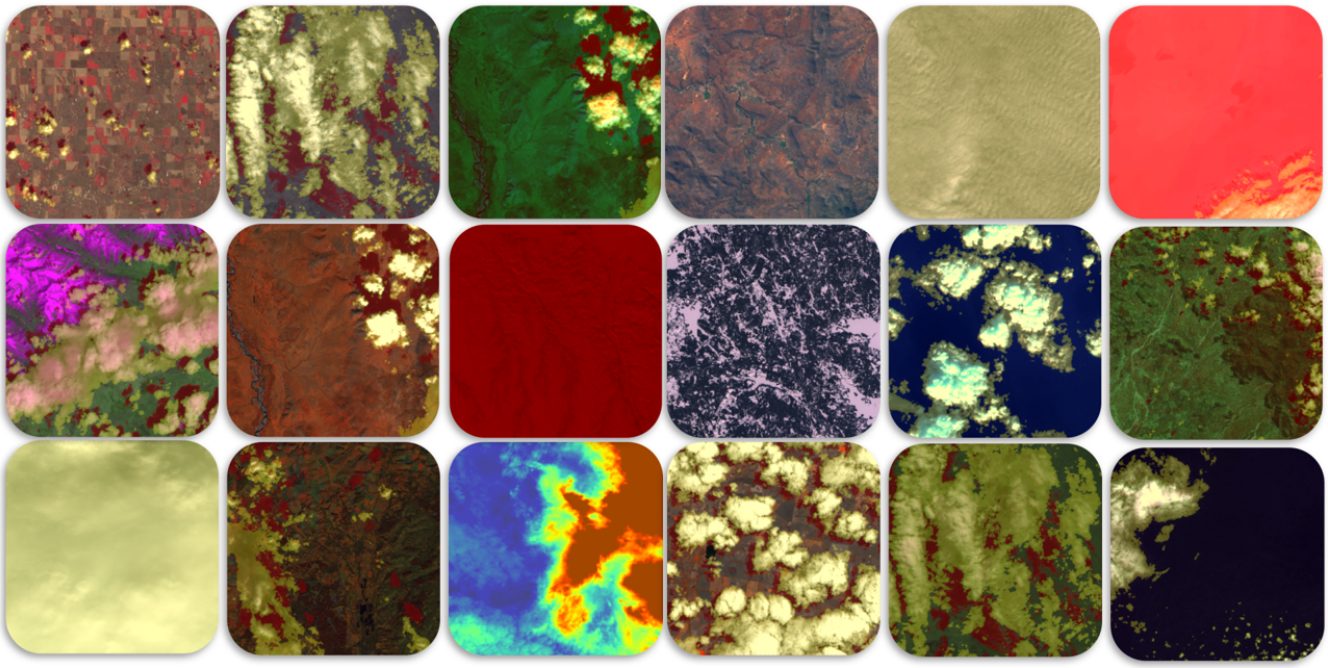
Sentinel-2 Cloud Mask Catalogue



1. General information

This dataset comprises cloud masks for 513 1022-by-1022 pixel subscenes, at 20m resolution, sampled random from the 2018 Level-1C Sentinel-2 archive. The design of this dataset follows from some observations about cloud masking: (i) performance over an entire product is highly correlated, thus subscenes provide more value per-pixel than full scenes, (ii) current cloud masking datasets often focus on specific regions, or hand-select the products used, which introduces a bias into the dataset that is not representative of the real-world data, (iii) cloud mask performance appears to be highly correlated to surface type and cloud structure, so testing should include analysis of failure modes in relation to these variables.

The data was annotated semi-automatically, using the IRIS toolkit, which allows users to dynamically train a Random Forest (implemented using LightGBM), speeding up annotations by iteratively improving it's predictions, but preserving the annotator's ability to make final manual changes when needed. This hybrid approach allowed us to process many more masks than would have been possible manually, which we felt was vital in creating a large enough dataset to approximate the statistics of the whole Sentinel-2 archive.

