Comprehensive quality assessment of optical satellite imagery using weakly supervised video learning

Valerie J. Pasquarella, Christopher F. Brown, Wanda Czerwinski, William J. Rucklidge
Google LLC
{valpasq, cfb, wczerwinski, wjr}@google.com

Abstract

Identifying high-quality (i.e., relatively clear) measurements of surface conditions is a near-universal first step in working with optical satellite imagery. Many cloud masking algorithms have been developed to characterize the likelihood that reflectance measurements for individual pixels were influenced by clouds, cloud shadows, and other atmospheric effects. However, due to the continuous density of the atmospheric volume, we argue that quantification of occlusion and corruption effects is better treated as a regression problem rather than a discretized classification problem as done in prior work. We propose a space-time context network trained using a bootstrapping procedure that leverages millions of automatically-mined video sequences informed by a weakly supervised measure of atmospheric similarity. We find that our approach outperforms existing machine learning and physical basis cloud and cloud shadow detection algorithms, producing state-of-the-art results for Sentinel-2 imagery on two different out-of-distribution reference datasets. The resulting product offers a flexible quality assessment (QA) solution that enables both standard cloud and cloud shadow masking via thresholding and more complex image grading for compositing or downstream models. By way of generality, minimal supervision, and scale of our training data, our approach has the potential to significantly improve the utility and usability of optical remote sensing imagery.

1. Introduction

With the increasing prevalence of cloud-based storage and analysis of Earth observation imagery [1, 16, 33, 36], reliable pre-processing algorithms are more critical than ever [56, 58]. Sensor-specific approaches for masking and correcting atmospheric effects in optical remote sensing imagery have been integrated into processing pipelines for major image providers such as the USGS and Copernicus/ESA [12, 13, 28, 37]. However, accurate and globally consistent cross-sensor image quality assessment remains an ongoing challenge [34, 45, 46, 64].

Most operational cloud masking efforts, including the Landsat TM/ETM+ Automatic Cloud Cover Assessment (ACCA) [21], Function of Mask (Fmask) algorithm [64, 65], Sentinel-2 QA60 bitmask [7], and Sen2Cor Scene Classification Layer (SCL) [27, 30] rely on a series of spectral tests and other heuristics to identify clouds and cloud shadows. Such approaches are defensible but rigid. Given variability in both atmospheric effects and global land surface types, it is exceedingly difficult to develop a universal set of rules that work for all cases. This makes so-called physical basis approaches prone to failure on corner cases such as mountainous regions [39, 51]; high-return surfaces including snow/ice, buildings, beaches, and salt flats that have similar spectral signatures to clouds [14, 22, 23, 66]; dark targets like wetlands and small water bodies that have similar reflectance properties as cloud shadows [18]; and high thin cirrus clouds that only partially occlude surface properties [41]. Furthermore, assumptions about physical relationships that would be expected from multi-spectral measurements are inherently linked to available spectral bands and band passes and must be adapted to new sensors, which has proven especially challenging for sensors lacking thermal and/or short-wave infrared bands [44, 62, 64].

Although physical basis approaches may be advantageous in their interpretability, we argue that a more general solution requires reframing atmospheric artifact detection as a learning problem. Given successes in other domains, deep learning approaches are increasingly being applied to Earth observation imagery for a variety of tasks, including generating better cloud masks [5, 24, 29, 63]. Most efforts to develop deep learning models for cloud masking rely on discrete classification via fully convolutional network (FCN) architectures [9, 10, 25, 59]. Unlike traditional physically-motivated expert systems, FCNs assign per-pixel labels based on implicit, often complex relationships across space and spectra [26]. However, moving away from physical basis approaches means the quality of results becomes inextricably tied to the quality and diversity of reference annotations in the training set, and collecting densely annotated image examples at scale is a massive undertaking [2, 3]. Furthermore, human perception of discretized atmospheric condition is highly subjective [34], and given the continuous nature of atmospheric artifacts, what constitutes useful data for one use case may be unacceptable in another. Without memorization of the underlying landscape, there is also inherent ambiguity in identifying cloud shadows using pure spatial context as is the case in FCNs: simple examples include a cloud shadow extending from a swath boundary that cannot be disambiguated from a topographic shadow, and large cirrus shadows that uniformly darken the observed area. Thus, while modern machine-learning-based cloud masking solutions have been operationalized [20, 50, 67], these approaches trade heuristics for an extensive active learning feedback loop to collect a sufficiently large training dataset.

In this paper, we approach the challenge of operational machine-learning-based optical image quality assessment (QA) via weakly supervised learning. Rather than work with individual images or time series of observations for a single-pixel location, we assemble sequences of Sentinel-2 imagery into short video clips to leverage spatial and temporal context for identifying atmospheric and other image artifacts. Our contributions can be summarized as follows:

- Propose a bootstrapping method using a weakly supervised atmospheric similarity metric to generate a globally-distributed training dataset
- Introduce a space-time context-based QA model trained on millions of video clips produced using the aforementioned bootstrapping method and capable of performing single-image reference-free QA assessments
- Demonstrate state-of-the-art performance of the resulting QA product, Cloud Score+ (CS+), on two independently collected cloud and cloud shadow reference datasets

2. Related work

Multi-temporal cloud masking. Previous work has shown that multi-temporal observations can be used to improve cloud and cloud shadow detection and classification [6, 31, 38, 54]. The key assumption of these approaches is that a clear reference image, image composite, or fitted trajectory can serve as statistical "ground truth" for anomaly detection where unlikely deviations from the baseline are ascribed to atmospheric artifacts, e.g., [17, 66]. Though intuitive, it is often the case that the true counterfactual, i.e., the image uncontaminated by the atmosphere, is unknowable due to unseen changes in ground condition or noise introduced by the satellite image acquisition process. While theoretically promising, cloud and shadow detection based on statistical comparisons has proven challenging in practice.

Multi-sensor fusion and image in-painting. Combining imagery from optical sensors with cloud-penetrating microwave observations from Synthetic Aperture Radar (SAR) instruments to generate cloud-free images has become an increasingly active area of research [43, 57]. The goal of SAR-to-optical translation is to use SAR measurements to reconstruct or in-paint cloud-contaminated portions of a target optical image. This is a fundamentally different task than identifying clouds and cloud shadows, and in many cases, these approaches are dependent on cloud masks as an input [15, 32, 61].

