	90.00	**85.35**	**86.65**	**6.88**
InterSSIM	93.30	72.50	77.60	80.90	**81.08**	**79.25**	**7.67**
LaSRC	94.60	94.50	79.80	95.30	**91.05**	**94.55**	**6.50**
MAJA	97.30	74.40	89.20		**86.97**	**89.20**	**9.48**
S2cloudless	93.90	73.00	82.10	84.50	**83.38**	**83.30**	**7.44**
sen2cor	91.60	72.80	78.30	79.70	**80.60**	**79.00**	**6.85**
Average	**93.06**	**90.43**	**79.43**	**82.90**			
median	**93.60**	**91.60**	**79.65**	**82.60**			
Std	**2.47**	**5.17**	**5.13**	**7.43**			
No. of pixels	~ 3.08 Bil	~ 46.26 Mio	13,490	1,517,387			
No. of products	30	28	29	59			
Average No. Of pixels per product	~ 102.94 Mio	~ 1.65 Mio	465	25,718			

3.4.2.3 Sentinel-2 user accuracy cloud

Results without thin clouds:

The results excluding thin clouds in Table 31 are only shown for completeness. As explained already in section 3.4.2.1, excluding thin clouds leads to quite incomparable datasets. The results without thin clouds can still be used to compare with the results including thin clouds in Table 32.

Table 31: Comparison of user accuracies for cloud classified pixels of all algorithms across all (Sentinel-2) reference datasets excluding thin clouds. The three best performing algorithms per dataset are highlighted in green and the least performing are highlighted in green and the least performing are highlighted in orange.

Reference dataset comparison						
Algo	**GSFC**	**PixBox**	**Hollstein**	**Average**	**median**	**Std**
ATCOR	100.00	81.40	93.20	**89.88**	**89.05**	**7.25**
CD-FCNN	99.90	87.90	96.70	**94.65**	**95.40**	**4.41**
Fmask 4.0 CCA	99.70	82.70	89.80	**90.75**	**90.30**	**6.04**
FORCE	98.50	73.60	89.10	**85.28**	**84.50**	**9.42**
Idepix	93.60	62.40	85.70	**82.15**	**86.30**	**11.79**
InterSSIM	99.70	91.30	97.50	**95.40**	**95.30**	**3.35**
LaSRC	97.80	51.30	76.00	**70.68**	**66.80**	**18.10**
MAJA	92.70	74.30		**79.90**	**74.30**	**9.07**
S2cloudless	99.30	86.40	94.30	**92.55**	**92.25**	**4.80**
sen2cor	99.70	78.60	89.80	**89.20**	**89.25**	**7.47**
Average	**98.09**	**76.99**	**90.23**			
median	**99.50**	**80.00**	**89.80**			
Std	**2.56**	**11.66**	**6.19**			
No. of pixels	~ 40.95 Mio	13,986	1.09 Mio			
No. of products	28	29	59			
Average No. Of pixels per product	~ 1.46 Mio	482	18,552			

Results incl. thin clouds:

Similar to the comparison between UA non-cloud with BOA, when comparing the results for UA cloud (Table 32) with the results for BOA (Table 28), it again quickly becomes obvious, that an overall good performance does not necessarily imply a good performance for a user having cloud conservative needs. It also shows again algorithms overall not achieving the best performance, can still be good at delivering cloud observations, with only low levels of non-cloud contamination. The limitations that come with this tendency have been already described in the previous section. While this imbalanced behavior should be minimized in general, for a certain set of users (cloud conservative) this could be favored.

When comparing UA cloud (Table 32) with UA non-cloud (Table 30) adverse performances can be recognized. But this is just a logical consequence of an imbalanced behavior. Algorithms like IdePix, LaSRC and MAJA, that have shown a good performance for UA non-cloud, show a weak performance for UA cloud. In contrast weak UA non-cloud performing algorithms like CD-FCNN or InterSSIM are very suitable for cloud conservative users, delivering cloud observations with only little clear contaminations.

With this knowledge of performance of UA non-cloud and cloud, as well as the results for BOA, we could revisit the question for best suitability for balanced user requirements. While Fmask 4.0 CCA, FORCE, InterSSIM and S2cloudless had shown good BOA performance, the analysis of UA had shown, that only Fmask 4.0 CCA and S2cloudless achieve this and still be balanced between cloud and non-cloud accuracy.

Table 32: Comparison of user accuracies for cloud classified pixels of all algorithms across all (Sentinel-2) reference datasets (incl. thin clouds for GSFC, PixBox, and Hollstein and excluding snow for PixBox). The three best performing algorithms per dataset are highlighted in green and the least performing are highlighted in orange

Reference dataset comparison							
Algo	**CESBIO**	**GSFC**	**PixBox**	**Hollstein**	**Average**	**median**	**Std**
ATCOR	84.90	100.00	87.10	96.50	**92.13**	**91.80**	**6.30**
CD-FCNN	94.10	99.90	91.30	97.70	**95.75**	**95.90**	**3.30**
Fmask 4.0 CCA	90.80	99.70	91.00	94.20	**93.93**	**92.60**	**3.60**
FORCE	79.90	98.60	84.70	93.80	**89.25**	**89.25**	**7.35**
Idepix	86.90	93.90	84.70	92.10	**89.40**	**89.50**	**3.74**
InterSSIM	93.10	99.70	94.50	98.60	**96.48**	**96.55**	**2.75**
LaSRC	57.60	98.20	71.40	86.70	**78.48**	**79.05**	**15.35**
MAJA	72.70	93.00	83.20		**82.97**	**83.20**	**8.29**
S2cloudless	90.20	99.30	91.60	96.80	**94.48**	**94.20**	**3.72**
sen2cor	88.70	99.70	86.10	94.30	**92.20**	**91.50**	**5.25**
Average	**83.89**	**98.20**	**86.56**	**94.52**			
median	**87.80**	**99.50**	**86.60**	**94.30**			
Std	**10.69**	**2.44**	**6.16**	**3.39**			
No. of pixels	~ 3.08 Bil	~ 46.26 Mio	13,490	1,517,387			
No. of products	30	28	29	59			
Average No. Of pixels per product	~ 102.94 Mio	~ 1.65 Mio	465	25,718			

3.4.2.4 *Landsat 8 balanced overall accuracy*

The results for Landsat 8 are a lot harder to interpret compared to Sentinel-2, as for Landsat 8, there is a strong difference between the datasets. Especially on the GSFC dataset all algorithms perform suspiciously well, with an average performance over 95% over all indicators (BOA, UA cloud, and UA non-cloud). This high accuracy suggests the GSFC dataset is relatively simply, especially when looking at some examples from the visual analysis (section 3.4.3.2) it will become quite obvious that no algorithm has an actual accuracy of above approx. 92%. The PixBox dataset on the other hand seems to be a challenging dataset for the algorithms with comparable low performances.

In section 3.4.3.2 we will show that the PixBox dataset is quite representative, even though it has an emphasis on coastal regions.

Reference dataset comparison						
Algo	**L8Biome**	**GSFC**	**PixBox**	**Average**	**median**	**Std**
ATCOR	88.20	97.30	86.25	**90.58**	**88.20**	**4.82**
CD-FCNN		97.29	78.20	**90.96**	**97.29**	**9.03**
Fmask 4.0 CCA	91.40	98.65	87.90	**92.65**	**91.40**	**4.48**
FORCE	89.50	98.14	79.10	**88.91**	**89.50**	**7.78**
LaSRC	90.90	96.45	67.75	**85.03**	**90.90**	**12.43**
Average	**91.48**	**97.57**	**79.84**			
median	**90.90**	**97.30**	**79.10**			
Std	**3.16**	**0.76**	**7.15**			
No. of pixels	~ 3.3 Bil	864,419	13,490			
No. of products	30.00	28.00	29.00			
Average No. Of pixels per product	~ 110 Mio	30872	465			

3.4.2.5 *Landsat 8 user accuracy non-cloud*

Reference dataset comparison						
Algo	**L8Biome**	**GSFC**	**PixBox**	**Average**	**median**	**Std**
ATCOR	85.90	95.18	90.80	**90.63**	**90.80**	**3.79**
CD-FCNN		94.97	86.70	**93.16**	**94.97**	**4.71**
Fmask 4.0 CCA	95.90	97.47	93.60	**95.66**	**95.90**	**1.59**
FORCE	96.30	96.71	90.20	**94.40**	**96.30**	**2.98**
LaSRC	92.60	95.03	81.60	**89.74**	**92.60**	**5.84**
Average	**93.70**	**95.87**	**88.58**			
median	**95.90**	**95.18**	**90.20**			
Std	**4.25**	**1.03**	**4.12**			
No. of pixels	~ 3.3 Bil	864,419	13,490			
No. of products	30.00	28.00	29.00			
Average No. Of pixels per product	~ 110 Mio	30872	465			

3.4.2.6 *Landsat 8 user accuracy cloud*

Reference dataset comparison						
Algo	**L8Biome**	**GSFC**	**PixBox**	**Average**	**median**	**Std**
ATCOR	90.90	99.75	97.20	**95.95**	**97.20**	**3.72**
CD-FCNN		99.99	89.40	**95.43**	**96.90**	**4.45**
Fmask 4.0 CCA	87.40	99.96	81.80	**89.72**	**87.40**	**7.59**
FORCE	84.20	99.73	61.30	**81.74**	**84.20**	**15.78**
LaSRC	89.20	98.04	59.50	**82.25**	**89.20**	**16.48**
Average	**89.72**	**99.49**	**77.84**			
median	**89.20**	**99.75**	**81.80**			
Std	**4.22**	**0.73**	**15.06**			
No. of pixels	~ 3.3 Bil	864,419	13,490			
No. of products	30.00	28.00	29.00			
Average No. Of pixels per product	~ 110 Mio	30872	465			

3.4.3 Visual analysis

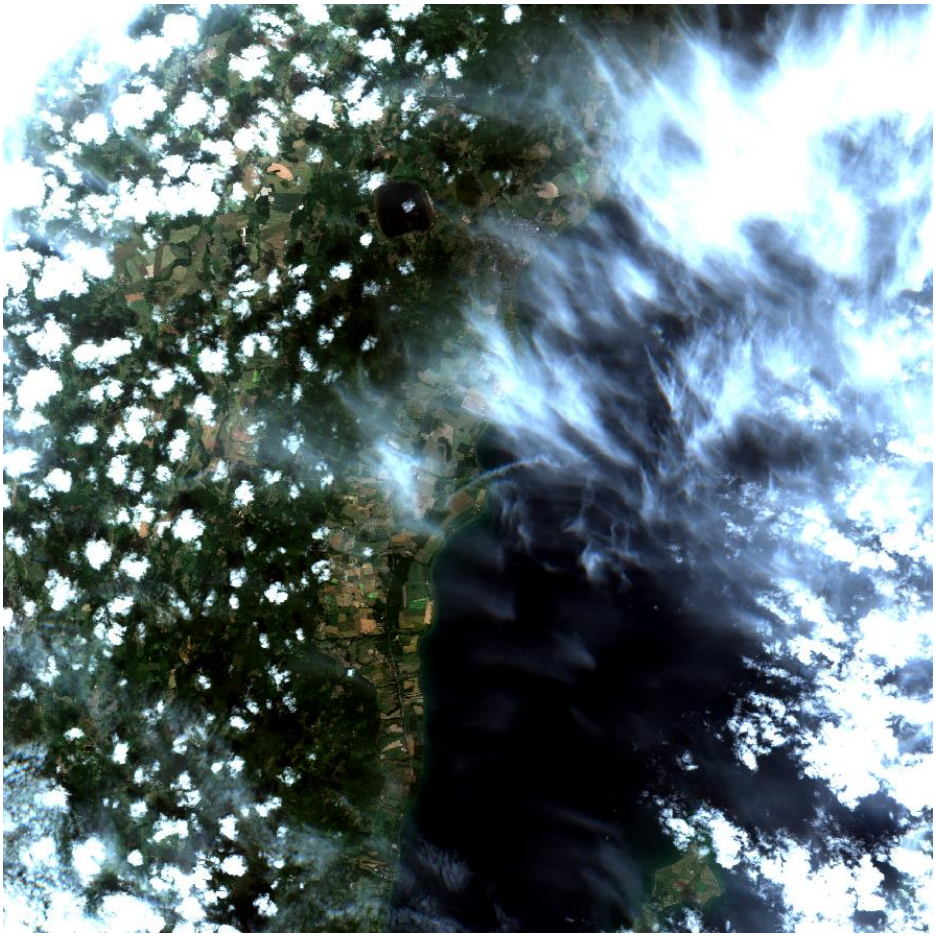
The goal of the visual analysis was to confirm or disprove the findings from the statistical analysis (pixel-based validation) and to identify specific behaviors or shortcomings of each algorithm. The later part is important for the algorithm producers to get feedback on certain features of the algorithm they might not be aware of, like systematic false detections, systematic under detections or misclassifications of certain surfaces.

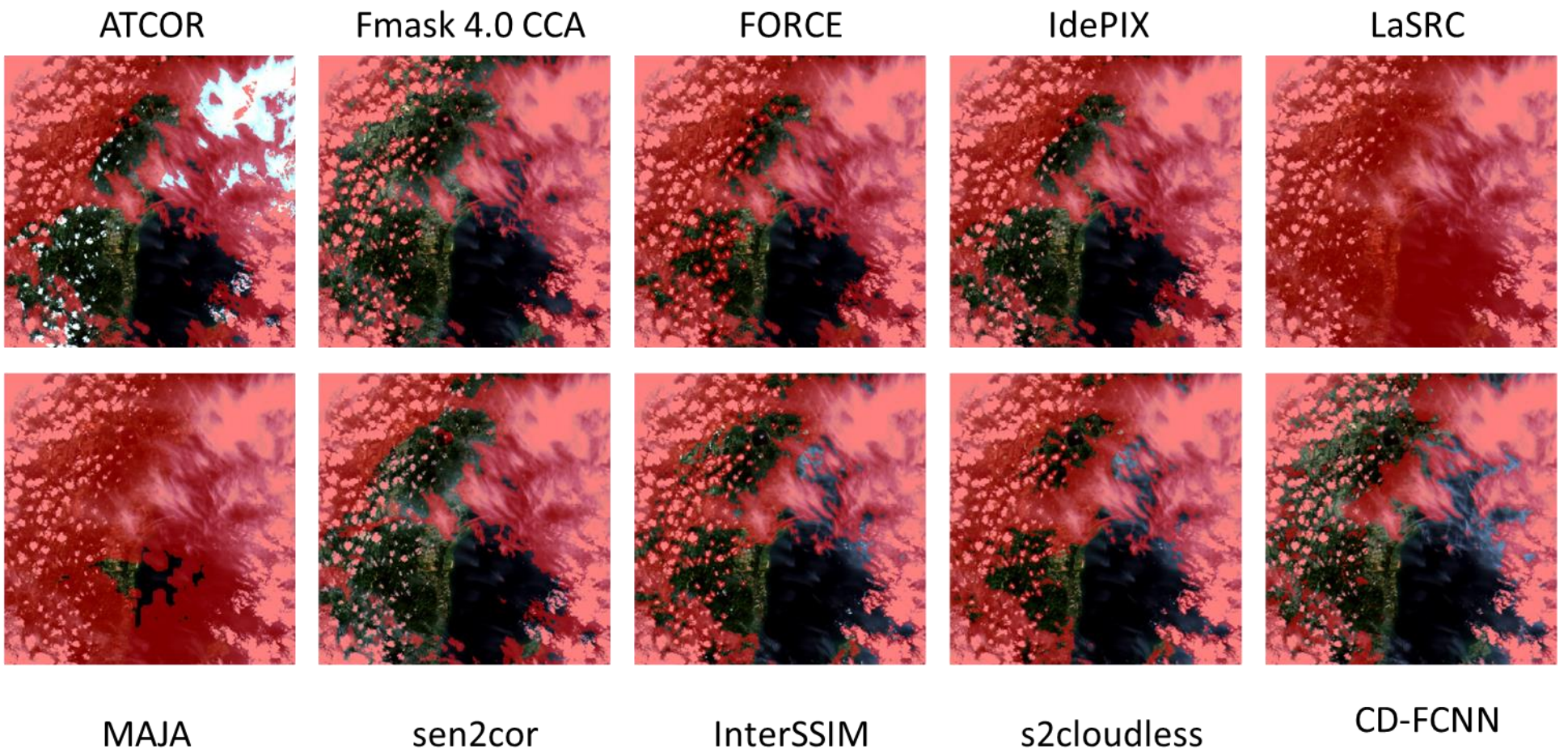
As the PixBox dataset for Sentinel-2 and Landsat 8 has been the dataset giving the most detail in terms of additional information for each pixel, these datasets have been used for visual analysis.

