KappaSet: Sentinel-2 KappaZeta Cloud and Cloud Shadow Masks

General information

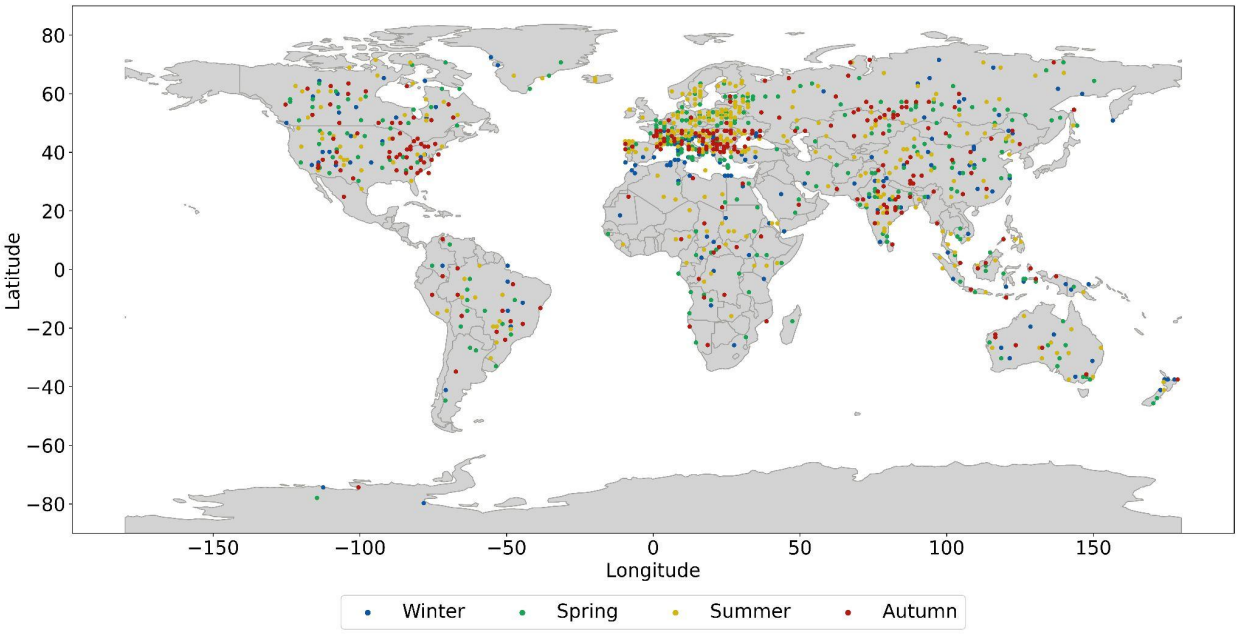
The dataset consists of 9251 labelled sub-tiles from 1038 Sentinel-2 (S2) Level-1C (L1C) products distributed over the globe. In terms of seasonal distribution, S2 products can be divided into the following groups:

- Winter products: 29 austral and 142 boreal S2 products
- Sprint products: 45 austral and 257 boreal S2 products
- Summer products: 30 austral and 293 boreal S2 products
- Autumn products: 29 austral and 213 boreal S2 products

Each S2 product was oversampled at 10 m resolution for 512 x 512 pixels sub-tiles. From each S2 product the most challenging ~5 sub-tiles per product were selected for labelling. Each selected L1C S2 product represents different clouds, such as cumulus, stratus, or cirrus, which are spread over various geographical locations around the world. The classification pixel-wise map consists of the following categories:

- 0 – UNDEFINED: pixels that the labeler is not sure which class they belong to;
- 1 – CLEAR: pixels without clouds or cloud shadows;
- 2 – CLOUD SHADOW: pixels with cloud shadows;
- 3 – SEMI TRANSPARENT CLOUD: pixels with thin clouds through which the land is visible; include cirrus clouds that are on the high cloud level (5-15km).
- 4 – CLOUD: pixels with cloud; include stratus and cumulus clouds that are on the low cloud level (from 0-0.2km to 2km).
- 5 – MISSING: missing or invalid pixels.

The dataset was labelled using Computer Vision Annotation Tool (CVAT) and Segments.ai. With the possibility of integrating an active learning process in Segments.ai, the labelling was performed semi-automatically. The distribution of the dataset is presented in the Figure below. Color represents the season from which the product was chosen.



The dataset limitations must be considered: the data mostly covers terrestrial regions (around 91%) and includes some water areas (around 9%); only around 7% of the dataset contains snow. Current sub-tiles do not have georeferencing.

Dataset Description

kappaset sub-folder contains ***<month>*** and ***test*** directories with **<tile_x_y.nc>** files.

Each **<tile_x_y.nc>** is the NetCDF file that includes the following series of bands: “B01” (443 nm), "B02" (490 nm), "B03" (560 nm), "B04" (665 nm), "B05" (705 nm), "B06" (740 nm), “B07” (783 nm), "B08" (842 nm), "B8A" (865 nm), “B09” (940 nm), “B10” (1375 nm), "B11" (1610 nm), "B12" (2190 nm), “Label”. The filename provides information about sub-tile coordinates which can be extracted by multiplying coordinates by 512 pixels (Figure 2). 512 x 512 pixels NetCDF sub-tiles are generated in tool developed by KappaZeta that is available at: https://github.com/kappazeta/cm-vsm

In addition to abovementioned bands, **<NetCDF.nc>** files in ***test*** also include FMC (Fmask classification map), SS2C (Sinergise S2Cloudless classification map), IDPX (IdePix cloud classification map), MAJAC (CNES MAJA cloud classification map) and SCL(Sen2Cor classification map). Note that MAJAC is available for a very limited number of products, as they should be located in 60° North and 56° South latitudes.

The features are resampled to the same 10 m resolution with Sinc Infinite Impulse Response (IIR) filter that is windowed with a Blackman filter.

Acknowledgements

The data were annotated by Olga Wold, Mariana Rohtsalu, Nikita Murin, Joosep Truupõld, Abdullah Toqeer, Catherine Akinyi Odera and Fariha Harun. The data verification and Software Development were performed by Indrek Sünter, Heido Trofimov, Anton Kostiukhin,

Marharyta Domnich, Mihkel Järveoja, Olga Wold and Tetiana Shtym. The methodology was developed by Kaupo Voormansik, Indrek Sünter, Marharyta Domnich and Tetiana Shtym.

The data were collected, processed, and checked as a part of “KappaMask: AI-based Cloudmask Processor for Sentinel-2” project. We thank Segments.ai team for providing a wonderful annotation tool that was actively used to prepare the dataset. In the end, we thank European Space Agency (ESA) for supporting, advising, and funding the project.

The project was funded by ***European Space Agency***, Contract No. 4000132124/20/I-DT.