Continuous quality assessment metrics. The conclusions of the first Cloud Masking Inter-comparison eXercise (CMIX-I), an international effort to compare the results of state-of-the-art cloud masking approaches for moderate resolution optical satellite imagery, suggested that vague definitions of clouds (including semi-transparent clouds and cloud boundaries) are generally problematic for most algorithms trained to identify discrete classes [46]. The workshop findings also noted that cloud shadow and terrain shadow are important to consider (including in validation datasets) and systematic errors (such as those over bright targets) should be identified. The s2cloudless algorithm [67] is a supervised single-date cloud detection approach that addresses issues with discrete cloud masks by predicting a per-pixel cloud probability score. There has also been work on post-processing techniques to reduce systematic errors, i.e., [14]. However, there is still no single product or metric that captures the full spectrum of atmospheric effects in a continuous manner and generalizes across geographies and sensors.

3. Methods

Our approach to building a space-time context-based QA model takes place over four distinct stages of model development, where at each stage the task becomes less abstracted from the ultimate goal: grading the quality, i.e., clarity, of a given observation relative to a theoretical clear reference image (Sec. 3.1). We first define a notion of "atmospheric similarity" (Sec. 3.2) and train a feature extractor to determine if two images are of the same location as a pre-training task (Sec. 3.4.1). We then fine-tune this feature extractor on a very small set of sparse annotations and synthetic images with known corruption values to establish relative values of QA scores (Sec. 3.4.2). We use the fine-tuned atmospheric similarity model (ASIM) to identify image sequences including a mix of relatively clear and cloudy frames through space and time (Sec. 3.5). Finally, we train a space-time context model that produces per-pixel reference-free QA scores for target images (Sec. 3.6). Although our models were developed and assessed using Sentinel-2 imagery (Sec. 3.3), the framework can be applied to other optical sensors with minimal-to-no modification to training and/or post-processing.

3.1. Quality Assessment (QA)

We define QA scores on [0, 1] and model each pixel as the linear combination of its true reflectance and some atmospheric corruption. This model assumes that at each pixel, (1) there exists a perfectly diffuse surface that may be shadowed and obscured by clouds and the atmosphere (corrupted), and (2) that the pixel imaged by the sensor (p_m) is a linear combination of this corrupted surface (c) and the "true" reflectance (p_t) at that location. Since there is some minimum optical thickness present in any imaging through Earth's atmosphere due to the presence of constituents like water vapor [19, 52], we consider "true" reflectance to be that measured from top-of-atmosphere in minimal optical depth conditions. The QA score (q) for a pixel (p) is then the coefficient on the true reflectance term under this simple model:

$$\mathrm{QA}(p_m) = q, \quad p_m = p_t q + c(1-q) \tag{1}$$

An exploitable side effect is that taking a linear combination of pixels with known QA scores produces a pixel with a QA score that is a linear combination of the individual QA scores of the combined pixels (see Sec. 3.4.2).

3.2. Atmospheric similarity

We hypothesize that it is a far simpler problem of perception to estimate "atmospheric similarity" (the pairwise maximum corruption between two images) than QA (the instantaneous corruption given a single image) directly. This is somewhat intuitive in that it can be difficult to discern clouds and cloud shadows in a single image. However, when observing a time series of acquisitions or a clear image at the same location, humans can easily perceive clouds and cloud shadows even without large spatial contexts.

We assume that cloud and cloud shadow conditions for a given location will almost always "look different" between acquisitions. It then follows that close-in-time satellite acquisitions of the same location will "look similar" given clear atmospheric conditions. We use this assumption to define a notion of atmospheric similarity between two images $\mathrm{ASIM}(\mathbf{x}, \mathbf{y})$ where for a given pixel i:

$$\mathrm{ASIM}_i(x_i, y_i) := \min_{p \in \{x_i, y_i\}} \mathrm{QA}(p) \tag{2}$$

Thus, ASIM provides a direct relationship between visual similarity and atmospheric quality: when $\mathrm{ASIM}(\mathbf{x}, \mathbf{y}) = \mathbf{1}$, the QA scores for all pixels in both images $\mathbf{x}$ and $\mathbf{y}$ must $=$ 1. It follows that for a corruption-free image $\mathbf{r}$, $\mathrm{QA}(\mathbf{r}) = \mathbf{1}$ and therefore $\mathrm{ASIM}(\mathbf{x}, \mathbf{r}) = \mathrm{QA}(\mathbf{x})$. This has two implications: first, $\mathrm{ASIM}(\mathbf{x}, \mathbf{y})$ is a sufficient primitive to identify clear (corruption-free) image pairs that we will call references. Second, given a reference $\mathbf{r}$, $\mathrm{ASIM}(\mathbf{x}, \mathbf{r})$ can directly yield the QA scores for an image $\mathbf{x}$. We therefore proceed to build our QA product by first modeling the two-image ASIM score, then use ASIM as a base primitive for training a QA model for which no reference is known. We believe this level of indirection is justified, as we will demonstrate, by our success in modeling ASIM in a low-shot learning regime.

3.3. Input imagery

We train both the ASIM and QA models on images from the Copernicus program's Sentinel-2 series of satellites [11]. Sentinel-2 imagery has thirteen spectral bands with resolutions ranging from 10m to 60m. Though numerous Sentinel-2 cloud detection algorithms exist [46], Sentinel-2 remains a high-value data stream to pursue for improving artifact detection. We focus entirely on Sentinel-2 L1C (top-of-atmosphere reflectance) products, though note that Sentinel-2 L2A (surface reflectance) products have identical registration and therefore our QA product applies to both. We log transform all model inputs using the following formula:

$$x' = \frac{\log{(x+1)}}{10} \tag{3}$$

This effectively compresses the long tail of reflectance values from high return surfaces (e.g., [18]) and normalizes inputs to [0, ∼1.1].

3.4. ASIM model

We produce a model that estimates ASIM from a pair of images taken at the same location in an order-independent way. For this we use an encoder/decoder FCN with two identical stems that are combined via element-wise sorting of the channels.

3.4.1 Pre-training

The ASIM pre-training task is to identify, from two pairs of images composited via a random mask, which pairs of pixels come from the same location at different times. This task requires robustness to snow, phenology, synthetic misregistration, and other non-atmospheric changes in value. We pre-train our ASIM model on an unlabeled set of ∼6M image pairs. These pairs were selected using the same general data mining protocol as described in Sec. 3.5 but with a Structural Similarity Index Measure (SSIM) [55], a widely used known-reference image similarity metric, rather than ASIM-based similarity measure.