3.4.3.1 S2 Pixbox dataset

S2A_MSIL1C_20170726T102021_N0205_R065_T33VVE_20170726T102259:

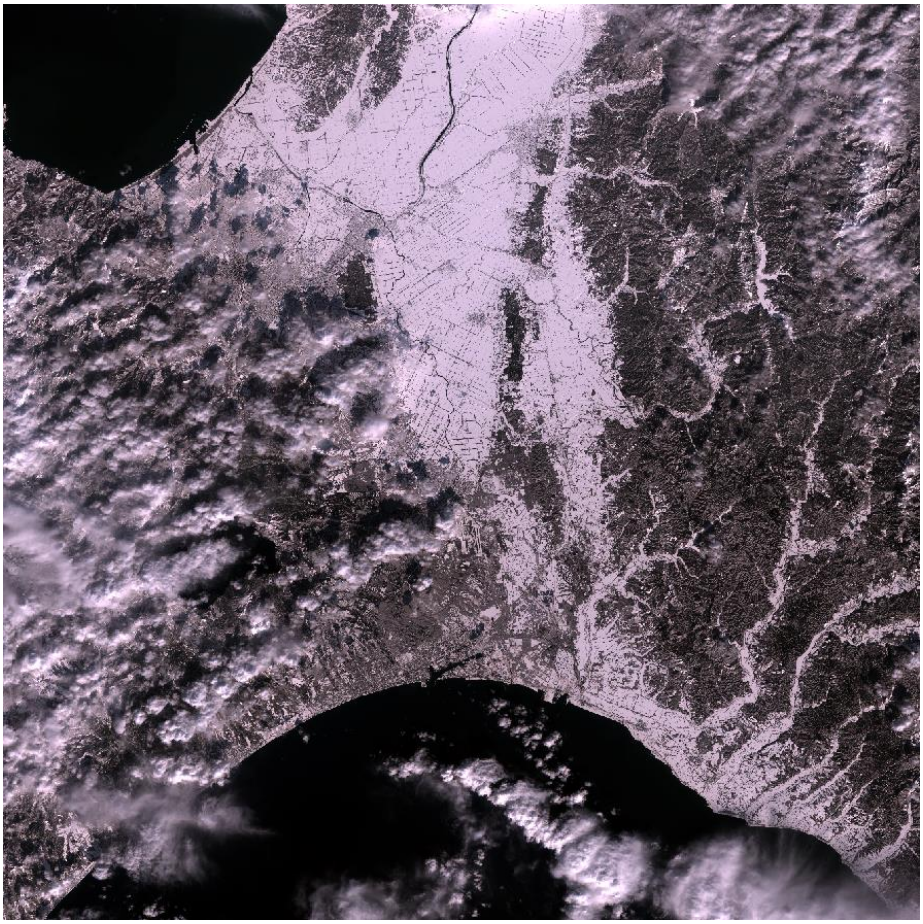
- Thin to medium semi-transparent clouds over water and land
- Small opaque cumulus clouds over land

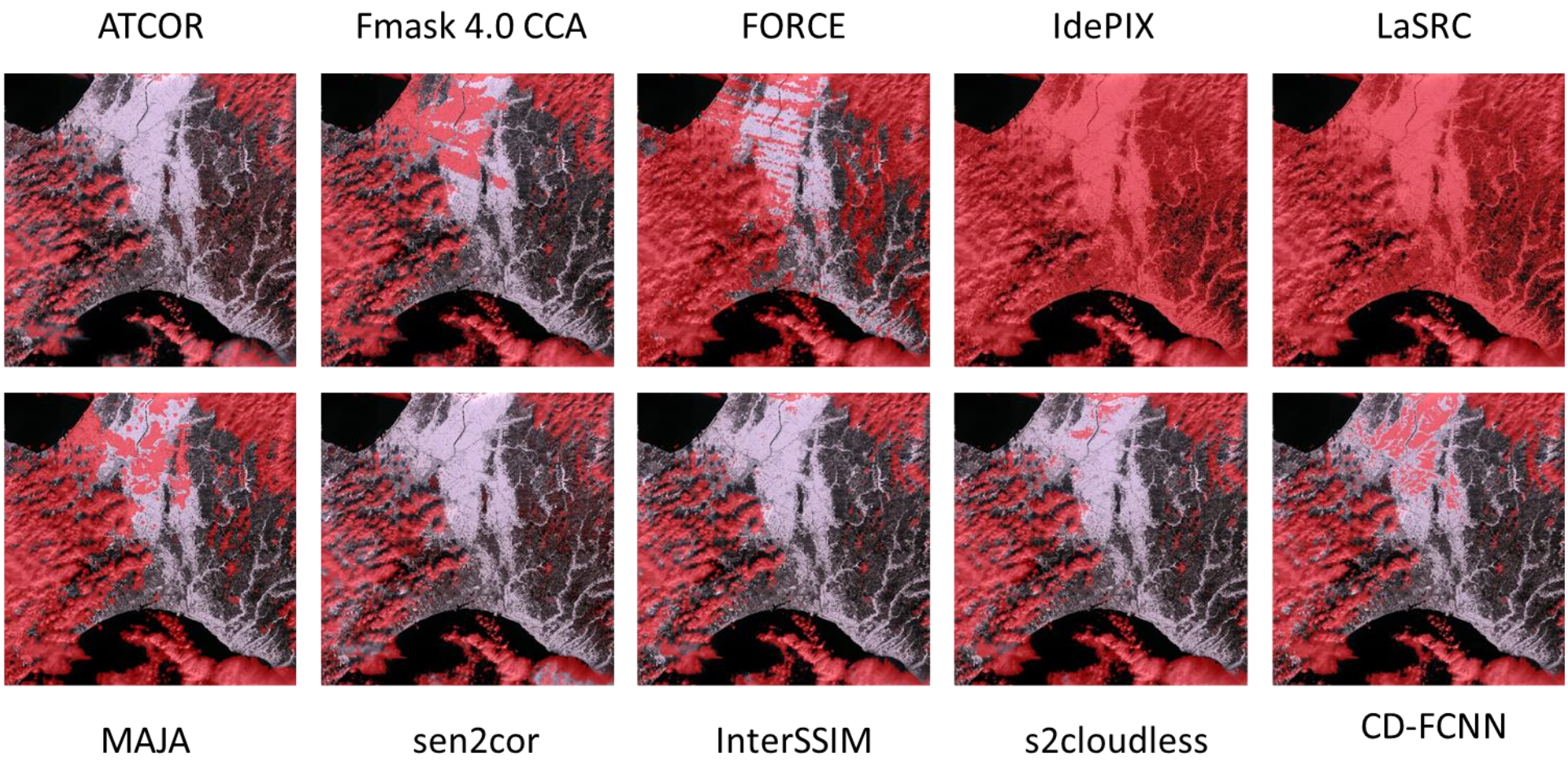



ATCOR
Fmask 4.0 CCA
FORCE
IdePIX
LaSRC
MAJA
sen2cor
InterSSIM
s2cloudless
CD-FCNN

S2A_MSIL1C_20180222T012651_N0206_R074_T54TWN_20180222T031349

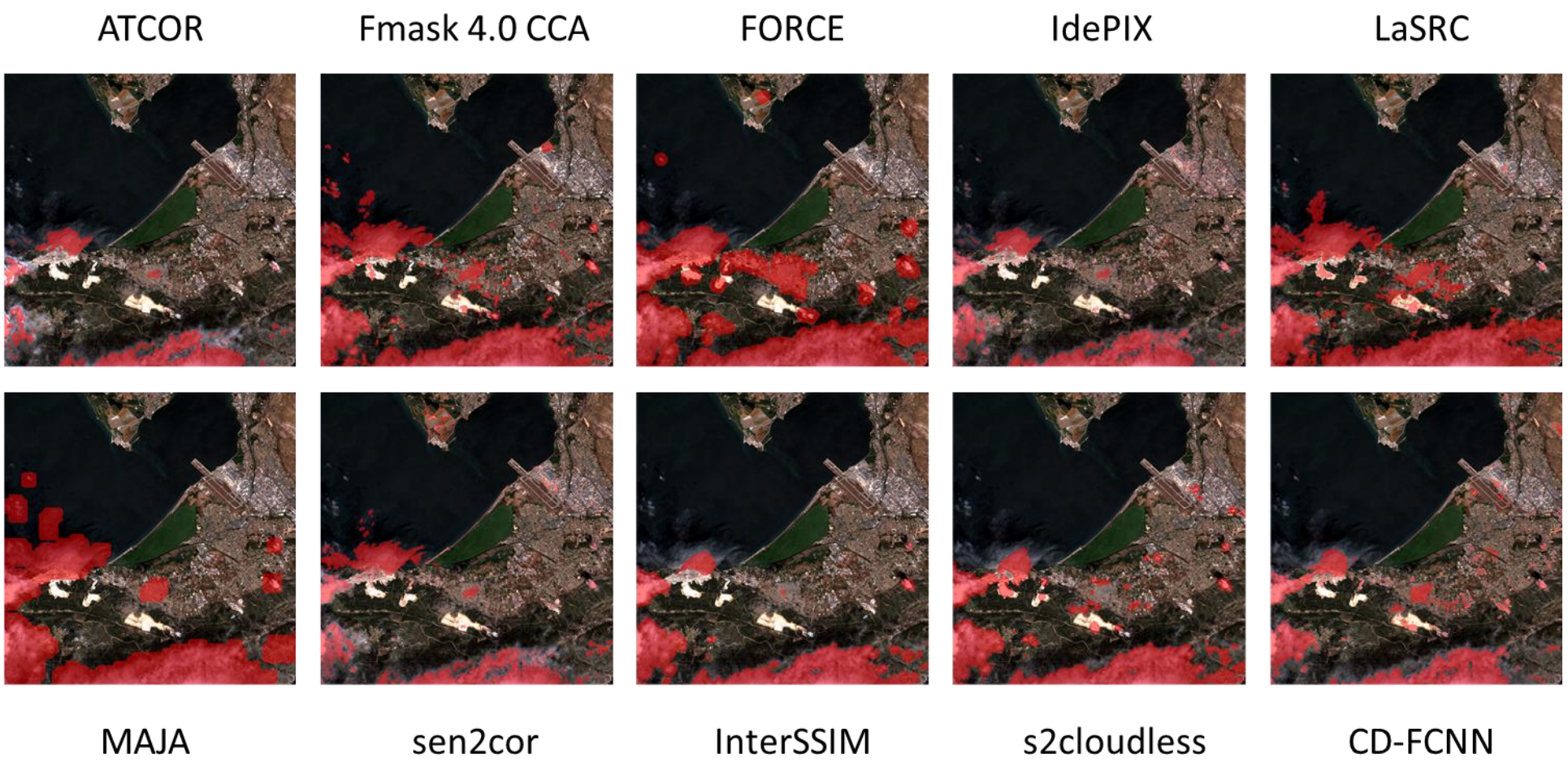
- Opaque clouds over snow




ATCOR
Fmask 4.0 CCA
FORCE
IdePIX
LaSRC
MAJA
sen2cor
InterSSIM
s2cloudless
CD-FCNN

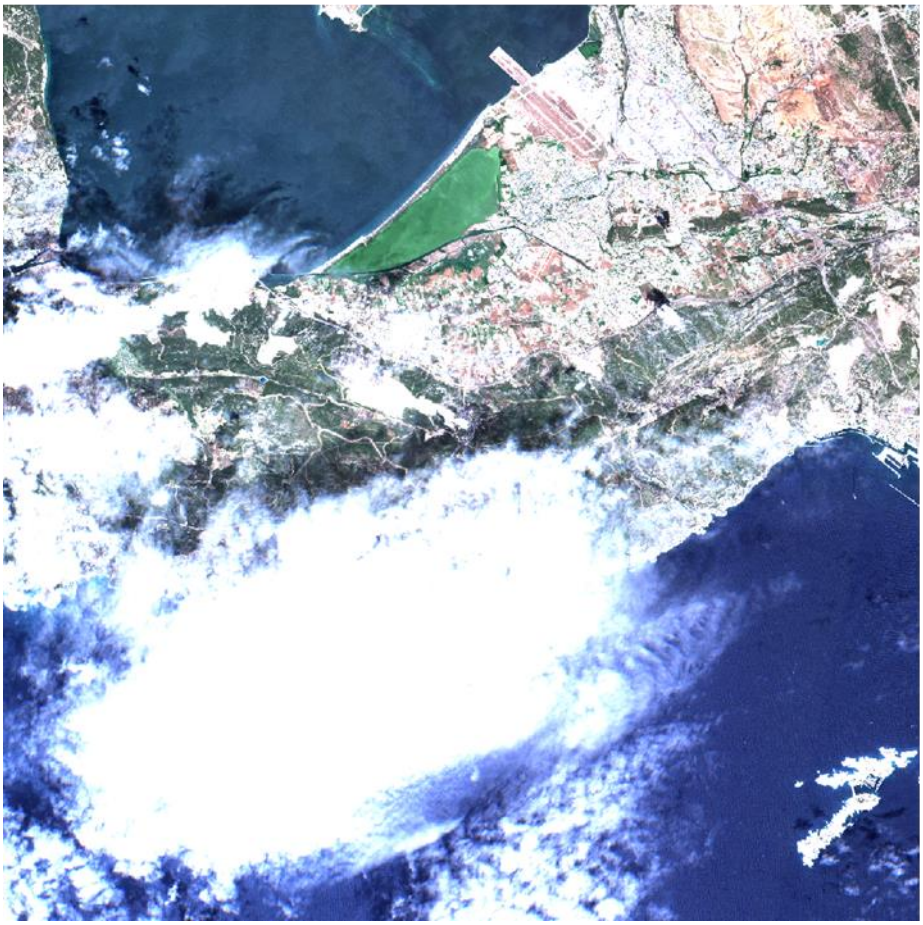
S2A_MSIL1C_20170629T103021_N0205_R108_T31TFJ_20170629T103020