In addition to the pixel-wise, 3 class (*CLEAR*, *CLOUD*, *CLOUD_SHADOW*) segmentation masks, we also provide users with binary classification "tags" for each subscene that can be used in testing to determine performance in specific circumstances. These include:

- **SURFACE TYPE**: *11 categories*
- **CLOUD TYPE**: *7 categories*
- **RELATIVE CLOUD HEIGHT**: *low, high*
- **CLOUD THICKNESS**: *thin, thick*
- **CLOUD EXTENT**: *isolated, extended*

Wherever practical, cloud shadows were also annotated, however this was sometimes not possible due to high-relief terrain, or large ambiguities. In total, 424 were marked with shadows (if present), and 89 have shadows that were not annotatable due to very ambiguous shadow boundaries, or terrain that cast significant shadows. If users wish to train an algorithm specifically for cloud shadow masks, we advise them to remove those 89 images for which shadow was not possible, however, bear in mind that this will systematically reduce the difficulty of the shadow class compared to real-world use, as these contain the most difficult shadow examples.

In addition to the 20m sampled subscenes and masks, we also provide users with shapefiles that define the boundary of the mask on the original Sentinel-2 scene. If users wish to retrieve the L1C bands at their original resolutions, they can use these to do so.

2. Dataset Description

We will now describe the structure of the overall dataset, and then go into more detail for each specific component.

```
DATASET/
|   README.pdf
|   classification_tags.csv
|
└───subscenes/
|   |   <scene1>.npy
|   |   <scene2>.npy     ---->  1022x1022x13, float32, TOA reflectance
|   |   <scene3>.npy
|   |   ....
└───masks/
|   |   <scene1>.npy
|   |   <scene2>.npy     ---->  1022x1022x3, bool, one-hot encoded classes
|   |   <scene3>.npy
|   |   ...
└───shapefiles/
|   └───<scene1>/
|   |   |   <scene1>.cpg  \
|   |   |   <scene1>.dbf   \
|   |   |   <scene1>.prj    }-> Esri shapefile for mask/subscene outline
|   |   |   <scene1>.shp   /
|   |   |   <scene1>.shx  /
|   └───<scene2>/
|   |   |   ...
|   |   ...
└───thumbnails/
|   |   <scene1>.png
|   |   <scene2>.png     ----> 512x512x3, uint8, downsampled image for quickly
|   |   <scene3>.png                            inspecting subscene
|   |   ...
└───alt_masks/
|   |   ALT_<scene1>.npy
|   |   ALT_<scene2>.npy
|   |   ALT_<scene3>.npy
|   |   ....
```

classification_tags.csv

The classification_tags.csv file contains several parameters that should be helpful for users. All the categorical tags are non-exclusive, and simply mean a given subscene contained some of that characteristic. For example, snow-covered mountain tops with forests in the lower valleys, would be tagged with *forest/jungle*, *snow/ice* and *hills/mountains*. Similarly multiple cloud types can exist within the same image. All tags were determined by visual inspection of both the image and, for **SURFACE TYPE**, other high-resolution imagery of the area (e.g. BingMaps and Google Earth). Objective classifications of this kind are not possible, and all our classifications here are made using subjective judgement. We describe each column (grouped into broad categories) as follows:

GENERAL INFORMATION:

- **scene**: string, Sentinel-2 product ID.
- **difficulty**: int, Subjective measure from 1->5 for difficulty of annotation.
  - 1. Near perfect
  - 2. Very good
  - 3. Mostly good
  - 4. Possibly some small errors
  - 5. Possibly large errors
- **annotator**: string, Which of the two annotators, *A* or *B*, created the mask used in *masks/* (note that *alt_masks/* may contain the other annotator's mask, if both labelled that subscene).
- **shadows_marked**: boolean, if 0 then no shadow markings were possible. If 1, shadows were marked where present.