3.4.2 Fine-tuning

ASIM model pre-training was followed by fine-tuning of probes on the decoder stages to make a probabilistic estimate of the actual ASIM score between pairs of images. This fine-tuning utilized a set of sparse annotations, synthetic artifacts, and linear mixup. Annotations were a combination of 388 hand-selected image pairs for which $\text{ASIM}(\mathbf{x}, \mathbf{y}) = \mathbf{1}$ uniformly (but no direct markup was provided), and 512 hand-selected image pairs with sparse markup across four grades [0, $1/3$, $2/3$, 1]. Synthetic clouds and shadows were generated to simulate a wide variety of cloud types, shadows, haze, elevations, and solar geometry with known ASIM scores, and both synthetic artifacts and annotations were combined using mixup [60], which preserves our definition of QA under a linear corruption model (Eq. (1)).

We produce a maximum likelihood estimation $(\mu, \log \sigma^2) = \text{ASIM}(\mathbf{x}, \mathbf{y}, \theta)$ (with model parameters θ) of the ground-truth QA score (μ') by minimizing a regularized log-likelihood:

$$z = \log \sigma^2|_{[a,b]} \tag{4}$$

$$r = (z - \log \sigma^2)^2 \tag{5}$$

$$l = \left(\log 2\pi + z + \frac{(\mu' - \mu)^2}{e^z}\right)\frac{1}{2} + r \tag{6}$$

Here r is introduced to prevent gradient plateau given that the domain of z is constrained to $[a, b]$ to improve numerical stability. We set $a = -16, b = 5$.

3.5. Video sequence sampling

To train a globally applicable QA model, we require a globally distributed sample that ideally balances both climatic variability and visually confusing scenarios: those that historically foil cloud and shadow detection methods. We accomplish this by balancing sampled data across the RESOLVE Ecoregions terrestrial biomes [8], and a set of pre-computed land conditions to facilitate sampling of confusing examples such as salt flats and snow, topographic shadows, cities and other high-return surfaces, oceans, and glacial lakes.

Our sampling initially targeted a set of twelve UTM zones selected to contain a high overall area of confusing categories (Fig. 1). Each UTM zone is gridded into 1.8 km cells in its planar projection, and then samples are drawn across yearly octants ($1/8$ of a year) with octant 0 centered on January 1. Octants are enlarged to encompass proportionally $1/5$ of a year to draw additional support images when necessary. Across each octant, we select all unique Sentinel-2 L1C datatakes at the latest processing baseline. We then use an ASIM model (see Sec. 3.4) to grade each video sequence to establish QA scores and determine suitability for inclusion in the training output.

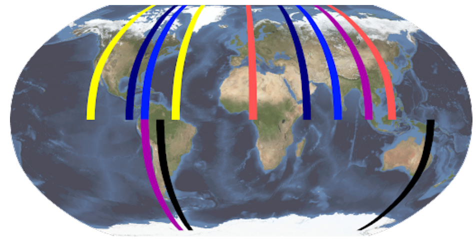


Figure 1. Twelve UTM zones sampled for training data, selected based on area of difficult surface types, specifically bright targets (yellow), topographic shadows (purple), urban/built-up areas (red), large bodies of water (dark blue), small bodies of water (bright blue), and ”other” (black).

We choose to observe image sequences sorted by mean spectral value with a sliding window in which every image is ASIM-scored to its two previous and following neighbors. Pairs for which the minimum ASIM score $(\mu, \log \sigma^2)$ achieves $\mu - \sigma > 0.75$ are nominally considered to be ”clear” references. Via inspection, the cutoff of 0.75 generally yielded clear references without conservative rejection of clear, yet uncertain pairs. If at least one pair of references exist in a sequence, all images are compared to the temporally closest reference to yield a score for each image under the previous assumption that for reference $\mathbf{r}$, $\text{ASIM}(\mathbf{x}, \mathbf{r}) = \text{QA}(\mathbf{x})$. In cases where no reference pair is found, the entire sequence is rejected even if it contains potentially viable clear images. Sequences may also be rejected for failing to meet minimum image criteria necessary to support all training tasks or for being ”too cloudy” or ”too clear” as these examples are less valuable for training. In general, liberal rejection is admissible given the scale of our data collection.

For each remaining sequence, additional data for model multi-tasking is computed, including a terrain illumination image derived from the ALOS World 3D 30m (AW3D30) global digital surface model [47]. See Fig. 2 for an example of a mined video sequence. After balancing our UTM-targeted sample to equalize across biomes and land conditions, the process of bootstrapping the ASIM primitive produced a training dataset of 2.2M video sequences with ASIM-based per-pixel QA grading for the years 2018 to 2021.

3.6. QA model

We use the large dataset of video sequences produced during sampling to train our reference-free QA model. This model takes a ”target” image to grade, and a ”query” set of images from the same location for support. Both the target and query set are first processed by an image feature extractor using the same FCN architecture as the ASIM model. Over the course of training, the length of the query is increased on a linear schedule from one to eight images,

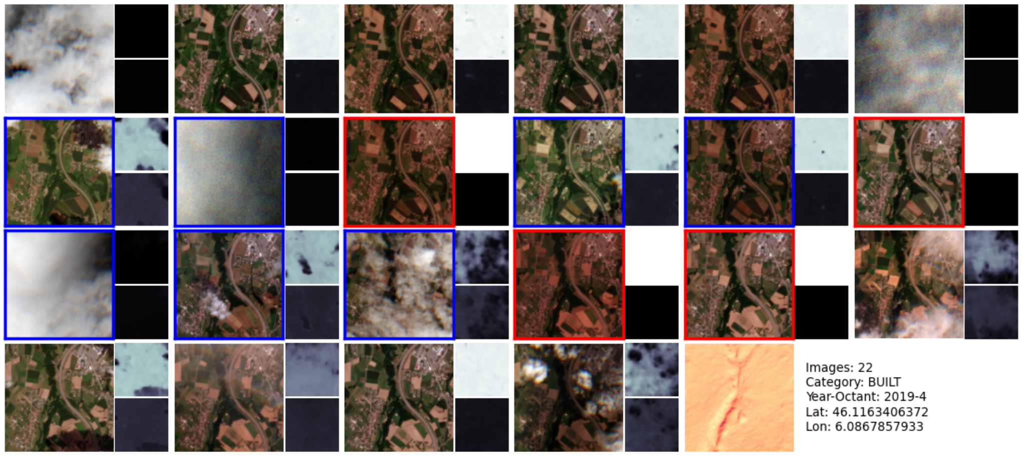


Figure 2. Example of a diagnostic filmstrip produced during sampling. Each image is 180x180 pixels at 10m resolution. Images outlined in red were selected as nominally clear reference images, images outlined in blue are used for training percentile outputs and images with no outlines are out-of-octant non-target images used only for prediction support. ASIM mean (top subplot, bright is higher QA and dark is lower QA) and standard deviation (bottom subplot) are shown to the right of each contrast-enhanced Sentinel-2 RGB, and additional properties and an illumination condition image are shown in the lower-right.