- Opaque clouds to semi-transparent clouds and urban

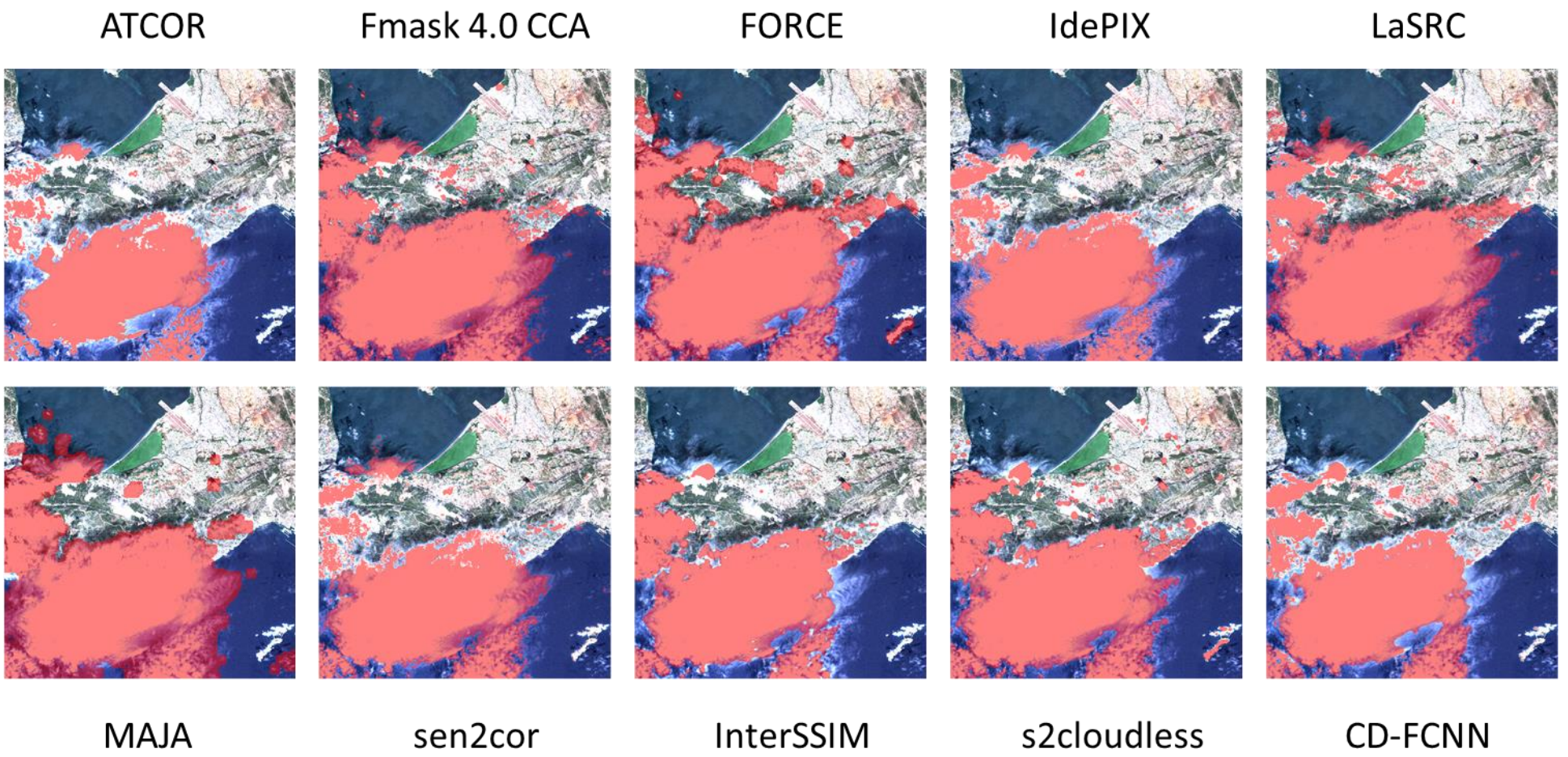

ATCOR
Fmask 4.0 CCA
FORCE
IdePIX
LaSRC
MAJA
sen2cor
InterSSIM
s2cloudless
CD-FCNN

S2A_MSIL1C_20170629T103021_N0205_R108_T31TFJ_20170629T103020

- Opaque clouds to semi-transparent clouds over water



RGB stretch:
0.14 – 0.16


ATCOR
Fmask 4.0 CCA
FORCE
IdePIX
LaSRC
MAJA
sen2cor
InterSSIM
s2cloudless
CD-FCNN

3.4.3.2 L8 Pixbox dataset

One interesting finding from the visual analysis of the Landsat 8 pixbox data was ATCOR being the only algorithm providing a cloud detection for all the products pixels, incl. the areas without thermal coverage. This finding is impossible to make when doing only a quantitative assessment. Nevertheless, this helps to put the results from the quantitative (confusion matrix) analysis in a relation, since the overall performance of ATCOR appears to be average with a very good amount of omission of clouds. Thus, ATCOR benefits of these additionally collected clouds in the non-thermal part of the product.

Another interesting finding was the very comparable detection of clouds from Fmask and FORCE, with FORCE having a higher commission error, due to a bigger buffer and over detection of bright surfaces. While the omission error of clouds seemed very much alike. This was surprising, as the numbers presented in sections 3.4.1.6 and 3.4.2.5 implied differently, with FORCE having a bit lower performance. This seems to be caused mostly by the size of the buffer, being much bigger for FORCE, as shown in Figure 46.

Figure 46: Comparison of Fmask 4.0 CCA and FORCE cloud buffer size

Visual analysis has also revealed semi-transparent clouds mostly being collected over water for the validation dataset. These semi-transparent clouds mostly consist of cirrus clouds This circumstance explains the low performance of CD-FCNN and LaSRC over water, shown in Table 23 of section 3.4.1.6, as both algorithms seem not to be using the cirrus band (1370 nm) for cloud detection. The visual analysis has shown that the performance of CD-FCNN and LaSRC, when neglecting cirrus clouds, is a lot better over water as the numbers in the confusion matrix have shown. Neglecting cirrus clouds, the performance is comparable to ATCOR.

The following main behaviors have been identified for each algorithm.

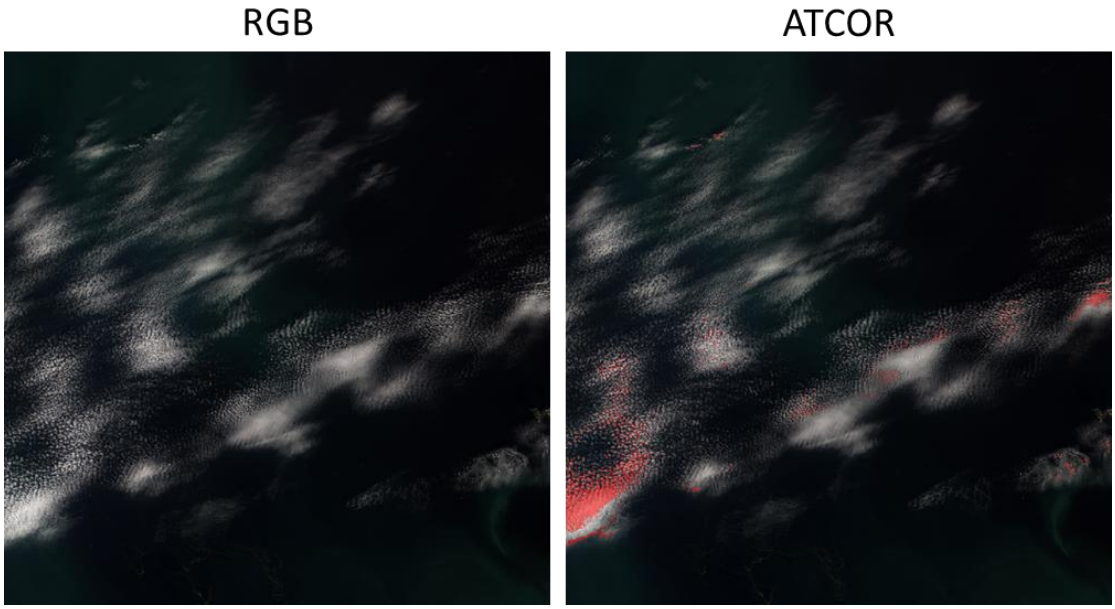
ATCOR:

Findings:

- Cloud conservative approach. Only little commission of non-clouds as cloud.
- Problem of detection small clouds, especially medium to thick semi-transparent cumulus and stratocumulus
- Missing fully opaque cloud parts
- In general under-detection of clouds

Recommendations:

- A buffer could improve performance quite a lot.



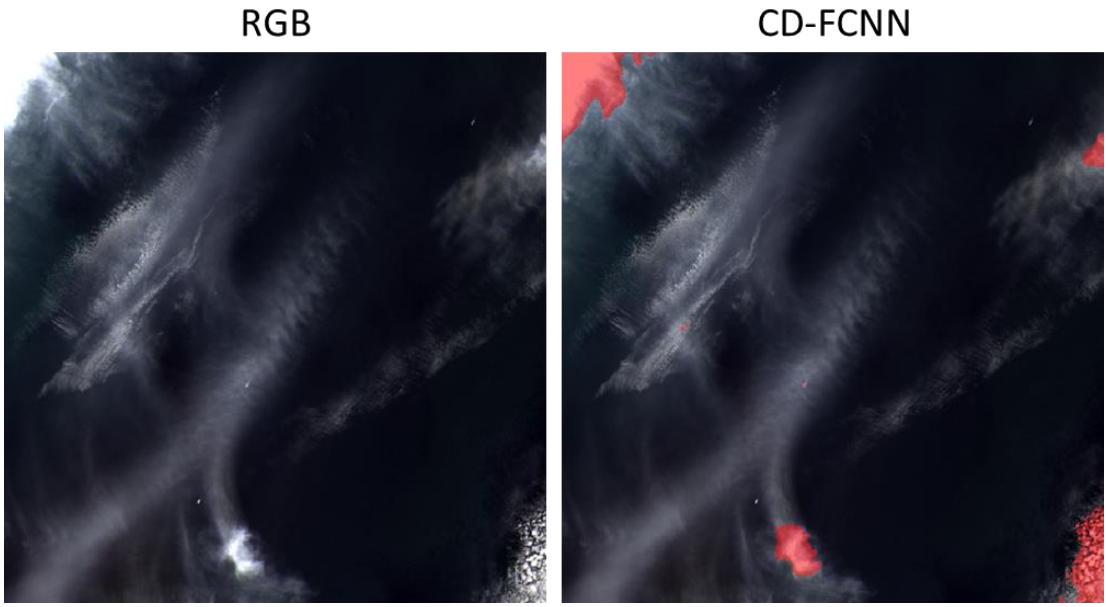
CD-FCNN:

Findings:

- biggest issue semi-transparent clouds
- weak performance over water
- cloud borders are mostly omitted

Recommendations:

- A buffer could improve the performance quite a bit



Fmask 4.0 CCA:

Findings:

- good detection of smaller clouds
- issue with mixed land/water pixels: coastal areas, along rivers

Recommendations:

- Nothing specific

FORCE

Findings:

- good detection of smaller clouds
- some commission error of bright surfaces (e.g., urban and beaches)
- good amount of commission error of non-cloud as cloud, due to large buffer

Recommendations:

- Potentially decreasing the buffer size a tiny bit to lower commission error.

RGB

FORCE

LaSRC

Findings:

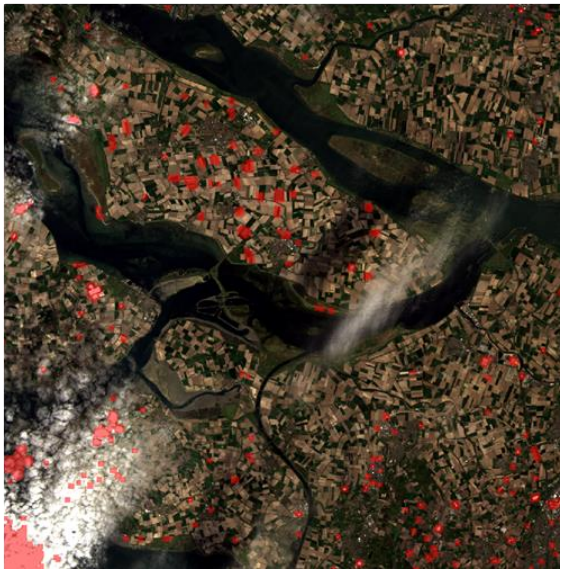
- weak performance in detecting semi-transparent clouds, especially over water.
- commissioning error of agricultural areas and urban

Recommendations:

- Add an additional snow flag

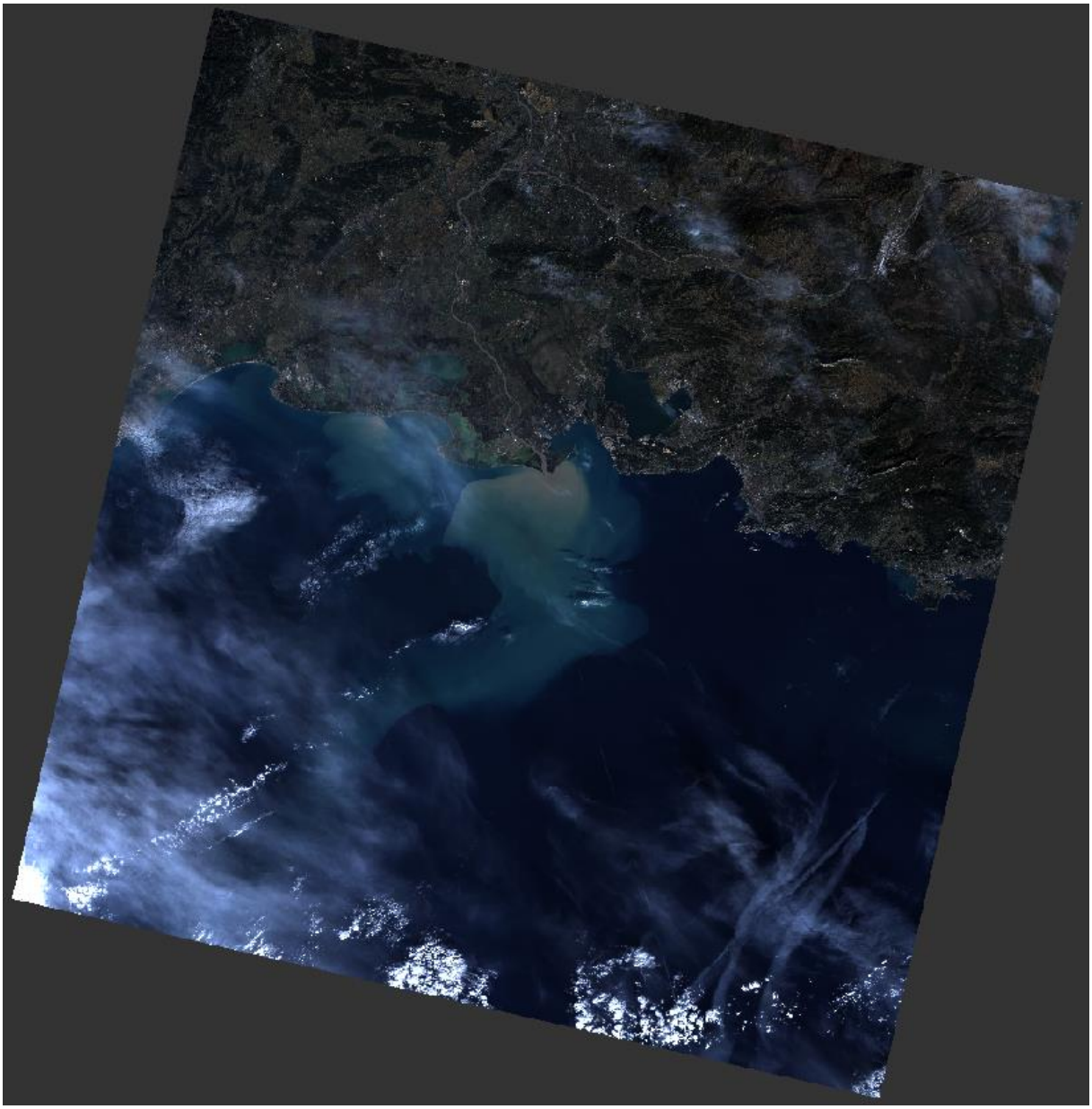
RGB

LaSRC



LC81960302014022

- Semi-transparent clouds over land and water
- Cumulus, stratocumulus, and altocumulus floccus clouds over water