- **clear_percent**: float, 0 -> 100 for percentage of pixels marked *CLEAR*
- **cloud_percent**: float, 0 -> 100 for percentage of pixels marked *CLOUD*
- **shadow_percent**: float, 0 -> 100 for percentage of pixels marked *CLOUD_SHADOW*
- **dataset**: string, referring to whether subscene was part of *CALIBRATION*, *MAIN*, or *VALIDATION* labelling stages (described in Section "Annotation Strategy")

SURFACE TYPES:

- **forest/jungle**: boolean, dense tree cover
- **snow/ice**: boolean, any snow or ice covering either land or water
- **agricultural**: boolean, farmlands or pastures
- **urban/developed**: boolean, human structures, conurbations, villages, industrial sites. Not including solitary roads.
- **coastal**: boolean, coastlines, specifically. (Not just water near to land, but the division between land and water itself)
- **hills/mountains**: boolean, noticeable hilly terrain
- **desert/barren**: boolean, completely arid, or very sparsely vegetated, drylands.
- **shrublands/plains**: boolean, somewhat sparsely vegetated areas, or non-forested areas without obvious signs or agriculture.
- **wetland/bog/marsh**: boolean, areas with large amounts of standing water (although)
- **open_water**: boolean, extended bodies of water, such as large lakes that continue beyond the image bounds, or the sea.
- **enclosed_water**: boolean, lakes, large rivers or contained areas of sea-water (e.g. fjords)

CLOUD THICKNESS:

- **thin**: boolean, any cloud where some surface colour/signal can be seen through it (not including at the edges of thick clouds, to avoid all thick tags being accompanied by thin ones)
- **thick**: boolean, any cloud where no surface can be seen through it (except possibly at its edges).

RELATIVE CLOUD HEIGHT:

- **low**: boolean, any cloud which has no obvious sign of high-altitude (high reflectance in the Cirrus (B10) band)
- **high**: boolean, any cloud with noticable reflectance in the Cirrus (B10) band

CLOUD EXTENT:

- **isolated**: boolean, clouds which are separated from one another, such that there are no areas of continuous cloud across the majority of the subscene.
- **extended**: boolean, clouds which span continuously across a majority of a subscene.

CLOUD TYPE:

- **cumulus**: boolean, rounded, clumpy and separate clouds.
- **cumulonimbus**: boolean, clouds with very large, steep, vertical profiles, reaching up very high.
- **altocumulus/stratocumulus**: boolean, most general cloud type, denoting any clouds that form periodic cells. Could be high or low, and thick or thin.
- **cirrus**: boolean, thin, extended clouds that have high reflectance in the Cirrus (B10) band. Sometimes almost invisible or very faint in any other band.
- **haze/fog**: boolean, low-lying, textureless, thin cloud. Often distinguishable because it looks similar to cirrus, but with no reflectance in B10.
- **ice_clouds**: boolean, any cloud that has a noticeable icy signature, seen as reddish-orange in the false-colour B1 (red), B11 (green), B12 (blue)
- **contrails**: boolean, exhaust trails left by planes.

subscenes

Each subscene is a 1022-by-1022-by-13 numpy array, with Top of Atmosphere reflectance values held as float32 numbers. These are taken directly from the Sentinel-2 L1C product that they share their name with, cropped randomly (but resampled until an area without any no-data values is found). Bands that are not at 20m are resampled to 20m using bilinear interpolation. Users who wish to use the bands at their original resolutions will need to download the L1C products and use the *shapefiles* provided to extract the area of the mask.

The order of Sentinel-2 bands in the numpy array's 3rd dimension is in numerical order, with band 8A coming between bands 8 and 9. All values are found by dividing the original L1C integer values by 10'000, to retrieve the reflectances, as advised in the Product Specification for Sentinel-2 L1C. Please note that many values go above 1, because apparent reflectance can be greater than 1 if a surface is at an angle which receives more light than is possible on a surface normal to the viewing angle.

masks

Each mask are a 1022-by-1022-by-3 numpy array, with boolean one-hot encoding (each pixel has exactly one True value across the last dimension). The class order in the last dimension is: *CLEAR*, *CLOUD*, *CLOUD_SHADOW*. Even if shadow markings were not possible for a given subscene, the third channel is still included. The masks are named using the corresponding product IDs of the scenes.