and swath boundaries, misregistration artifacts, and detector failure artifacts are simulated in both the target image and query images. Following feature extraction, a modification to a standard self-attention layer [4] with three heads uses the target features as the key to produce a single time-axial attended output from the query. This simplification is in part because our QA model is designed to run in a stateless system, one for which no previous model results are available and there is little benefit to saving cross-attended results for each query image. Finally, the attended query features and target features are combined in a per-pixel MLP for each task. Our model multi-tasks to a maximum likelihood estimate of terrain illumination at the target image following Eq. (6), a maximum likelihood estimate of QA $(\mu, \log\sigma^2)$ also following Eq. (6), and an estimate of the QA μ percentile $\mathrm{p}(x,\theta)$. The latter is achieved by minimizing:

$$\left(\mathrm{p}(x,\theta) - \left\{ \begin{array}{ll} 1 & \text{if } \mathrm{QA}_\mu(x) \geq \mathrm{QA}_\mu(y) \\ 0 & \text{o/w} \end{array} \right\}\right)^2 \quad (7)$$

As both images $\mathbf{x}$ and $\mathbf{y}$ are drawn from the same distribution of targets at a given location, the expectation $\mathbb{E}\left[\mathrm{QA}_\mu(x) \geq \mathrm{QA}_\mu(y)\right]$ is the cumulative distribution function (CDF) of QA for x. We also include a self-consistency penalty where model outputs for a random crop of the target image do not match those for the same region in the uncropped image. Of the aforementioned task head outputs, the estimated QA distribution is what is surfaced as our QA product, hereafter referred to as Cloud Score+ (CS+).

3.7. Inference

Generating CS+ QA products for new and historic Sentinel-2 L1C imagery is accomplished first by standard overtiling of the target image. For each sub-tile, a query set is established by searching for complete (not masked) Sentinel-2 datatakes overlapping the sub-tile area. The search concludes when a query limit is reached; in practice we use 16. The search is optimized to select acquisitions that are proximal to the target time of year, not necessarily in the target year. Once inference is complete for all sub-tiles, it is possible that adjacent sub-tiles exhibit seams related to the query search when unique sets of acquisitions are selected. We therefore perform a global optimization to compute a gain and bias (γ, β) for four control points at the sub-tile corners that define a per-tile bi-linear parameter interpolation to minimize the seam artifacts. QA μ is adjusted to $\mu' = \gamma\mu + \beta$ and σ is adjusted to $\sigma' = \gamma\sigma$. Along each overlapping seam for adjacent tiles P and Q with QA score means μ_p and μ_q, given an interpolated γ_{T_i} and β_{T_i} and position i, we minimize the following linear system for (γ, β):

$$\begin{aligned} \min_{\gamma,\beta} \ \lambda \left(\sum_T (\gamma_\mathrm{T} - 1)^2 + \beta_\mathrm{T}^2 \right) + \\ \sum_i \left(\gamma_{\mathrm{Q}_i} \mathrm{Q}_i + \beta_{\mathrm{Q}_i} - \gamma_{\mathrm{P}_i} \mathrm{P}_i - \beta_{\mathrm{P}_i} \right)^2 \end{aligned} \quad (8)$$

We use $\lambda = 1$. We found this modulation had no significant effect on our validation, but greatly reduced visual artifacts.

4. Evaluation

We assessed QA model results using two independently collected reference datasets [35, 48]. To compare our continuous CS+ QA products with reference labels and results from other state-of-the-art cloud masking algorithms, we reclassify all reference labels and processor output values to binary *clear* and *not clear* labels. We extracted mask values for each labeled point location in the reference datasets and computed a variety of standard classification metrics including F1 score, overall accuracy, omission error, commission error, precision, and recall. For continuous products, we apply a threshold (t) to create a binary mask. We tested thresholds between 0 and 1 at 0.01 intervals and we present results for threshold values that achieved the most balanced omission and commission error rates. In addition to accuracy metrics, we visually compared source imagery and mask results and the distributions of our CS+ QA scores for *clear* and *not clear* labels.

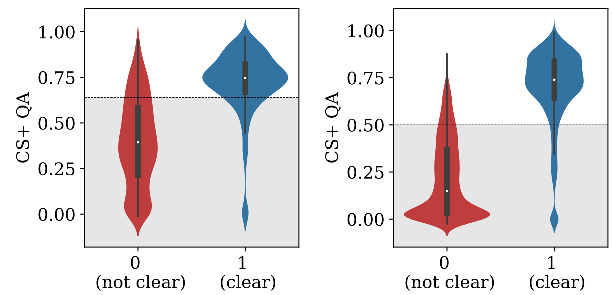
4.1. Tarrio reference dataset

The Tarrio reference dataset [48] includes 2,681 interpreted points and cloud masks for 28 products (images) from six S2 tiles in the Eastern Hemisphere. Interpreted points were labeled as cloud, cloud shadow, and clear, and we combined the cloud and cloud shadow labels into our *not clear* category. A total of 50 points were removed from the original set due to irreconcilable mismatches between the CLASS label and CLASS_ID value. Though the points were selected using a stratified sample over algorithm agreement, we do not account for strata weights in our assessments given differences in objectives and labeling schemes.

The original Tarrio et. al study [49] compared five different cloud-masking algorithms: Sen2Cor [42], MAJA [17], LaSRC [53], Fmask 1.0 [65], Fmask 2.0, and Fmask 4.0 [40], and Tmask [66]. Of these algorithms, Sen2Cor, LaSRC and Fmask are single-date, while MAJA and Tmask use temporal context. Masks were provided with a standardized legend with classes for cloud and cloud shadows, which we relabeled *not clear*, and clear land, clear water, and snow/ice, which we relabeled *clear*. In addition to the seven masks included in the Tarrio set, we also include the Sentinel-2 QA60 bitmask and s2cloudless in our comparisons. The QA60 bitmask is a standard Sentinel-2 "quality assurance" band included with all L1C images [2, 7, 11] and represents a current operational baseline, while s2cloudless represents an existing state-of-the-art machine-learning-based image QA solution [46,67]. Because s2cloudless does not include a cloud shadow class, we produce a variation on our reclassifications that labels shadow points as *clear*.

Our QA approach (CS+) had the highest F1 score (0.8466) and overall accuracy (0.8096) on the Tarrio reference dataset with a recommended threshold of 0.64 (Tab. 1). We improve on both errors of omission and commission relative to the next-best processor, Fmask v2, with a large margin of improvement in F1 score and overall accuracy relative to other top processors, while widely available products exhibit among the worst performance.