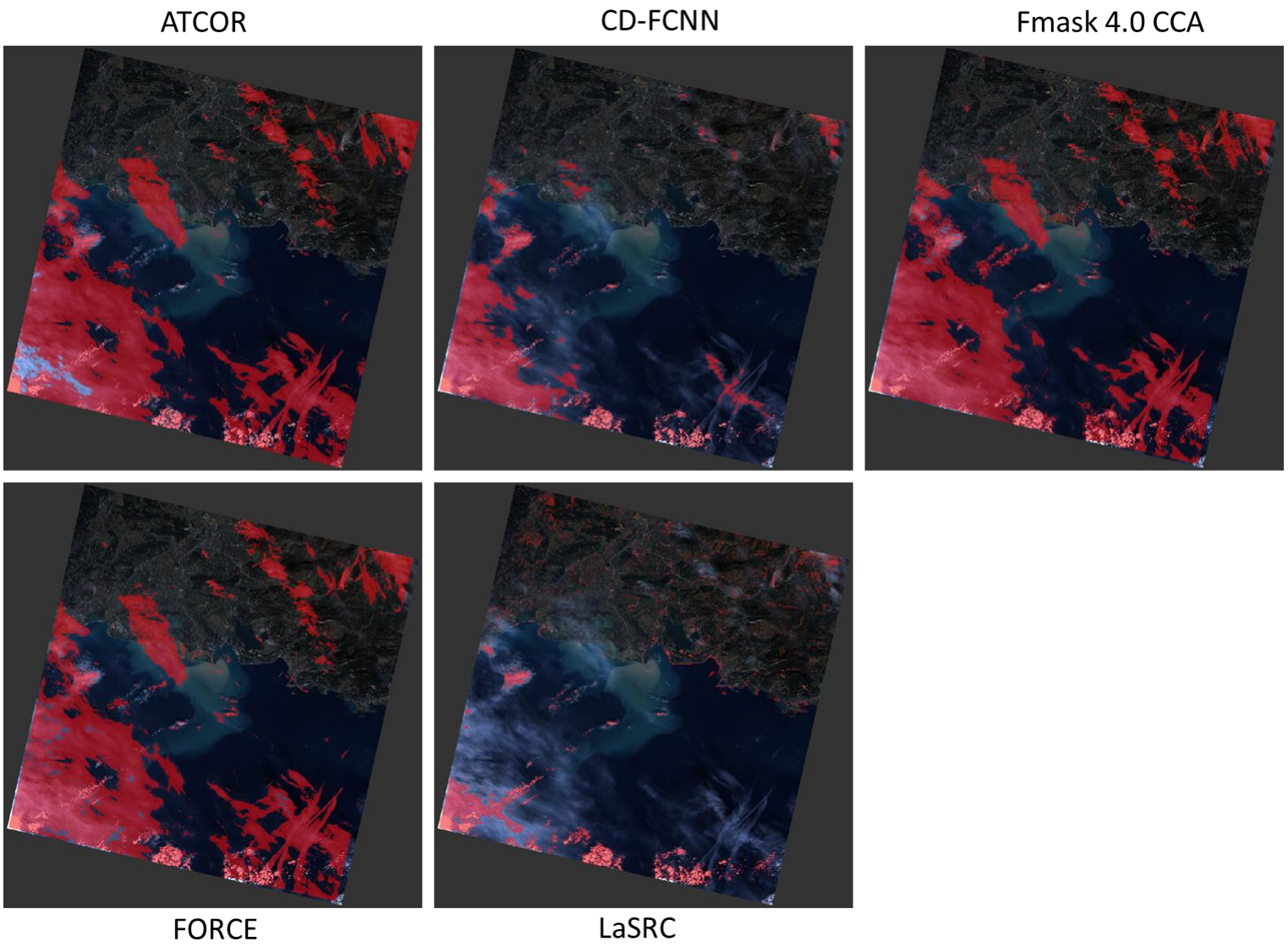
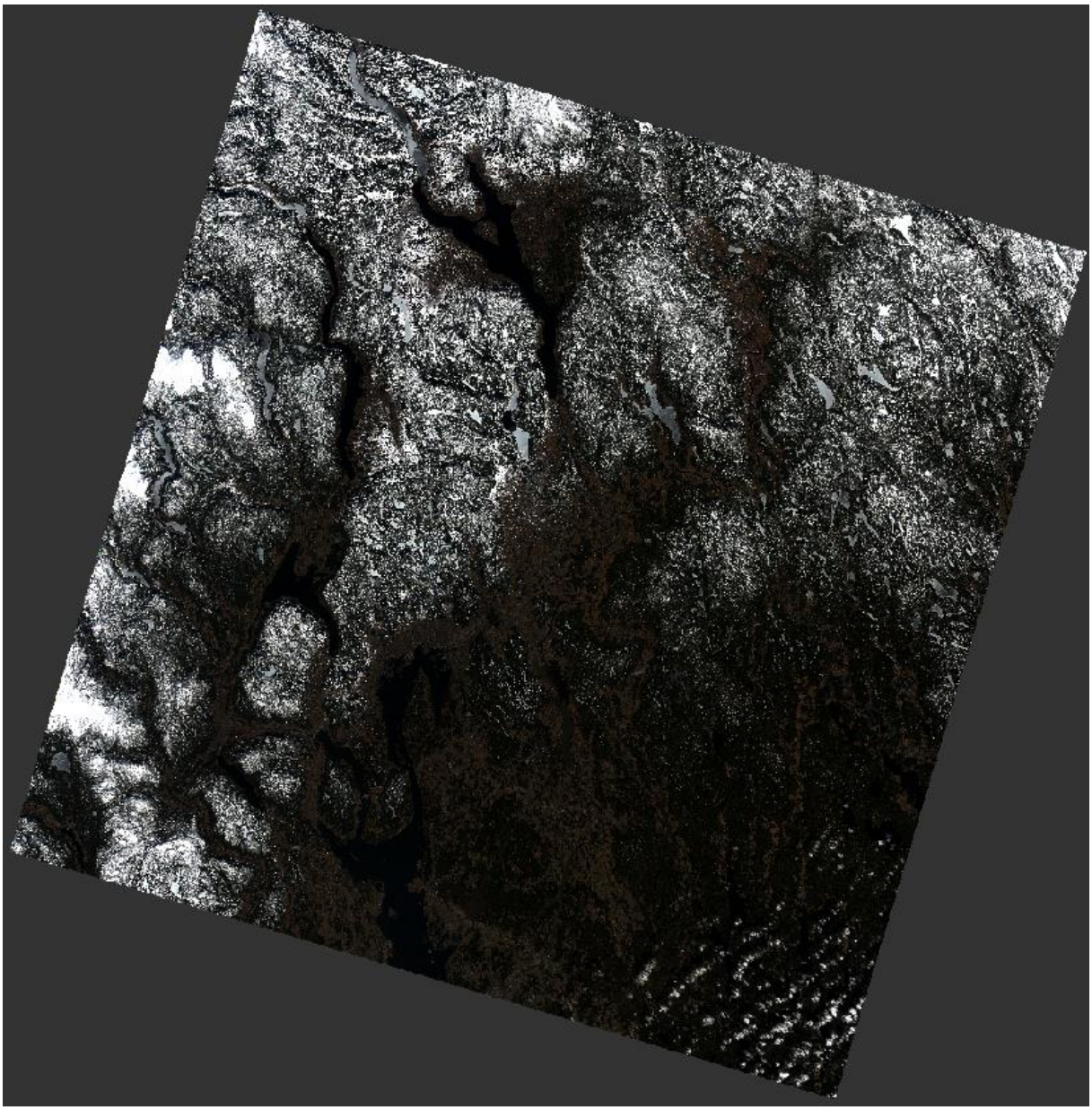


Figure 47: Algorithms' performance for LC81960302014022

LC81970182015080

- Non-cloud snow covered land surface
- Bands of cumulus clouds over snow free land surface