shapefiles

An *Esri* shapefile for each scene which describes the polygon of the extracted subscene. These can be used to extract the original band data from the Sentinel-2 L1C scenes if users need them.

thumbnails

A set of downsampled .png images showing the subscenes, plus a small area around them. Not to be used for any processing, they are included to give a simple way to browse the data.

alt_masks

For scenes which were annotated by both annotators (10 in *CALIBRATION* and 50 in *VALIDATION*—see *Appendix B* for more details about this process) the other annotators mask is provided for completeness. This can be used to verify the statistical tests we performed using them, or as backup masks should you spot one that you would prefer to use for any reason. The selection was random, so they are just as valid as those found in the mask folder.

3. Statistics

Class Prevalances

The following section gives a break down of the frequency with which classes occur in the dataset. For the segmentation masks, this is stated as an overall percentage of every pixel across all masks. Then, for classification tags, they refer to the fraction of subscenes within the dataset that contain that classification.

Segmentation masks (pixel-wise over the whole dataset):

```
Total marked as CLEAR:       45.955%   |------------------>                    |
Total marked as CLOUD:       52.579%   |---------------------->                |
Total marked as CLOUD_SHADOW: 1.466%   |>                                      |
                                       |———————————————————————————————————————|
                                       0                  50                 100
```

Difficulty:

```
1: 52.632%   |---------------------->                |
2: 6.238%    |-->                                    |
3: 23.782%   |---------->                            |
4: 13.06%    |----->                                 |
5: 4.288%    |->                                     |
             |———————————————————————————————————————|
             0                  50                 100
```

Classification tags:

```
forest/jungle:             17.154%   |------>                                |
snow/ice:                  22.027%   |-------->                              |
agricultural:              13.840%   |----->                                 |
urban/developed:            8.967%   |--->                                   |
coastal:                    4.094%   |->                                     |
hills/mountains:           11.501%   |---->                                  |
desert/barren:             14.035%   |----->                                 |
shrublands/plains:         13.255%   |----->                                 |
wetland/bog/marsh:          2.924%   |->                                     |
open_water:                23.392%   |---------->                            |
enclosed_water:             6.823%   |-->                                    |
thin:                      42.690%   |----------------->                     |
thick:                     61.793%   |------------------------->             |
low:                       53.801%   |---------------------->                |
high:                      46.979%   |------------------>                    |
isolated:                  34.113%   |------------->                         |
extended:                  52.632%   |---------------------->                |
cumulus:                   56.920%   |----------------------->               |
cumulonimbus:              10.916%   |---->                                  |
altocumulus/stratocumulus: 35.088%   |-------------->                        |
cirrus:                    36.647%   |-------------->                        |
haze/fog:                   3.119%   |->                                     |
ice_clouds:                18.519%   |------>                                |
contrails:                  0.780%   |>                                      |
                                     |———————————————————————————————————————|
                                     0                  50                 100
```

4. Errors and Uncertainties

Unfortunately, cloud masking is an inherently ambiguous task. When framed as a binary segmentation task, an annotator must decide how to distinguish between cloud and clear, and differences between their definition and another annotator's is inevitable. We have sought, using our *CALIBRATION* process, to make sure that both annotators were as similar as possible, and that our decisions were as consistent as possible. However, errors still exist throughout the dataset. Some of these can be estimated, whilst others are known but cannot necessarily be accounted for. In this section, we first outline some of the known issues that are not easily quantified, then later, we use some statistical metrics to quantify the level of inter-annotator agreement.

Known issues

Mask boundaries

As discussed in *Appendix B*, we discovered a small bug that affected the border pixels of each mask, making them all clear. To resolve this issue, we have cropped all masks and subscenes to be 1022 pixels across in both dimensions.