Considering the distribution of CS+ QA scores for *clear* and *not clear* label aggregations, we find that the 0.64 threshold adequately distinguished *clear* observations while there was greater ambiguity for the *not clear* class (Fig. 3a). Visually inspecting results, we see continuous metrics provide more nuanced information on per-pixel usability, capturing a range of atmospheric interference and occlusion with greater precision than categorical masks, e.g., Fig. 4.



(a) Tarrio reference dataset. (b) CMIX-I reference dataset.

Figure 3. Violin plots of QA scores for *clear* and *not clear* labels. Dotted lines show recommended thresholds for each dataset.

4.2. CMIX-I reference dataset

The Sentinel-2 PixBox dataset was used as a validation reference for CMIX-I assessments and comparisons [46]. The dataset consists of 17,351 interpreted pixels for 29 Sentinel-2 Level 1C products [35]. Trained experts manually classified each point in the PixBox dataset using a very detailed set of categories describing surface conditions, cloud characteristics, shadows, aerosols, sun glint, water bodies, and types of ice such that each point is assigned a classification within each category. For the CMIX-I reference dataset, we used a combination of CLOUD_CHARACTERISTICS_ID, SHADOW_ID, and AEROSOL TYPE_ID to determine the final binary label.

CS+ consistently outperformed s2cloudless across all labeling schemes and mask thresholds considered for the CMIX-I dataset (Tab. 2). The greatest difference in performance was observed when using the full set of reference points (*all points*), with CS+ achieving an F1 score of 0.8816 and overall accuracy of 0.8768, while s2cloudless had an F1 score of 0.8162 and overall accuracy of 0.8094, which is unsurprising given that s2cloudless does not identify cloud shadows as "bad QA". When removing shadows from the reference dataset (*no shadows*), CS+ performance decreases slightly, but still exceeds that of s2cloudless by at least 0.03 in terms of both F1 score and overall accuracy. We also find that treating shadows as clear (*clear shadows*) in order to keep the number of points in the refer-

Name	Threshold	F1	Overall	Omission	Commission	Precision	Recall
CS+	0.64	**0.8466**	**0.8096**	0.1889	0.1933	0.8854	0.8111
Fmask v2	-	0.8133	0.7742	0.2411	0.1976	0.8761	0.7589
MAJA	-	0.8002	0.7218	**0.1402**	0.5324	0.7483	**0.8598**
Tmask	-	0.7861	0.7290	0.2317	0.3434	0.8047	0.7683
LaSRC	-	0.7784	0.7210	0.2440	0.3434	0.8021	0.7560
s2cloudless	0.12	0.7589	0.7111	0.2985	0.2711	0.8265	0.7015
Fmask v4	-	0.7444	0.7233	0.3783	0.0896	0.9274	0.6217
s2cloudless*	0.15	0.6650	0.6994	0.2889	0.3091	0.6245	0.7111
Sen2Cor	-	0.5190	0.5675	0.6399	0.0508	0.9289	0.3601
Fmask v1	-	0.5094	0.5817	0.6504	0.0379	0.9379	0.3496
QA60	-	0.4579	0.5401	0.7003	**0.0173**	**0.9696**	0.2997

Table 1. Accuracy metrics for interpreted points from the Tarrio reference dataset. Asterisks (*) indicate experiments where s2cloudless was assessed with shadows considered *clear*. Thresholds are selected to balance omission and commission error.

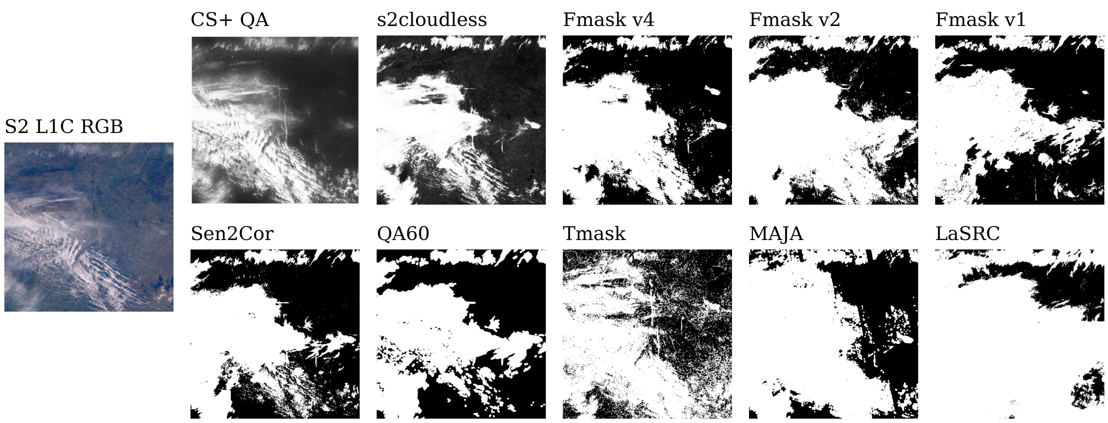


Figure 4. Example results from different processors for a Sentinel-2 image (32TLT, 2016-09-12) from the Tarrio reference dataset. White (0) indicates not clear and black (1) indicates clear.

ence dataset consistent or using the provider-recommended threshold of 0.40 instead of the optimal value based on balanced omission-commission error did not significantly affect results, though we do find the lowest rates of commission for s2cloudless using a 0.40 threshold for the *no shadows* dataset.

The optimal threshold for the CMIX-I reference dataset was relatively consistent whether or not shadows were included (0.50 versus 0.48) and was notably lower than the threshold identified for the Tarrio dataset (0.64). The greater separability of the *clear* and *not clear* groupings for the CMIX-I dataset (Fig. 3b) is likely indicative of higher-quality labels, especially since the CMIX-I image set represents more diverse geographic and atmospheric conditions. Although tuned thresholds for individual reference datasets serve as a recommendation for creating binary masks, users may select thresholds that work best for their specific use cases. Alternatively, continuous CS+ QA values can be used directly for building ”clearest pixel” composites or as weights on individual observations. Visualizing the CS+ results for select images from the CMIX-I dataset, we note that CS+ is able to characterize a variety of cloud types, including high cirrus and haze (Fig. 5). CS+ also performs well on very challenging examples, including a scene-level gradient in cloud cover over snow and ice, and a scene dominated by ice-capped mountainous terrain with small, dense clouds along ridges and valleys (bottom two rows of Fig. 5).