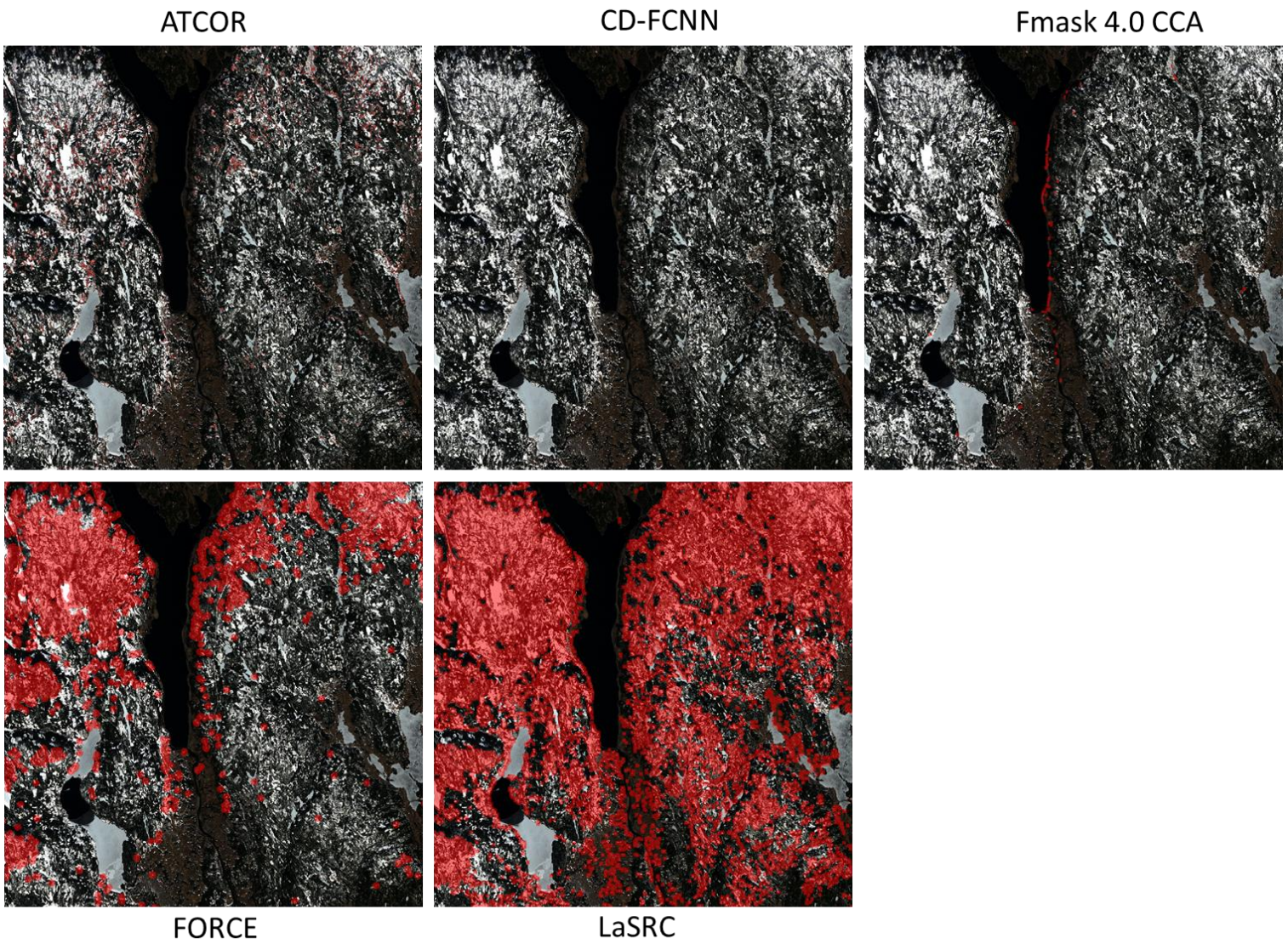


Figure 48: Algorithms' performance on LC81970182015080 over snow

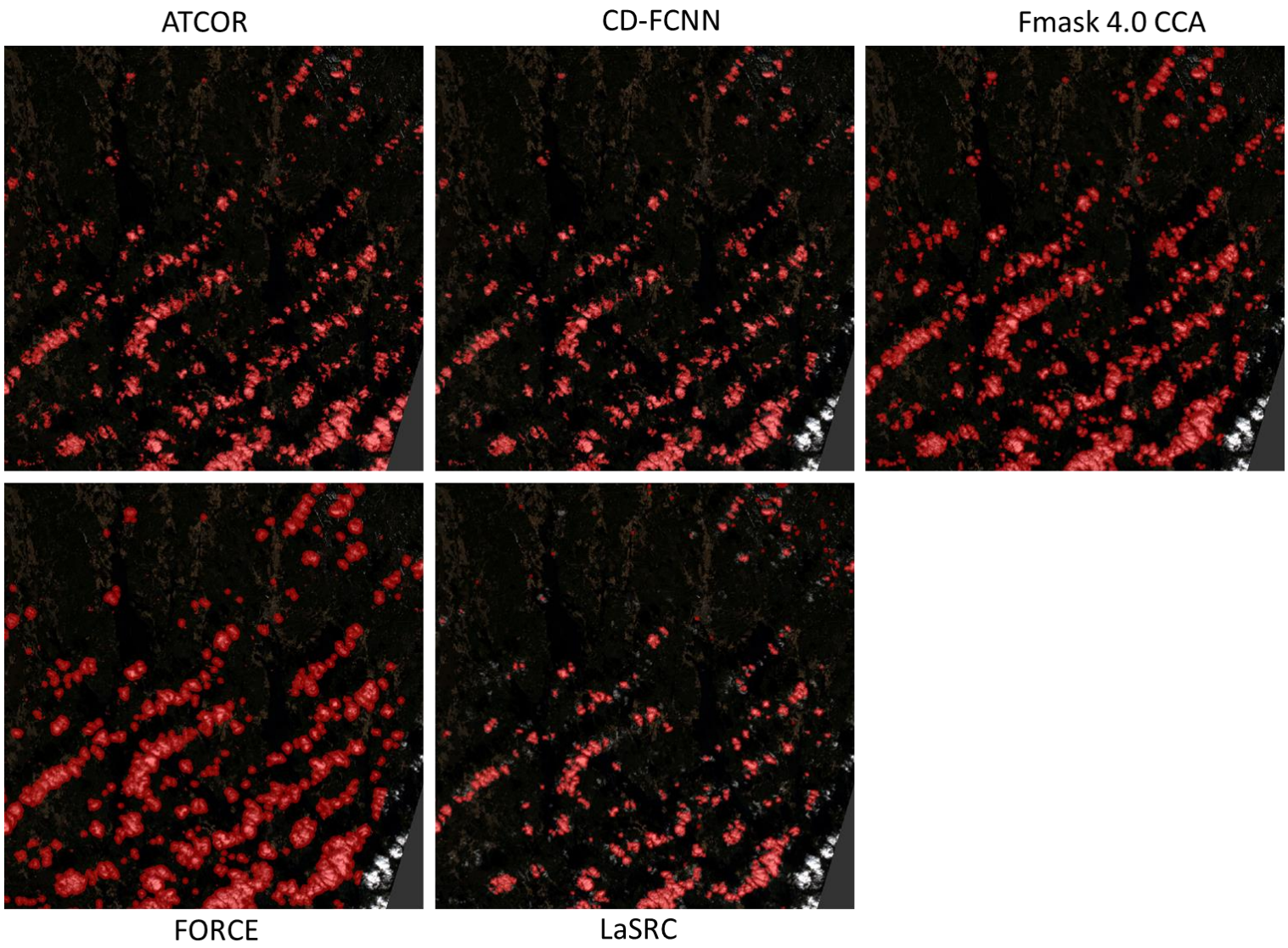
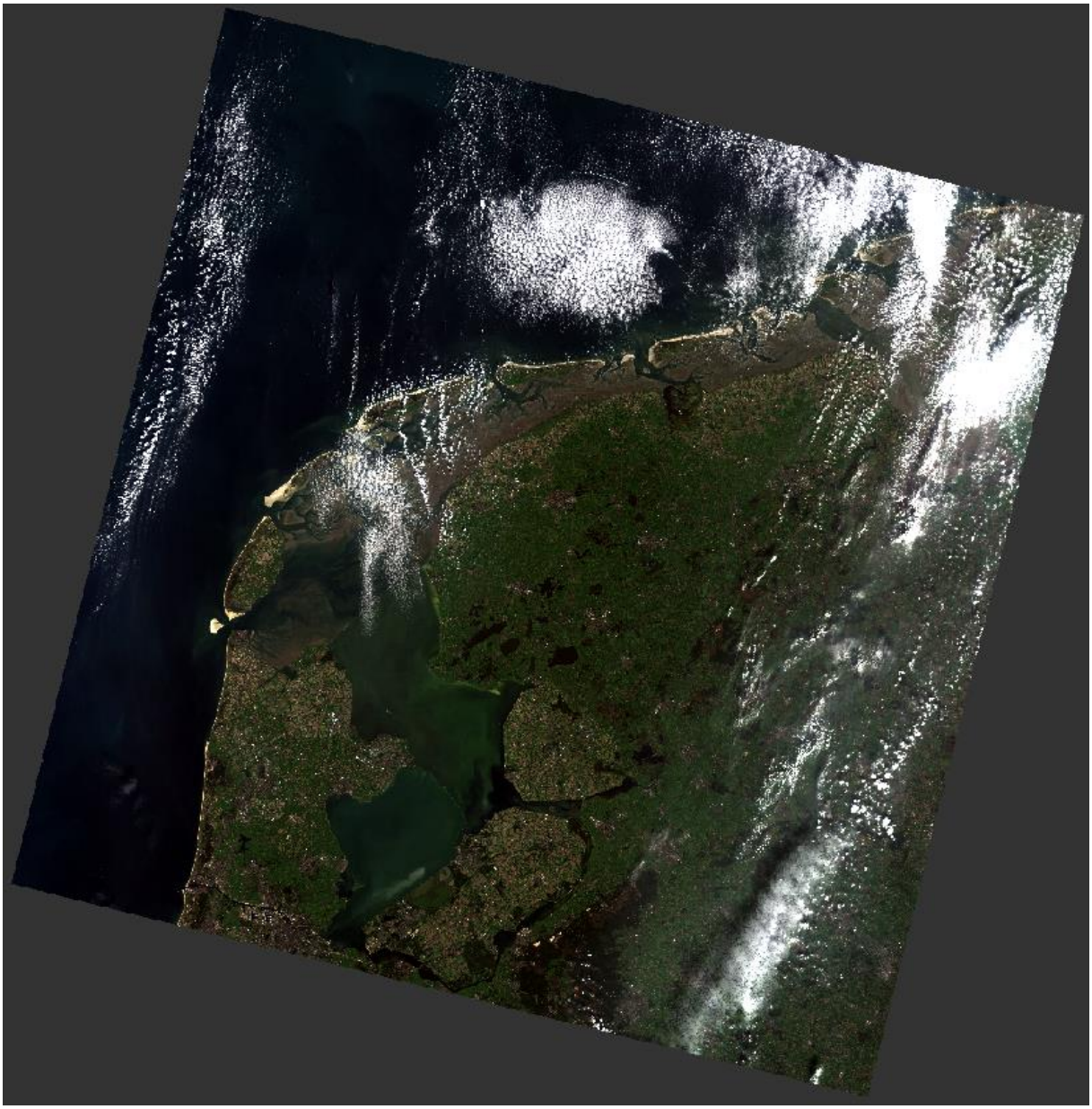


Figure 49: Algorithms' performance on LC81970182015080; No detection of clouds outside thermal band coverage except for ATCOR

LC81980232014276

- Cumulus, stratocumulus, and altocumulus floccus clouds over land and water
- Stratus and cirrus clouds over land



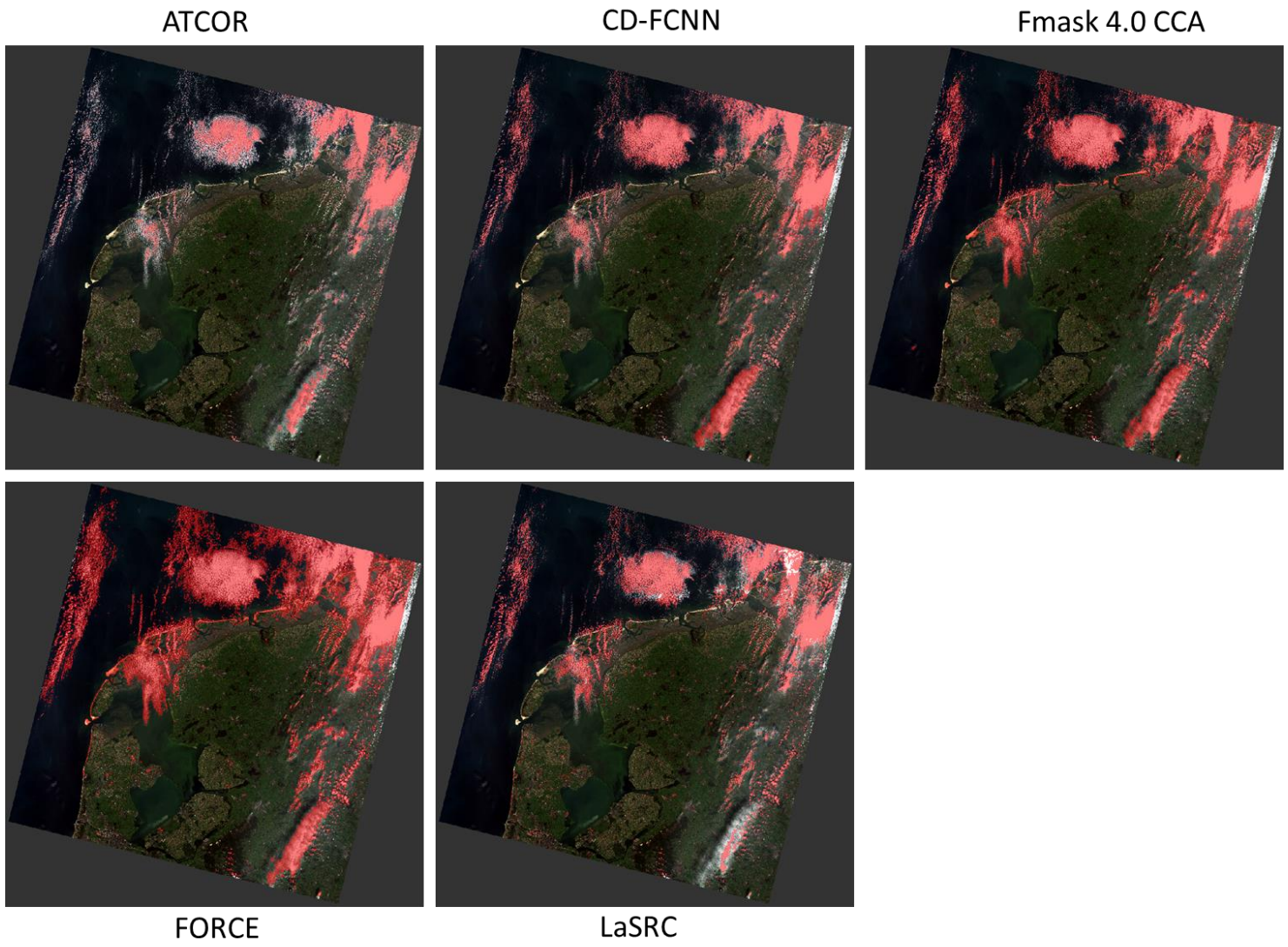


Figure 50: Algorithms' performance on LC81980232014276

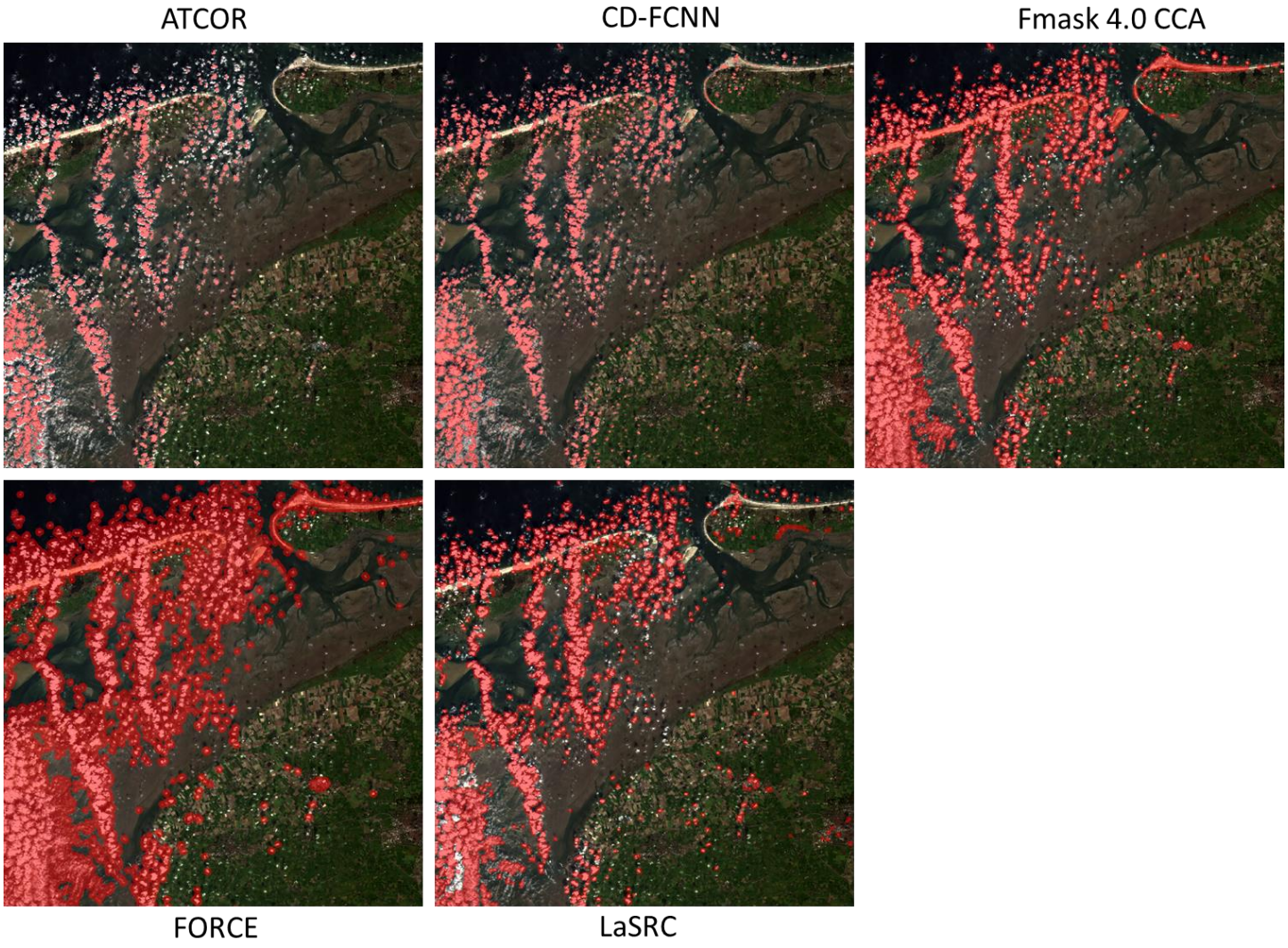
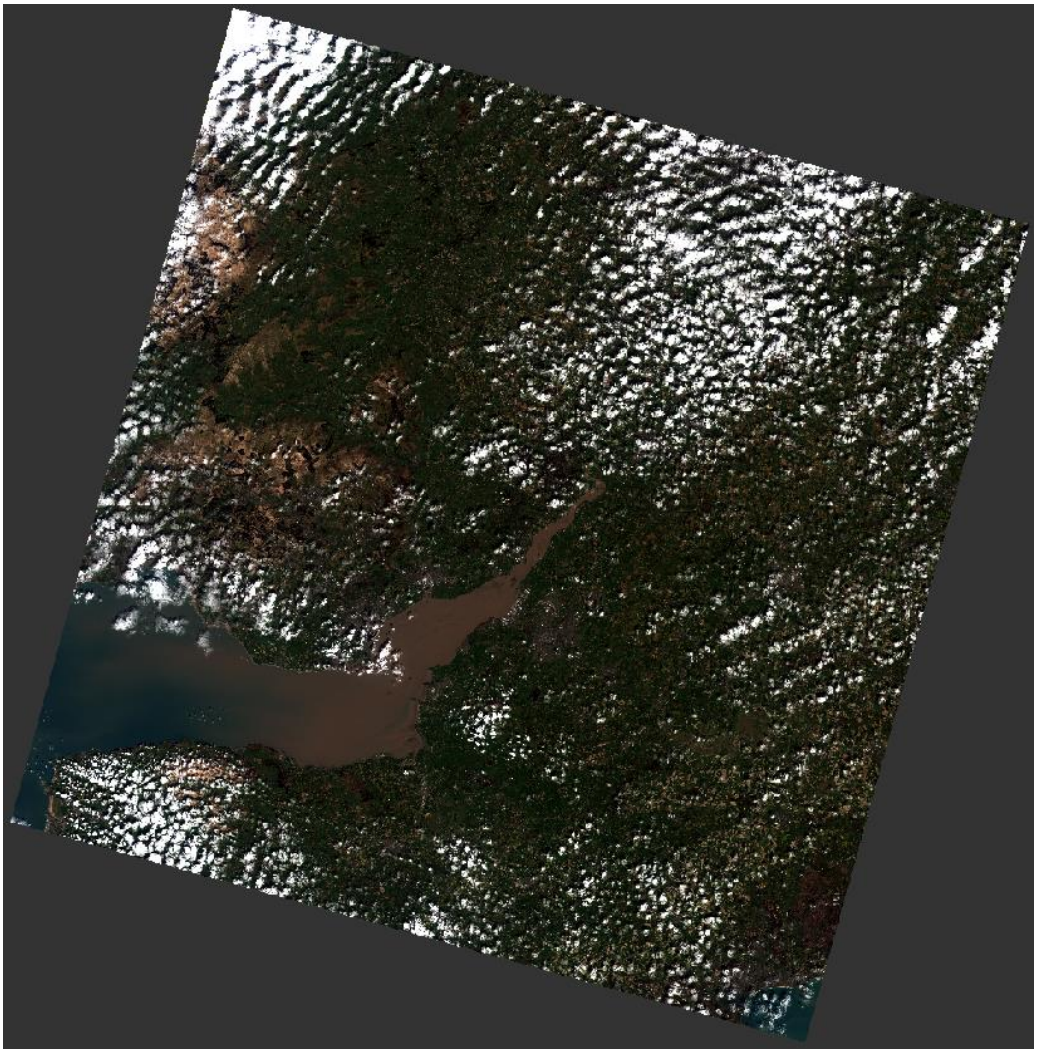