Cloud pixels surrounding shadows

We noticed in many scenes that, at the boundary between cloud shadow and clear pixels, a thin line of cloud pixels could sometimes erroneously appear in the Random Forest prediction. We corrected these errors manually when they were spotted during annotation, however, because it affected very small areas, some were missed. It is likely that this occurred because the edge of cloud shadows resembled the darker, shaded areas of clouds to the Random Forest. We considered attempting to use a post-processing technique to remove these artifacts, however we determined that it would likely cause more errors than it would solve. Overall, these are most often a few isolated pixels at the edges of some shadows, and so should not effect an algorithm's training (or test accuracy) too dramatically.

Large, thin cloud boundaries

If thin clouds (often cirrus) appeared over a large section of the image, it was often incredibly difficult to tell exactly where the boundary between clear and cloud lay. Ultimately this is a subjective exercise, and we approached it with the understanding that, if in doubt, mark it as cloud. This is because most end-users of cloud masks desire high cloud recall, though we appreciate it may not suit users who prefer high precision.

High-reflectance icy subscenes

Some subscenes, often over the Antarctic, were essentially completely white in all bands, except the SWIR bands, indicating that the entire subscene was dominated by ice or snow. It was often difficult to tell if this was either clear, high-altitude ice shelves, or smooth icy clouds. We therefore usually marked these with high difficulty (4 or 5), as we were never completely certain that we hadn't marked the entire scene incorrectly.

Classification tag ambiguities

The classification tag scheme was developed to be a quick-and-easy way to understand how an algorithm performs in a given context. They are based on visual cues (described in *Section 2*), rather than physical measurements. As a result, the categories used will not be suitable to every user, and some human errors will more than likely still exist. In particular, **CLOUD TYPE** should be seen as much more of a subjective classification, that delineates clouds by certain visual characteristics, but without any formal meteorological rigour. We recommend that if users have specific needs of their own they should develop their own classification scheme to compliment the one provided here.

Inter-annotator agreement

Now we consider the consistency of annotations between the two annotators. This does not give us a measure of actual accuracy, but rather relative consistency between annotations. We present statistics here for both *CALIBRATION* (10 subscenes) and *VALIDATION* (50 subscenes). See *Appendix B* for details on how these two stages were carried out.

F1-score here refers to the harmonic mean of recall and precision, and is therefore symmetrical with respect to the annotators (the recall of one annotator is the precision of the other, and vice versa). **Accuracy** simply refers to the percentage of pixels that are in agreement between the two annotators.

CALIBRATION DATASET:

```
CLEAR F1-SCORE:             94.478%   |------------------------------------>  |
CLOUD F1-SCORE:             93.149%   |------------------------------------>  |
CLOUD_SHADOW F1-SCORE:      85.024%   |---------------------------------->     |
TOTAL PIXEL-WISE ACCURACY:  93.739%   |------------------------------------>  |
                                      |————————————————————————————————————|
                                      0                  50                 100
```

VALIDATION DATASET:

```
CLEAR F1-SCORE:             94.574%   |------------------------------------>  |
CLOUD F1-SCORE:             95.969%   |-------------------------------------> |
CLOUD_SHADOW F1-SCORE:      77.947%   |------------------------------->        |
TOTAL PIXEL-WISE ACCURACY:  94.978%   |------------------------------------>  |
                                      |————————————————————————————————————|
                                      0                  50                 100
```

We were surprised to find a higher overall agreement for *VALIDATION*, given that we worked on the masks separately, without consultation. However, this may be because the *CALIBRATION* images were hand picked because of their usefulness in demonstrating certain features that may be difficult to annotate consistently. We encourage users to split the dataset such that the test set includes all *CALIBRATION* and *VALIDATION* images, so that an approximate comparison can be made between model performance and inter-annotator agreement.

APPENDIX A: Working with the dataset

The dataset uses numpy arrays to hold the data, therefore the only dependency for reading the data in is to have a python installation with the numpy package. We use python 3 and the latest numpy release (1.19.3) but it is unlikely that any versioning issues will arise when loading the data.