4.3. Limitations & future work

Our QA approach generally demonstrates strong performance for historically challenging use cases, however, known weaknesses include detection of cloud shadows over water, mis-characterization of moving boats as bad QA, and high uncertainty over active ice floes. Because the QA model estimates both the mean and variance of the QA prediction, the variance can be used to further constrain or filter results. This both enhances interpretability and allows users greater flexibility in navigating potential failure cases.

Name	Reference	Threshold	F1	Overall	Omission	Commission	Precision	Recall
CS+	all points	0.50	**0.8816**	0.8768	0.1213	0.1253	**0.8846**	0.8787
CS+	no shadows	0.48	0.8800	**0.8841**	**0.1156**	0.1162	0.8756	**0.8844**
s2cloudless	no shadows	0.24	0.8419	0.8471	0.1523	0.1536	0.8362	0.8477
s2cloudless	clear shadows	0.24	0.8376	0.8453	0.1535	0.1558	0.8289	0.8465
s2cloudless	no shadows	0.40*	0.8263	0.8423	0.2189	**0.1011**	0.8772	0.7811
s2cloudless	clear shadows	0.40*	0.8218	0.8413	0.2232	0.1013	0.8724	0.7768
s2cloudless	all points	0.18	0.8162	0.8094	0.1899	0.1913	0.8223	0.8101

Table 2. Accuracy metrics for interpreted points from the CMIX-I PixBox reference dataset. Asterisks (*) indicate experiments using the provider's recommended threshold, otherwise thresholds are selected to balance omission and commission error.

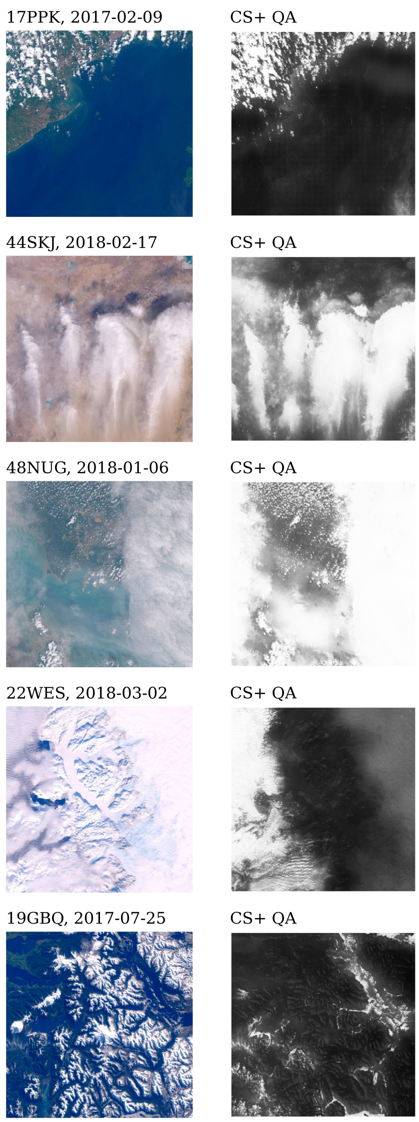


Figure 5. Sentinel-2 L1C images and corresponding Cloud Score+ QA results for six images from the CMIX-I PixBox dataset.

In future work, we plan to further validate performance on additional benchmarks, specifically the recently released CloudSEN12 dataset [2]. Our approach is also designed to generalize to other sensors even in the absence of dense image time-series, i.e., we are able to make single-date predictions for other sensors by using Sentinel-2 as support. Early results for Landsat 8 and 9 top-of-atmosphere images with the exact model and parameters used for Sentinel-2 have been very promising (Figure Fig. 6) and formal evaluation is currently underway.

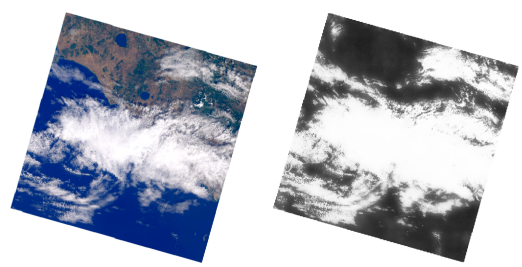


Figure 6. Example of Cloud Score+ QA results for a Landsat 8 image (Path 191, Row 31, 2021-09-03).

5. Conclusions

In this paper, we introduce a weakly supervised video analysis approach for characterizing the quality of observations acquired by optical satellite instruments, specifically Sentinel-2. Our continuous image QA results are designed to circumvent limitations of categorical cloud masks by instead scoring the usability of a given observation on a continuous scale. We demonstrate state-of-the-art performance on two independently collected reference datasets. We expect to generate CS+ QA products for all historic and newly acquired Sentinel-2 L1C images, providing a novel and flexible solution for identifying the most useful pixels for terrestrial monitoring.

References

[1] Tom Augspurger. Scalable sustainability with the planetary computer. In *AGU Fall Meeting Abstracts*, volume 2021, pages U51B–14, 2021. 1

[2] Cesar Aybar, Luis Ysuhuaylas, Jhomira Loja, Karen Gonzales, Fernando Herrera, Lesly Bautista, Roy Yali, Angie Flores, Lissette Diaz, Nicole Cuenca, et al. Cloudsen12, a global dataset for semantic understanding of cloud and cloud shadow in sentinel-2. *Scientific Data*, 9(1):782, 2022. 2, 6, 8

[3] Louis Baetens, Camille Desjardins, and Olivier Hagolle. 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(4):433, 2019. 2

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 5

[5] John E Ball, Derek T Anderson, and Chee Seng Chan. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *Journal of Applied Remote Sensing*, 11(4):042609–042609, 2017. 1

[6] Danang Surya Candra, Stuart Phinn, and Peter Scarth. Cloud and cloud shadow masking for Sentinel-2 using multitemporal images in global area. *International Journal of Remote Sensing*, 41(8):2877–2904, 2020. 2

[7] Rosa Coluzzi, Vito Imbrenda, Maria Lanfredi, and Tiziana Simoniello. A first assessment of the Sentinel-2 Level 1-C cloud mask product to support informed surface analyses. *Remote Sensing of Environment*, 217:426–443, 2018. 1, 6

[8] Eric Dinerstein, David Olson, Anup Joshi, Carly Vynne, Neil D Burgess, Eric Wikramanayake, Nathan Hahn, Suzanne Palminteri, Prashant Hedao, Reed Noss, et al. An ecoregion-based approach to protecting half the terrestrial realm. *BioScience*, 67(6):534–545, 2017. 4

[9] Marharyta Domnich, Indrek Sünter, Heido Trofimov, Olga Wold, Fariha Harun, Anton Kostiukhin, Mihkel Järveoja, Mihkel Veske, Tanel Tamm, Kaupo Voormansik, et al. KappaMask: Ai-based cloudmask processor for Sentinel-2. *Remote Sensing*, 13(20):4100, 2021. 1