Figure 51: Algorithms' performance on LC81980232014276 (details)

LC82030242014103

- Cumulus, stratocumulus, altocumulus
- Mostly land with coastal water (high TSM)



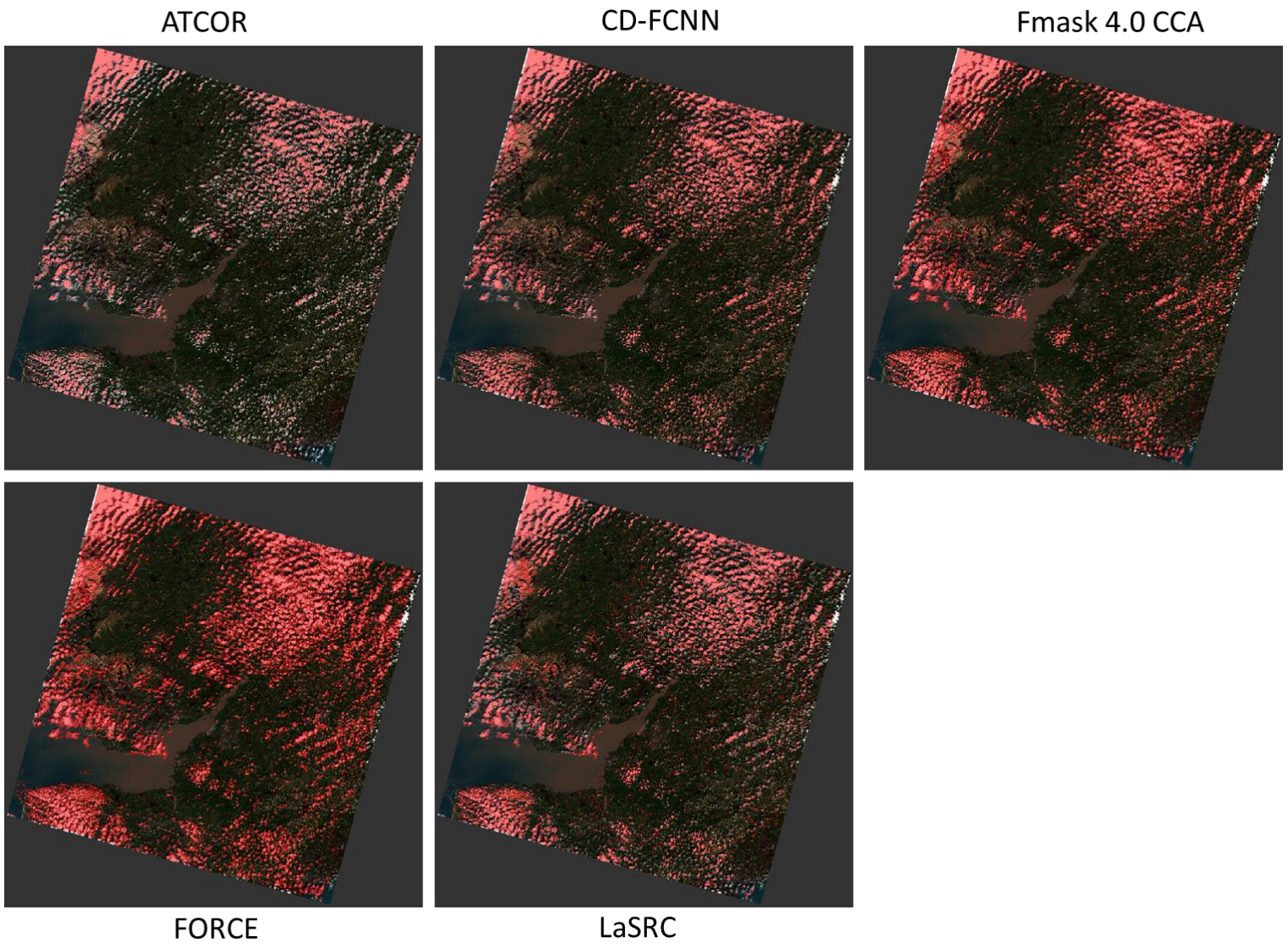
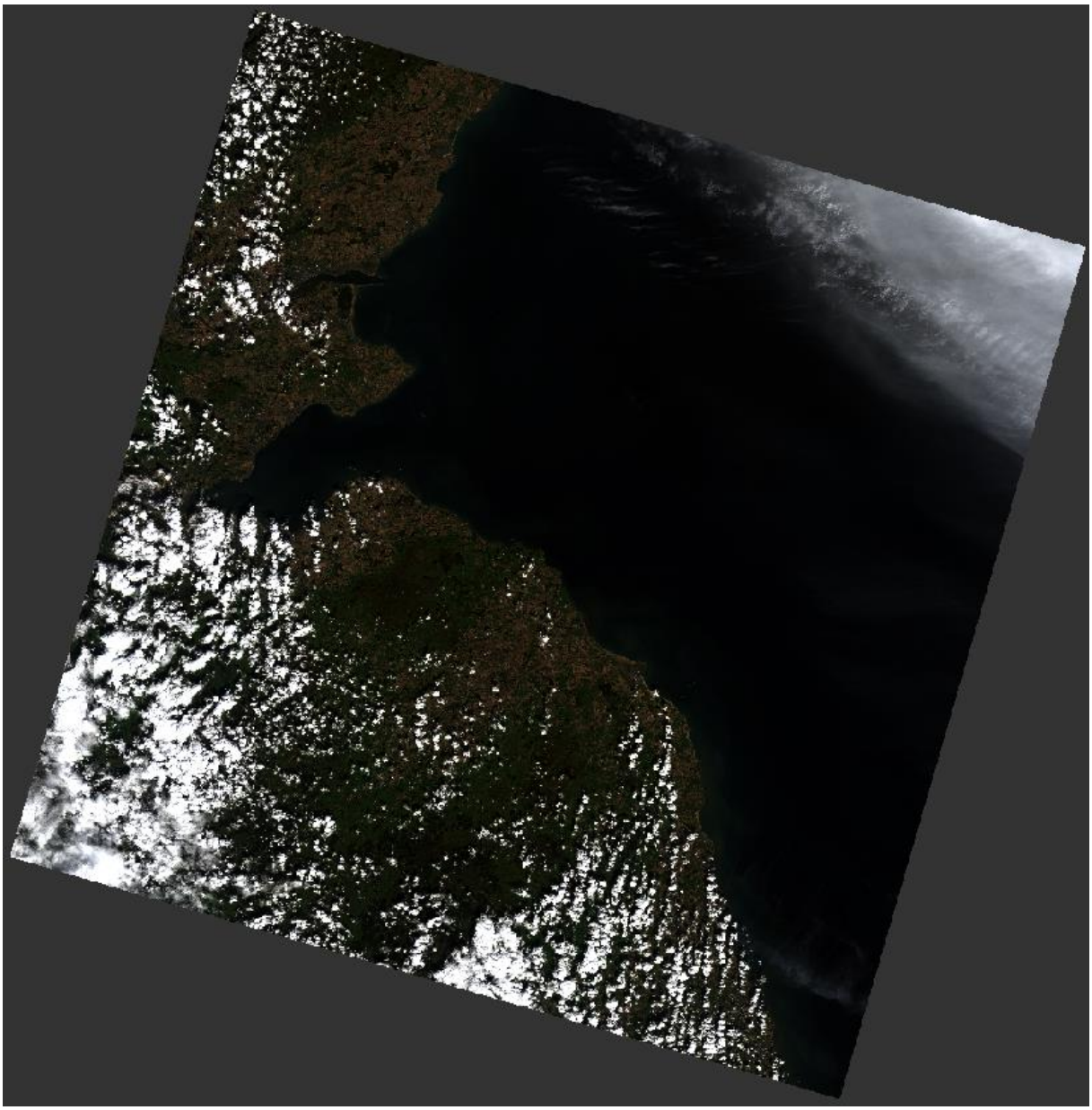


Figure 52: Algorithms' performance on LC82030242014103

LC82040212013251

- Stratus, and multiple types of cumulus over land
- Altostratus and cirrus clouds over water



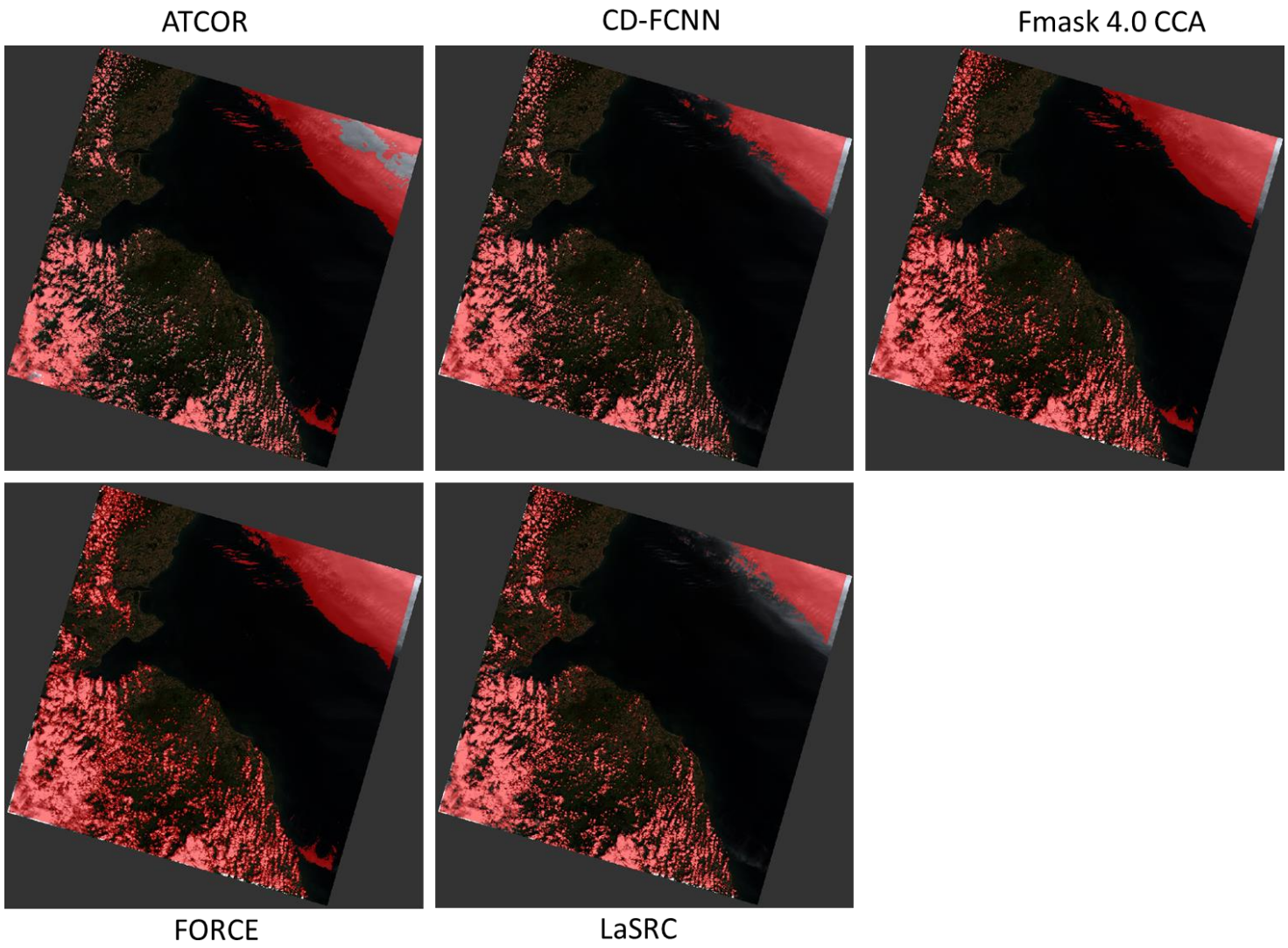


Figure 53: Algorithms' performance on LC82040212013251

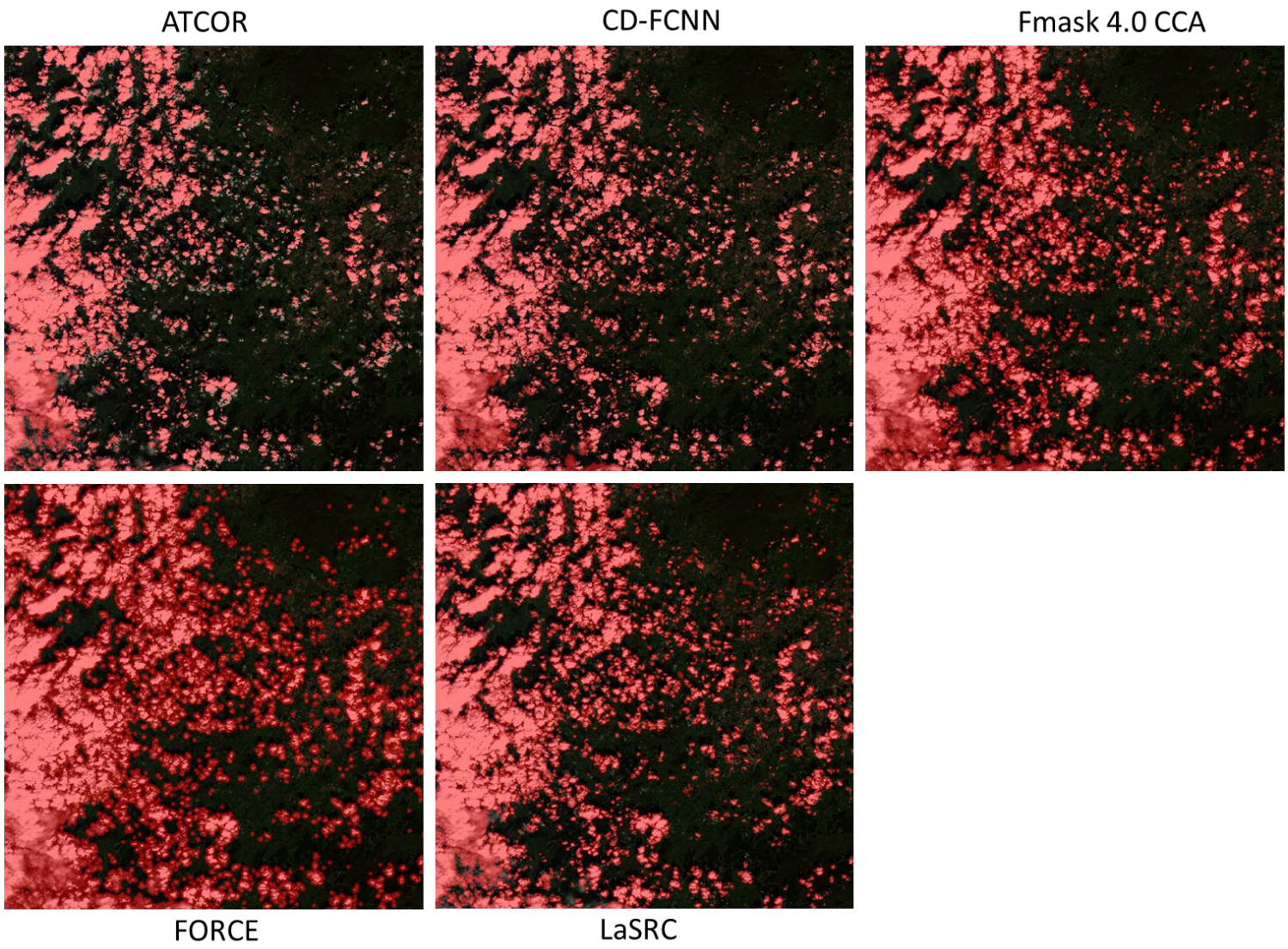


Figure 54: Algorithms' performance on LC82040212013251, detailed view

4 Feedback from the second CMIX workshop

In this section we will summarize the feedback we had collected during the second CMIX workshop in ESRIN.