To load a subscene array in python using the numpy package:

```
import numpy as np

subscene = np.load(<path/to/subscene.npy>)
```

To load an associated mask:

```
mask = np.load(<path/to/mask.npy>)
```

For an RGB composite of a subscene one would select:

```
RGB = subscene[...,[3,2,1]]
```

Several python packages are able to read "classification_tags.csv". We recommend the pandas package as it provides lots of functionality for filtering, querying and analysing the data as a DataFrame object (for more information, see here):

```
import pandas as pd

table = pd.read_csv("classification_tags.csv",index_col="index")
```

APPENDIX B: Annotation Strategy

To annotate the images, we first randomly sampled images from the Sentinel-2 2018 archive. By "random", we mean that every product in the 2018 archive had an equal chance of being selected (rather than randomising the location or some other variable). For each selected product, we then extracted random areas of 1152-by-1152 pixels at 20m (note the difference between this and the final 1022-by-1022 subscenes, explained in the next paragraph). This was completed using a virtual machine using the CREODIAS system, which allowed us to quickly query and access the Sentinel-2 data without needing to download it.

Once our subscenes were extracted, we began annotating some test images, and configured IRIS. Several different views were defined, using different band combinations, which could be quickly switched between. It also provided a BingMaps view of the location, which was invaluable when subscenes were highly ambiguous, and gave context for where we were actually looking. During annotation, we padded the subscenes by 64 pixels on each side (resulting in the 1152-by-1152 image used in IRIS), so that there was no disadvantage to annotations close to the edge of the final 1022-by-1022 window. We had originally intended to export 1024-by-1024 pixel masks, however we noticed too late a bug which caused all edge pixels of the mask to be unfilled, hence the final 1022-by-1022 window size.

The Random Forest (RF) could also be configured dynamically. As a general rule, we began annotating a subscene with a simple, small RF that just used the reflectance values. If the model was seen to be struggling, then edge filters could be added as input, and the RF size increased in terms of number of trees, branches and leaves. A suppression filter could also be activated, which smoothed the mask by removing cloudy/shadowy pixels which were not in the vicinity of others. We always checked carefully over the image to check that the RF had done a reasonable job, and often found that significant numbers of iterations and manual corrections were required.

We annotated with high recall in mind. If we felt totally split between annotating an area as cloudy or clear, we defaulted to marking as cloudy. There is no perfect solution to this, as cloud masking is an inherently ambiguous task, but we felt most users of cloud masks desired high recall, and thus our cloud masks should be cautious.

Our annotation style was formalised using 10 hand-picked images (which contained difficult or instructive examples of the kinds of decisions that would need to be made). These 10 images constitute the *CALIBRATION* set, on which we refined our annotations in an attempt to get as high an agreement between us as possible, within a session of a few hours.

Once *CALIBRATION* was completed, we began annotating subscenes individually, served to us randomly by IRIS. This comprised the scenes with the *MAIN* dataset tag, resulting in 453 annotated subscenes. After this, we then decided to both annotate another 50 subscenes in parallel (without consulting one another like we had in *CALIBRATION*). This *VALIDATION* set can then be used to measure the level of inter-annotator agreement, and thus gives users a rough guide of human-level performance.

Once all masks were complete, we then went through every image again and added the classification tags. These were entered into a spreadsheet using a Google Form, in which the product ID was entered, along with checkboxes for each of the tags. Some logical checks were then carried out to reduce human errors in this process (e.g. no image can contain a **CLOUD TYPE**, but not a **CLOUD THICKNESS**, **RELATIVE CLOUD HEIGHT** and **CLOUD EXTENT** tag). Undoubtedly, there will remain some errors in the classification tags, and we encourage users to double-check them if they are especially important in their application.

Contributions & Acknowledgements

The data were collected, annotated, checked, formatted and published by Alistair Francis and John Mrziglod.

Support and advice was provided by Prof. Jan-Peter Muller and Dr. Panagiotis Sidiropoulos, for which we are grateful.

We would like to extend our thanks to Dr. Pierre-Philippe Mathieu and the rest of the team at *ESA PhiLab*, who provided the environment in which this project was conceived, and continued to give technical support throughout.

Finally, we thank the *ESA Network of Resources* for sponsoring this project by providing ICT resources.