[10] Johannes Drönner, Nikolaus Korfhage, Sebastian Egli, Markus Mühling, Boris Thies, Jörg Bendix, Bernd Freisleben, and Bernhard Seeger. Fast cloud segmentation using convolutional neural networks. *Remote Sensing*, 10(11):1782, 2018. 1

[11] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 120:25–36, 2012. 3, 6

[12] John L Dwyer, David P Roy, Brian Sauer, Calli B Jenkerson, Hankui K Zhang, and Leo Lymburner. Analysis ready data: enabling analysis of the Landsat archive. *Remote Sensing*, 10(9):1363, 2018. 1

[13] Steve Foga, Pat L Scaramuzza, Song Guo, Zhe Zhu, Ronald D Dilley Jr, Tim Beckmann, Gail L Schmidt, John L Dwyer, M Joseph Hughes, and Brady Laue. Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote Sensing of Environment*, 194:379–390, 2017. 1

[14] David Frantz, Erik Haß, Andreas Uhl, Johannes Stoffels, and Joachim Hill. 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, 2018. 1, 2

[15] Jianhao Gao, Qiangqiang Yuan, Jie Li, Hai Zhang, and Xin Su. Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks. *Remote Sensing*, 12(1):191, 2020. 2

[16] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017. 1

[17] Olivier Hagolle, Mireille Huc, D Villa Pascual, and Gérard Dedieu. 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, 2010. 2, 6

[18] André Hollstein, Karl Segl, Luis Guanter, Maximilian Brell, and Marta Enesco. 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, 2016. 1, 3

[19] Robert Horvath, John G Braithwaite, and Fabian C Polcyn. Effects of atmospheric path on airbornemultispectral sensors. *Remote Sensing of Environment*, 1(4):203–215, 1970. 3

[20] H Huang and DP Roy. Characterization of Planetscope-0 Planetscope-1 surface reflectance and normalized difference vegetation index continuity. *Science of Remote Sensing*, 3:100014, 2021. 2

[21] Richard R Irish, John L Barker, Samuel N Goward, and Terry Arvidson. Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm. *Photogrammetric Engineering & Remote Sensing*, 72(10):1179–1188, 2006. 1

[22] Viktoria Kristollari and Vassilia Karathanassi. Artificial neural networks for cloud masking of Sentinel-2 ocean images with noise and sunglint. *International Journal of Remote Sensing*, 41(11):4102–4135, 2020. 1

[23] Viktoria Kristollari and Vassilia Karathanassi. Fine-tuning Self-Organizing Maps for Sentinel-2 imagery: Separating clouds from bright surfaces. *Remote Sensing*, 12(12):1923, 2020. 1

[24] Liyuan Li, Xiaoyan Li, Linyi Jiang, Xiaofeng Su, and Fansheng Chen. A review on deep learning techniques for cloud detection methodologies and challenges. *Signal, Image and Video Processing*, 15(7):1527–1535, 2021. 1

[25] Cheng-Chien Liu, Yu-Cheng Zhang, Pei-Yin Chen, Chien-Chih Lai, Yi-Hsin Chen, Ji-Hong Cheng, and Ming-Hsun Ko. Clouds classification from Sentinel-2 imagery with deep residual learning and semantic image segmentation. *Remote Sensing*, 11(2):119, 2019. 1

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Pro-*

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3431–3440, 2015. 1

[27] Jérôme Louis, Bringfried Pflug, Vincent Debaecker, Uwe Mueller-Wilm, Rosario Quirino Iannone, Valentina Boccia, and Ferran Gascon. Evolutions of Sentinel-2 Level-2A cloud masking algorithm Sen2Cor prototype first results. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 3041–3044. IEEE, 2021. 1

[28] Jérôme Louis, Bringfried Pflug, Magdalena Main-Knorn, Vincent Debaecker, Uwe Mueller-Wilm, Rosario Quirino Iannone, Enrico Giuseppe Cadau, Valentina Boccia, and Ferran Gascon. Sentinel-2 global surface reflectance level-2A product generated with Sen2Cor. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 8522–8525. IEEE, 2019. 1

[29] Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, and Brian Alan Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152:166–177, 2019. 1

[30] Magdalena Main-Knorn, Bringfried Pflug, Jerome Louis, Vincent Debaecker, Uwe Müller-Wilm, and Ferran Gascon. Sen2Cor for Sentinel-2. In *Image and Signal Processing for Remote Sensing XXIII*, volume 10427, pages 37–48. SPIE, 2017. 1

[31] Gonzalo Mateo-García, Luis Gómez-Chova, Julia Amorós-López, Jordi Muñoz-Marí, and Gustau Camps-Valls. Multitemporal cloud masking in the Google Earth Engine. *Remote Sensing*, 10(7):1079, 2018. 2

[32] Andrea Meraner, Patrick Ebel, Xiao Xiang Zhu, and Michael Schmitt. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:333–346, 2020. 2

[33] Grega Milcinski, Matej Batic, Miha Kadunc, Primoz Kolaric, Rok Mocnik, et al. SENTINEL-2 Services Library-efficient way for exploration and exploitation of EO data. In *EGU General Assembly Conference Abstracts*, page 19502, 2017. 1

[34] Fabrizio Niro, Philippe Goryl, Steffen Dransfeld, Valentina Boccia, Ferran Gascon, Jennifer Adams, Britta Themann, Silvia Scifoni, and Georgia Doxani. European Space Agency (ESA) calibration/validation strategy for optical land-imaging satellites and pathway towards interoperability. *Remote Sensing*, 13(15):3003, 2021. 1, 2

[35] Michael Paperin, Jan Wevers, Kerstin Stelzer, and Carsten Brockmann. Pixbox Sentinel-2 pixel collection for CMIX, June 2021. 5, 6

[36] Pedro Pérez-Cutillas, Alberto Pérez-Navarro, Carmelo Conesa-García, Demetrio Antonio Zema, and Jesús Pilar Amado-Álvarez. What is going on within google earth engine? a systematic review and meta-analysis. *Remote Sensing Applications: Society and Environment*, 29:100907, 2023. 1

[37] Bringfried Pflug, Jérôme Louis, Vincent Debaecker, Uwe Mueller-Wilm, Carine Quang, Ferran Gascon, and Valentina Boccia. Next updates of atmospheric correction processor Sen2Cor. In *Image and Signal Processing for Remote Sensing XXVI*, volume 11533, pages 12–18. SPIE, 2020. 1