The following 7 points are the main results:

1. All 5 validation datasets (VD) have different strengths and weaknesses -> results vary depending on the validation dataset
    a. Subjectivity of detecting/photo-interpreting clouds, especially thin clouds, in VD should be minimized, e.g. though the use of a network of sky images
2. CMIX had shown that there is no clear superiority of any methodology (Spectral tests vs. AI, mono vs. multitemporal)
3. A buffer and its size have a strong influence on the validation results -> Bigger buffer = better results for UA non cloud, but not for BOA.
4. All results show high accuracies (> 80%) for all processors -> cloud screening is good but can be improved.
5. It might be better to use 'UA clear' as a quality indicator compared to OA (Without neglecting the commission errors), but it also depends on the application. Also, in case of imbalanced classes, the balanced OA should be considered as well.
6. Thin semi-transparent clouds and cloud boundaries are an issue for mostly all algorithms. -> How to define a transparent cloud? boundary of a cloud?
7. The validation dataset and method does not allow for detecting systematic errors.

During discussion with the participants the following points had been discussed that should be considered by the algorithm developers or by the organization team for a next CMIX:

1. Usage of a buffer seems to be beneficial.
    - S2 bands look at surface under different angles, leading to uncertainty at the cloud edge.
    - Adjacency effect next to a cloud also impacts the signal.
    - Buffer included in cloud mask versus extra mask.
2. Cloud class definition
    - Cloud class should be uniquely defined, and consistent between algorithms' flag and validation data.
    - Cloud/no cloud too simplistic; more refined (at least slightly) should be used.
3. Validation data sets (VD)
    - VD critically influences the results.
    - Current VD were used because they were available, some having limitations; it would be better to have a dedicated VD for CMIX purpose.
    - We should compare the cloud masking in scenes available in different VDs.
    - Identify critical scenes in the VD (biggest differences in results between algorithms).
    - Dedicated VDs are needed for a CMIX II.

4. Cloud shadow and terrain shadow are important to consider, including the validation dataset.
5. Systematic errors should be identified, e.g. over bright surfaces[20]
6. Critical cases should get special attention: clouds over snow, urban, coastline, desert, …
7. It was suggested to linking CMIX with ACIX, to prove the „fitness for purpose“ of the cloud masking algorithms. This means the clouds masks would be used as input for an atmospheric correction.

[20] https://www.mdpi.com/2072-4292/10/10/1570

5 Consolidation of results

The goal of CMIX was the evaluation and intercomparison of multiple cloud screening algorithms. Before consolidating the results of the previous chapters, we need to step back a bit to reevaluate purpose, methods and datasets and see how they are fit to serve the purpose. To do this we need to answer a few questions:

1. Who is the addressee of the analysis?
2. What are the needs of the addressee?
3. Which indicators of the CMIX analysis are suitable to address these needs?
4. Are there more indicators that have not been considered that are important within real world application scenarios?

In the context of CMIX the first question cannot be answered clearly, as the results are important to users as well as to the producers. For the producers, the results give feedback on performance and potential sources for improvement. For the users, they can be a guidance for selecting a cloud mask. Nevertheless, the user's perspective should be the priority in any case, as the producers should aim at delivering the best cloud mask possible for the users. To evaluate this, we need to address the second question.

The needs of the users vary depending on the application. In terms of cloud masking there are three major preferences: cloud conservative, non-cloud conservative and balanced. Non-cloud conservative approaches are mostly needed for applications that do not allow cloud contamination (e.g., the remote sensing of land-, sea- and ice surface temperatures, determination of vegetation biophysical variables, phenology of vegetation status, elaboration of monthly composites of surface reflectance, determination of total column water vapor or of aerosol optical properties), while cloud conservative approaches are mostly needed for cloud remote sensing applications. Needs for balanced approaches are the hardest to serve. They are needed where each pixel of a product is important. This can only be achieved with a balanced overall good performance of a cloud mask, while the previous two needs can be achieved by accepting a certain amount of omission error.[21]

The main source of analysis for CMIX has been validation datasets which consist of manually or automatically classified samples used as the "truth" for comparison with the algorithms' results. For comparing the algorithms' results with the "truth" standard confusion matrices have been used. Confusion matrices provided different measures to quantify the performance of a classification. It gives performance indicators for both sides, the users as well as producers for each analyzed class, as well as information on the overall performance of an algorithm/classification. If we revisit the finding of the previous paragraph (needs) and consult the definitions of confusion matrix indicators (section 3.3.1), it becomes obvious that user accuracies are of great importance, as well as overall accuracies to analyze the balance of an algorithm. In addition to this method, single products have been analyzed visually, to get a better impression of the performance of the algorithms in the spatial domain. While these two methods are sufficient to answer the above described needs reasonably well, some questions cannot be answered, as: 1) are there any systematic detection issues of any algorithm?, 2) are detection issues bound to specific land cover types?, 3) do high cloud frequencies have influence

[21] As Sentinel-2 provides time series of surface reflectances with a good revisit, it might sometimes be preferable to wrongly discard a cloud free observation than to include a thin cloud in reflectance time series that might reduce the accuracy of the analyses and retrievals.

on temporal based cloud detection algorithms?, or 4) what is the actual effect of differences in cloud screening on real world applications?

When talking about real world applications, it becomes obvious that the current CMIX did not cover this scenario. The used comparison methods and indicators are limited to comparing the outputs (cloud masks) to evaluate the performance. While this is a fair method to start with, it does not cover different scenarios in which these algorithms can potentially be used. And thus, it gives no indication of costs for the users. You could think about four different potential user types with different "cost functions":

1. Desktop user working on single products
2. Desktop user within a cloud environment
3. Private processing cluster
4. Big data processing / service in the cloud (AWS, GE, ESA TEPs)

Depending on data needs (multi-temporal vs. mono-temporal approaches) and processing time[22], the cost for processing will vary for each user type.

Desktop user working locally.
While processing time might be not a big cost factor for a locally working user, if smaller amounts of products are processed, the need for long time series could be demanding, as this type of user usually downloads the products manually.

Desktop user within a cloud environment
Working in a cloud environment usually implies being close to your data, therefore bigger amounts of input data (time series) might be a smaller cost factor for this type of user, while processing time usually is, when CPU time is charged.

Private processing cluster
These types of clusters usually combine the issues of the first two examples. As usual a great amount of data is processed and data is stored locally, each additional download of later unused products generates costs, as well as smaller increases of processing time have a bigger impact, when processing big amounts of data.

Big data processing / service in the cloud (AWS, GE, ESA TEPs)
This fourth type of user is mostly affected by processing time and consequentially costs.

Besides these cost factors, the technical implementation and operations aspects of the algorithms has not been evaluated. For example:

1. How easy is the algorithm implemented in my environment?
2. Which operating systems are supported?
3. Is it GUI based, script based or both?
4. Does it provide an API to be integrated in other applications?
5. Is it free and open available or must be purchased?

[22] In this section to be understood as CPU time

Note:

Products from some algorithms are already available for end users to consume on various platforms, like MAJA cloud masks on https://www.theia-land.fr/en/product/sentinel-2-surface-reflectance/, sen2cor's and s2cloudless' cloud masks are available on Sentinel Hub (https://www.sentinel-hub.com/) and Google Earth Engine (https://earthengine.google.com/), as well as sen2cor being provided as ESA official L2A product via Copernicus open access hub (https://scihub.copernicus.eu/).

Nevertheless, all platforms do not provide cloud mask for the entire globe for the entire time span of L1C product availability.

With these four questions (audience, needs, indicators, real-world) addressed we can try to consolidate the finding from this first CMIX exercise.

- Pixel based analysis had given us:
    - a first insight into the functionality, strengths, and weaknesses of all compared algorithms,
    - an overview of the strengths and especially the weaknesses of the used validation datasets
    - A good basis to better design the next CMIX
- The visual analysis had shown that:
    - Specific algorithm behaviour cannot always be identified using statistical means.
    - Even if statistic results are good, issues of a method can be existing that are unfavourable for certain applications.

6 Conclusion and lessons learned

CMIX was a valuable lesson for the participants as well as for the organization team. It was the first exercise of its kind, and helped to better understand what is needed to compare and validate different cloud masks. Especially shortcomings and limitations in used reference datasets have been identified, while the exercise still revealed the differences in the compared cloud screening algorithms. Therefore, it helped to identify strengths and weaknesses of single algorithm/method.

A lot was learned from this first exercise, which will be used to prepare CMIX II. Therefore, recommendations for a next exercise are given in the following chapter.

7 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.

The first area of improvements should aim at providing, first of all, on a definition of “cloud” that is passed beforehand to participants and validation dataset originators. Ideally this would be an objective definition of clouds, which would include a numerical metric. As results from CMIX-I showed existing validation datasets varied in how a cloud was defined, and it influenced performance of the algorithms. One potential metric to define the cloud would be the cloud optical thickness, for example. However, this poses the questions at which wavelength to be defined. While there was a consensus between algorithms and developers in defining thick non-transparent clouds, there was a disagreement (sometimes by design and depending on the intended applications) in transparent (semi-transparent) clouds, such as cirrus and stratus, and cloud edges. Also, the effect of those clouds can vary with wavelengths, which adds complexity to the analysis.

Based on the cloud definition, the second area of improvements would include generation of new reference/validation datasets. The strengths and weaknesses of existing cloud reference datasets were thoroughly analyzed and discussed within this report, and new datasets should substantially address those weaknesses. Some of the recommendations include:

- Implementing consistently the cloud definition, and adding cloud shadows to the analysis. Recommended practices for labelling clouds shall be developed and implemented for new datasets, whether through visual interpretation or ground measurements or ancillary data (e.g. geostationary satellites). Clouds shadows should be also part of the analysis, since inaccurate cloud mask can lead to substantial artefacts in the downstream products.
- Increasing the number of sites of ground-based imagery of the sky and use them in coordination with Aeronet measurements.
- Acquire multiple datasets over the same area to analyse 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:

- Sample-based approach versus area-based approach, when comparing reference cloud mask with the predicted one. The problem with an area-based approach is that more weights would be given to large clouds (which cover the larger area), whereas smaller clouds might have a small impact on the performance metrics. But sampled based approaches can also miss some

specific land cover features, and sometimes do not address the limits of the clouds. Both approaches are therefore necessary.

- 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. Nevertheless, undetected clouds add noise on time series, therefore 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 analyse efficiency of the cloud/shadow masks is to "validate" them indirectly within the downstream products. An example can include a generic land cover mapping workflow, when the same set of satellite data will be processed by various cloud detection algorithms, and used an input to the classification algorithm. The derived land cover maps will be validated using the same validation data and inter-compared.

And finally, CMIX-I was limited to Landsat 8 and Sentinel-2 data. Future activities could include adding coarse resolution data, such as MODIS, VIIRS, Sentinel-3, and commercial very high spatial resolution satellites, such as Planet.

8 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 Sensing*, 11, 433.

Baetens Louis, & Hagolle Olivier. (2018). Sentinel-2 reference cloud masks generated by an active learning method [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1460961

Claverie, M., Ju, J., Masek, J. G., Dungan, J. L., Vermote, E. F., Roger, J. C., ... & Justice, C. (2018). The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote Sensing of Environment*, *219*, 145–161.

Congalton, R.G. 1991: A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*. 37: 35-46.

Congalton, R.G.; Green, K. 1999. Assessing the accuracy of remotely sensed data: principles and practices. Boca Raton, FL: Lewis Publishers. 137 p.

Congalton, R.G. 2007: Thematic and Positional Accuracy Assessment of Digital Remotely Sensed Data. In: McRoberts, Ronald E.; Reams, Gregory A.; Van Deusen, Paul C.; McWilliams, William H., eds. Proceedings of the seventh annual forest inventory and analysis symposium; October 3-6, 2005; Portland, ME. Gen. Tech. Rep. WO-77. Washington, DC: U.S. Department of Agriculture, Forest Service: 149-154.

Doxani, G., Vermote, E., Roger, J. C., Gascon, F., Adriaensen, S., Frantz, D., ... & Louis, J. (2018). Atmospheric correction inter-comparison exercise. *Remote Sensing*, *10*(2), 352.

Foga, S., Scaramuzza, P. L., Guo, S., Zhu, Z., Dilley Jr, R. D., Beckmann, T., ... & Laue, B. (2017). Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote Sensing of Environment*, *194*, 379-390.

Frantz, D. (2019). FORCE—Landsat + Sentinel-2 Analysis Ready Data and Beyond. *Remote Sensing*, 11, 1124. DOI: 10.3390/rs11091124

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 Sensing of Environment*, 215, 471-481. DOI: 10.1016/j.rse.2018.04.046

Frantz, D., Röder, A., Stellmes, M., & Hill, J. (2016). An Operational Radiometric Landsat Preprocessing Framework for Large-Area Time Series Applications. *IEEE Transactions on Geoscience and Remote Sensing*, 54, 3928-3943. DOI: 10.1109/TGRS.2016.2530856

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 Geoscience and Remote Sensing Letters*, 12, 1242-1246. DOI: 10.1109/lgrs.2015.2390673

Gascon, F., Bouzinac, C., Thépaut, O., Jung, M., Francesconi, B., Louis, J., ... & Languille, F. (2017). Copernicus Sentinel-2A calibration and products validation status. *Remote Sensing*, *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 Sensing of Environment*, 114(8), 1747-1755.

Hagolle Olivier, Huc Mireille, Desjardins Camille, Auer Stefan, & Richter Rudolf. (2017, December 7). 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., ... & Lavenu, F. (1998). AERONET—A federated instrument network and data archive for aerosol characterization. *Remote Sensing of Environment*, *66*(1), 1–16.

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 Sensing*, *8*(8), 666.

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. *Applied Optics*, *45*(26), 6762–6774.

López-Puigdollers, D., Mateo-García, G., & Gómez-Chova, L. (2021). Benchmarking Deep Learning Models for Cloud Detection in Landsat-8 and Sentinel-2 Images. *Remote Sensing*, *13*(5), 992.

Mateo-García, G., Laparra, V., López-Puigdollers, D., Gómez-Chova, L., (2020) Transferring deep learning models for cloud detection between Landsat-8 and Proba-V. ISPRS J. Photogramm. *Remote Sens.*, 160, 1–17.

Qiu, Shi, et al. "Improving Fmask cloud and cloud shadow detection in mountainous area for Landsats 4–8 images." *Remote Sensing of Environment* 199 (2017): 107-119.

Qiu, Shi, Zhe Zhu, and Binbin He. "Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery." *Remote Sensing of Environment* 231 (2019): 111205.

Richter, R., "A fast atmospheric correction algorithm applied to Landsat TM images", *Int. J. Remote Sensing*, Vol. 11, 159-166 (1990).

Richter, R., Wang. X., Bachman, M., and Schläpfer, D., "Correction of cirrus effects in Sentinel-2 type of imagery", *Int. J. Remote Sensing*, Vol. 32, 2931-2941 (2011).

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. *International Journal of Applied Earth Observation and Geoinformation*, *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 Geoscience and Remote Sensing Letters*, *14*(12), 2408–2412.

Skakun, S., Vermote, E. F., Roger, J. C., Justice, C. O., & Masek, J. G. (2019). Validation of the LaSRC cloud detection algorithm for Landsat 8 images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *12*(7), 2439–2446.

Vermote, E., Justice, C., Claverie, M., & Franch, B. (2016). Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. *Remote Sensing of Environment*, *185*, 46–56.

Vermote, E. F., Tanré, D., Deuze, J. L., Herman, M., & Morcette, J. J. (1997). Second simulation of the satellite signal in the solar spectrum, 6S: An overview. *IEEE Transactions on Geoscience and Remote Sensing*, *35*(3), 675–686.

Vermote, E. F., and S. Kotchenova (2008), Atmospheric correction for the monitoring of land surfaces, *Journal of Geophysical Research-Atmospheres*, *113*(D23).

Vermote, E., Justice, C., & Csiszar, I. (2014). Early evaluation of the VIIRS calibration, cloud mask and surface reflectance Earth data records. *Remote Sensing of Environment*, *148*, 134–145.

Zhu, Zhe, Shixiong Wang, and Curtis E. Woodcock. "Improvement and expansion of the Fmask algorithm: Cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images." *Remote Sensing of Environment* 159 (2015): 269-277

Zhu, Z., & Woodcock, C.E. (2012). Object-Based Cloud and Cloud Shadow Detection in Landsat Imagery. *Remote Sensing of Environment, 118*, 83-94. DOI: 10.1016/j.rse.2011.10.028

9 Annex

9.1 S2 Pixbox Detailed results

9.1.1 Complete dataset – no thin clouds

Figure 55 and Figure 56 show the confusion matrices for all algorithms on the complete dataset but excluding thin clouds.

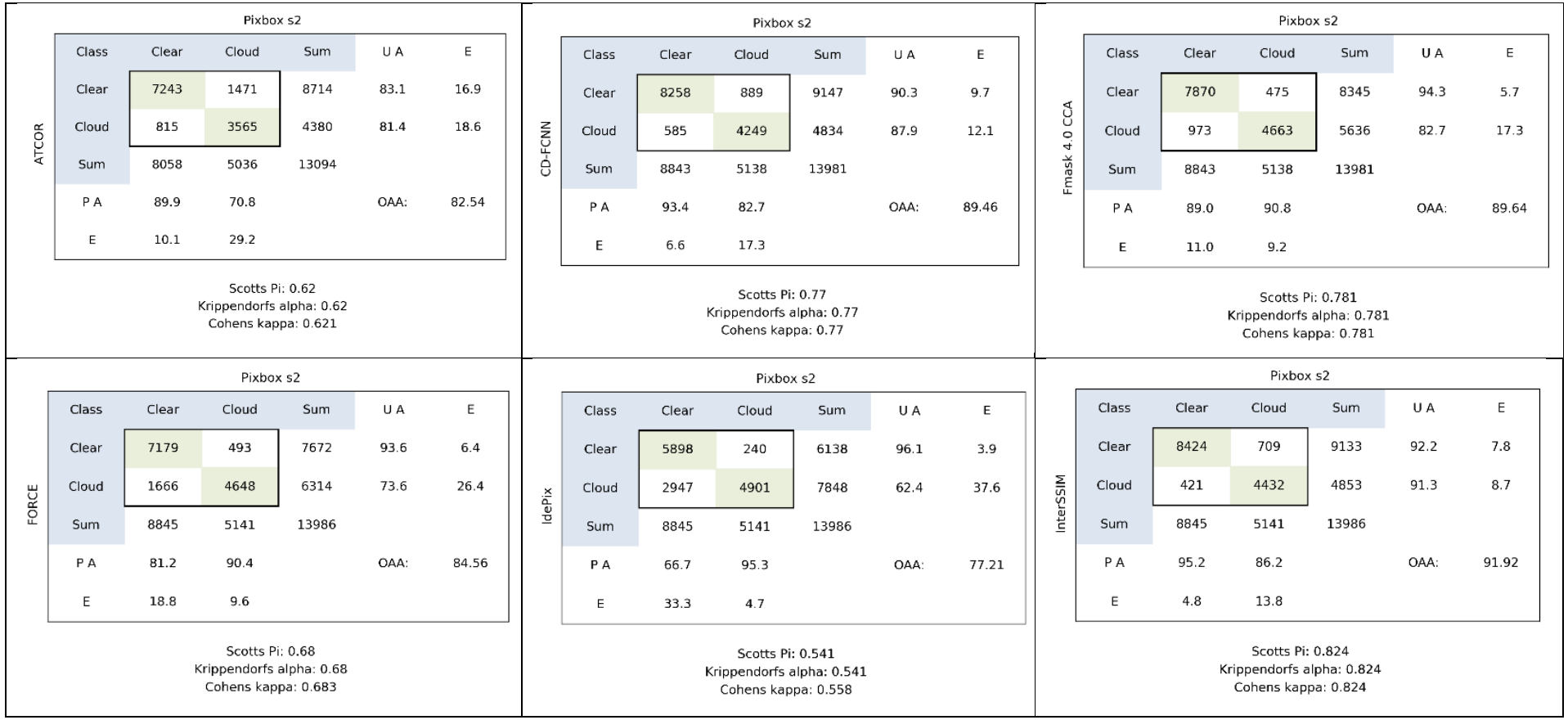


Figure 55: Confusion matrices for the complete dataset without thin clouds – part 1

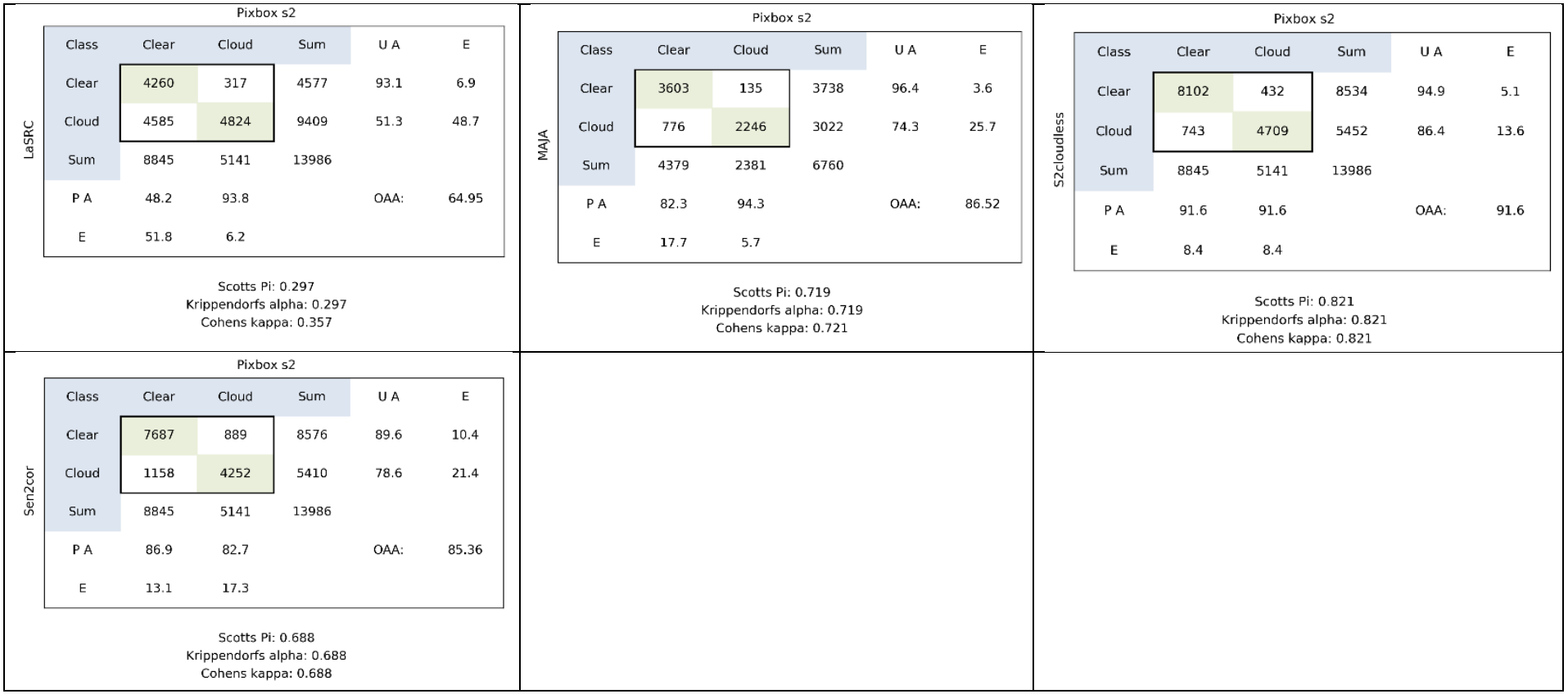


Figure 56: Confusion matrices for the complete dataset without thin clouds – part 2

9.1.2 Detailed view of classifications over different clear surfaces

ATCOR

Class	Urban	Desert, sandy soil	Dry/Salt lake	Other land	Road	Vessel, offshore platform	Spatially mixed land/snow-ice	Spatially mixed land/water	River	Ship canal	Lake	Coastal	Open Ocean	Snow/ice	Glint
Clear	483	208	234	2359	159	27	58	143	200	1	631	626	333	1616	165
Cloud	63	51	36	251	28	0	8	56	9	0	6	50	13	152	92
Sum	546	259	270	2610	187	27	66	199	209	1	637	676	346	1768	257

CD-FCNN

Class	Urban	Desert, sandy soil	Dry/Salt lake	Other land	Road	Vessel, offshore platform	Spatially mixed land/snow-ice	Spatially mixed land/water	River	Ship canal	Lake	Coastal	Open Ocean	Snow/ice	Glint
Clear	491	165	152	2550	173	27	160	195	202	1	640	688	341	2216	257
Cloud	55	94	118	146	14	0	0	4	7	0	2	8	5	132	0
Sum	546	259	270	2696	187	27	160	199	209	1	642	696	346	2348	257

Fmask 4.0 CCA

Class	Urban	Desert, sandy soil	Dry/Salt lake	Other land	Road	Vessel, offshore platform	Spatially mixed land/snow-ice	Spatially mixed land/water	River	Ship canal	Lake	Coastal	Open Ocean	Snow/ice	Glint
Clear	473	238	153	2587	174	27	158	143	195	1	640	646	330	1934	171
Cloud	73	21	117	109	13	0	2	56	14	0	2	50	16	414	86
Sum	546	259	270	2696	187	27	160	199	209	1	642	696	346	2348	257

FORCE

Class	Urban	Desert, sandy soil	Dry/Salt lake	Other land	Road	Vessel, offshore platform	Spatially mixed land/snow-ice	Spatially mixed land/water	River	Ship canal	Lake	Coastal	Open Ocean	Snow/ice	Glint
Clear	412	226	166	2354	159	25	144	136	191	1	621	623	313	1693	115
Cloud	134	33	104	343	28	2	16	63	18	0	21	73	33	656	142
Sum	546	259	270	2697	187	27	160	199	209	1	642	696	346	2349	257

IdePix

Class	Urban	Desert, sandy soil	Dry/Salt lake	Other land	Road	Vessel, offshore platform	Spatially mixed land/snow-ice	Spatially mixed land/water	River	Ship canal	Lake	Coastal	Open Ocean	Snow/ice	Glint
Clear	420	150	64	2479	158	18	150	129	189	1	618	587	326	473	136
Cloud	126	109	206	218	29	9	10	70	20	0	24	109	20	1876	121
Sum	546	259	270	2697	187	27	160	199	209	1	642	696	346	2349	257

Figure 57: Detailed view of clear in-situ classes classified as cloud or clear by the algorithms – part 1

InterSSIM

Class	Urban	Desert, sandy soil	Dry/Salt lake	Other land	Road	Vessel, offshore platform	Spatially mixed land/snow-ice	Spatially mixed land/water	River	Ship canal	Lake	Coastal	Open Ocean	Snow/ice	Glint
Clear	482	246	230	2637	187	27	159	153	199	1	642	644	330	2232	255
Cloud	64	13	40	60	0	0	1	46	10	0	0	52	16	117	2
Sum	546	259	270	2697	187	27	160	199	209	1	642	696	346	2349	257

LaSRC

Class	Urban	Desert, sandy soil	Dry/Salt lake	Other land	Road	Vessel, offshore platform	Spatially mixed land/snow-ice	Spatially mixed land/water	River	Ship canal	Lake	Coastal	Open Ocean	Snow/ice	Glint
Clear	367	88	55	1653	126	27	58	108	156	0	549	591	311	153	18
Cloud	179	171	215	1044	61	0	102	91	53	1	93	105	35	2196	239
Sum	546	259	270	2697	187	27	160	199	209	1	642	696	346	2349	257

MAJA

Class	Urban	Desert, sandy soil	Dry/Salt lake	Other land	Road	Vessel, offshore platform	Spatially mixed land/snow-ice	Spatially mixed land/water	River	Ship canal	Lake	Coastal	Open Ocean	Snow/ice	Glint
Clear	260	114	124	1159	78	4	26	15	102	1	291	208	261	960	0
Cloud	128	19	15	240	21	0	0	0	12	0	66	6	47	222	0
Sum	388	133	139	1399	99	4	26	15	114	1	357	214	308	1182	0

S2cloudless

Class	Urban	Desert, sandy soil	Dry/Salt lake	Other land	Road	Vessel, offshore platform	Spatially mixed land/snow-ice	Spatially mixed land/water	River	Ship canal	Lake	Coastal	Open Ocean	Snow/ice	Glint
Clear	439	224	185	2558	181	27	158	142	197	1	641	644	329	2125	251
Cloud	107	35	85	139	6	0	2	57	12	0	1	52	17	224	6
Sum	546	259	270	2697	187	27	160	199	209	1	642	696	346	2349	257

Sen2cor

Class	Urban	Desert, sandy soil	Dry/Salt lake	Other land	Road	Vessel, offshore platform	Spatially mixed land/snow-ice	Spatially mixed land/water	River	Ship canal	Lake	Coastal	Open Ocean	Snow/ice	Glint
Clear	448	187	150	2486	165	14	150	137	186	1	634	640	325	2072	92
Cloud	98	72	120	211	22	13	10	62	23	0	8	56	21	277	165
Sum	546	259	270	2697	187	27	160	199	209	1	642	696	346	2349	257

Figure 58: Detailed view of clear in-situ classes classified as cloud or clear by the algorithms – part 2