[38] Jernej Puc and Lojze Žust. On cloud detection with multi-temporal data, 2019. https://medium.com/sentinel-hub/on-cloud-detection-with-multi-temporal-data-f64f9b8d59e5. Accessed on 2023-01-14. 2

[39] Shi Qiu, Binbin He, Zhe Zhu, Zhanmang Liao, and Xingwen Quan. Improving Fmask cloud and cloud shadow detection in mountainous area for Landsats 4–8 images. *Remote Sensing of Environment*, 199:107–119, 2017. 1

[40] Shi Qiu, 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*, 231:111205, 2019. 6

[41] Shi Qiu, Zhe Zhu, and Curtis E Woodcock. Cirrus clouds that adversely affect Landsat 8 images: What are they and how to detect them? *Remote Sensing of Environment*, 246:111884, 2020. 1

[42] R Richter, J Louis, and U Müller-Wilm. Sentinel-2 MSI—Level 2A products algorithm theoretical basis document. *European Space Agency,(Special Publication) ESA SP*, 49(0):1–72, 2012. 6

[43] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. SEN12MS–a curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion. *arXiv preprint arXiv:1906.07789*, 2019. 2

[44] Fernando Sedano, Pieter Kempeneers, Peter Strobl, Jan Kucera, Peter Vogt, Lucia Seebach, and Jesús San-Miguel-Ayanz. A cloud mask methodology for high resolution remote sensing data combining information from high and medium resolution optical sensors. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(5):588–596, 2011. 1

[45] Yuri Shendryk, Yannik Rist, Catherine Ticehurst, and Peter Thorburn. Deep learning for multi-modal classification of cloud, shadow and land cover scenes in PlanetScope and Sentinel-2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 157:124–136, 2019. 1

[46] Sergii Skakun, Jan Wevers, Carsten Brockmann, Georgia Doxani, Matej Aleksandrov, Matej Batič, David Frantz, Ferran Gascon, Luis Gómez-Chova, Olivier Hagolle, et al. Cloud Mask Intercomparison eXercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2. *Remote Sensing of Environment*, 274:112990, 2022. 1, 2, 3, 6

[47] T Tadono, H Ishida, F Oda, S Naito, K Minakawa, and H Iwamoto. Precise global DEM generation by ALOS PRISM. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2(4):71, 2014. 4

[48] Xiaojing Tang, Katelyn Tarrio, Jeffrey Masek, Martin Claverie, Junchang Ju, Shi Qiu, Zhe Zhu, Shijuan Chen, Qiyuan Fu, Yihao Liu, Xianfei Shen, Yetianjian Wang, Yingtong Zhang, Chongyang Zhu, and Curtis Woodcock. Reference dataset for comparison of cloud detection algorithms for Sentinel-2 imagery, Nov. 2021. This research was funded by NASA through both the Harmonized Landsat Sentinel effort and the Making Earth System Data Records for Use in Research Environments (MEaSUREs) Program, as well as the USGS through the Landsat Science Team. 5, 6

[49] Katelyn Tarrio, Xiaojing Tang, Jeffrey G Masek, Martin Claverie, Junchang Ju, Shi Qiu, Zhe Zhu, and Curtis E

Woodcock. Comparison of cloud detection algorithms for Sentinel-2 imagery. *Science of Remote Sensing*, 2:100010, 2020. 6

[50] Planet Team. UDM 2, 2023. https://developers.planet.com/docs/data/udm-2/. Accessed on 2023-02-24. 2

[51] Dirk Tiede, Martin Sudmanns, Hannah Augustin, and Andrea Baraldi. Investigating ESA Sentinel-2 products' systematic cloud cover overestimation in very high altitude areas. *Remote Sensing of Environment*, 252:112163, 2021. 1

[52] Robert E Turner. Atmospheric effects in multispectral remote sensor data. Technical report, Environmental Research Institute of Michigan, May 1975. ERIM 109600-15-F. 3

[53] Eric Vermote, Chris Justice, Martin Claverie, and Belen Franch. Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. *Remote Sensing of Environment*, 185:46–56, 2016. 6

[54] Bin Wang, Atsuo Ono, Kanako Muramatsu, and Noboru Fujiwara. Automated detection and removal of clouds and their shadows from Landsat TM images. *IEICE Transactions on information and systems*, 82(2):453–460, 1999. 2

[55] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3

[56] Michael A Wulder, Joanne C White, Thomas R Loveland, Curtis E Woodcock, Alan S Belward, Warren B Cohen, Eugene A Fosnight, Jerad Shaw, Jeffrey G Masek, and David P Roy. The global Landsat archive: Status, consolidation, and direction. *Remote Sensing of Environment*, 185:271–283, 2016. 1

[57] Quan Xiong, Guoqing Li, Xiaochuang Yao, and Xiaodong Zhang. SAR-to-optical image translation and cloud removal based on conditional generative adversarial networks: Literature survey, taxonomy, evaluation indicators, limits and future directions. *Remote Sensing*, 15(4):1137, 2023. 2

[58] Nicholas E Young, Ryan S Anderson, Stephen M Chignell, Anthony G Vorster, Rick Lawrence, and Paul H Evangelista. A survival guide to Landsat preprocessing. *Ecology*, 98(4):920–932, 2017. 1

[59] Yongjie Zhan, Jian Wang, Jianping Shi, Guangliang Cheng, Lele Yao, and Weidong Sun. Distinguishing cloud and snow in satellite images via deep convolutional network. *IEEE Geoscience and Eemote Sensing Letters*, 14(10):1785–1789, 2017. 1

[60] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 4

[61] Mingmin Zhao, Peder A Olsen, and Ranveer Chandra. Seeing through clouds in satellite images. *arXiv preprint arXiv:2106.08408*, 2021. 2

[62] Xiaolin Zhu and Eileen H Helmer. An automatic method for screening clouds and cloud shadows in optical satellite image time series in cloudy regions. *Remote Sensing of Environment*, 214:135–153, 2018. 1

[63] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017. 1

[64] Zhe Zhu, 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*, 159:269–277, 2015. 1

[65] Zhe Zhu and Curtis E Woodcock. Object-based cloud and cloud shadow detection in landsat imagery. *Remote Sensing of Environment*, 118:83–94, 2012. 1, 6

[66] Zhe Zhu and Curtis E Woodcock. Automated cloud, cloud shadow, and snow detection in multitemporal Landsat data: An algorithm designed specifically for monitoring land cover change. *Remote Sensing of Environment*, 152:217–234, 2014. 1, 2, 6

[67] A Zupanc. Improving cloud detection with machine learning, 2017. https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13. Accessed on 2023-01-14. 2, 6