ENV

A labelled dataset description for SAR images segmentation

Reference :	CLS_ENV-NT-22-0578
Nomenclature :	[Nomenclature]
Issue :	2.0
Date :	Nov.21, 2022

Distribution:

Amine Benchaabane, Charles Peureux, François Soulat, Gerald Dibarboure (CNES).

	Written by:	*Approved by:*
Names:	A. Benchaabane, C. Peureux	François Soulat
Signature: (Visa or ref.)		[Approver]

Outline

FORM-CR-GB-7-1

1. Table of figures

2. Introduction

Among the various space-borne sensors, Synthetic Aperture Radar (SAR) sensors yield high-resolution imprints of ocean surface and provide significant geophysical parameters for global weather and climate analysis. SAR sensors have several acquisition modes such as the "Wave Mode" (WV), which is dedicated to the measurement of ocean waves in the open ocean. This mode was first introduced on Earth observation mission by the European Space Agency (ESA) for the European Remote Sensing (ERS-1/2) missions (1991-2003) (Kerbaol, et al., 1998). A few years later, the ESA started the Sentinel-1 (S1) mission which is a constellation of two satellites A and B launched in 2014 and 2016 (Torres, et al., 2012). Every month, S-1A and S-1B collect nearly 120,000 WV vignettes with 20 × 20 km ocean surface footprints (~ 5 m of resolution) at two alternate incidence angles of 23.8° (WV1) and 36.8° (WV2). The acquisition can be obtained in VV (Vertical transmitter – Vertical receiver) or HH (Horizontal transmitter – Horizontal receiver) polarization.

Despite the extensive use of learning methods over the past decade, little work has been done on the application of these methods to the detection of ocean features in SAR images that have a significant textural noise, and the various oceanic or atmospheric phenomena over the open ocean seem to be quite difficult to extract from the rest of the image. New learning techniques can be used to effectively address this topic, such as classification and segmentation techniques, where classification aims to assign one or more categories to an image, and segmentation goes further by providing a pixel-by-pixel classification, allowing to delineate the categories in space.

Recently, Wang (Wang, et al., 2019) proposed a first database of SAR images classification called TenGeoP-SARwv for 'Ten Geophysical Phenomena from SAR wave mode', which is provided by IFREMER (L'Institut Français de Recherche pour l'Exploitation de la MER) and publicly available at sea scientific open data publication (SEANOE): http://www.seanoe.org/data/00456/56796/. The studied categories by TenGeoP are pure ocean waves (POW), wind streaks (WS), micro convective cells (MCC), rain cells (RC), biological slicks (BS), sea ice (SI), icebergs (IB), low wind areas (LWA), atmospheric fronts (AF) and oceanic fronts (OF). More details about this database are discussed in section 3 of this document. However, such classification is limited to a single

label per image, whereas several phenomena often coexist. A switch to segmentation has made it possible to overcome this limitation. It is in this context that Colin (Colin, et al., 2022) focused on the development of a SAR image segmentation algorithm based on a deep learning approach. Due to the lack of ground truth for the studied phenomena, a manual annotation of SAR images was necessary to perform the segmentation which is not yet available in TenGeoP database. Colin (Colin, et al., 2022) developed about 110 manually annotated images per category to perform the segmentation, but these annotations do not seem sufficient to capture the signatures of all studied phenomena at different incidence angles. To improve such segmentation, additional work was done by a CLS team to extend the labeling of the TenGeoP database; an estimation of the accuracy of the existing supervised segmentation shows that we can expect to increase its performance by about 30% by doubling the number of annotated images. This newly labelled database can be useful to a wide range of users and communities in the fields of deep learning, remote sensing, oceanography and meteorology. Such a data set would be useful for the finer determination of the radar signature of such phenomena, and to estimate their global cartography. The methodology used to extend the labeled dataset is described in Section 4.

3. TenGeoP-SARwv description

TenGeoP is a subset of more than 37,553 SAR images selected from Sentinel-1A WV acquisition (VV polarization) in 2016. Among the ten categories, only one geophysical phenomenon was assigned to each image of the database (Figure 1). The definition and properties of each phenomenon are well described by Wang (Wang, et al., 2019).
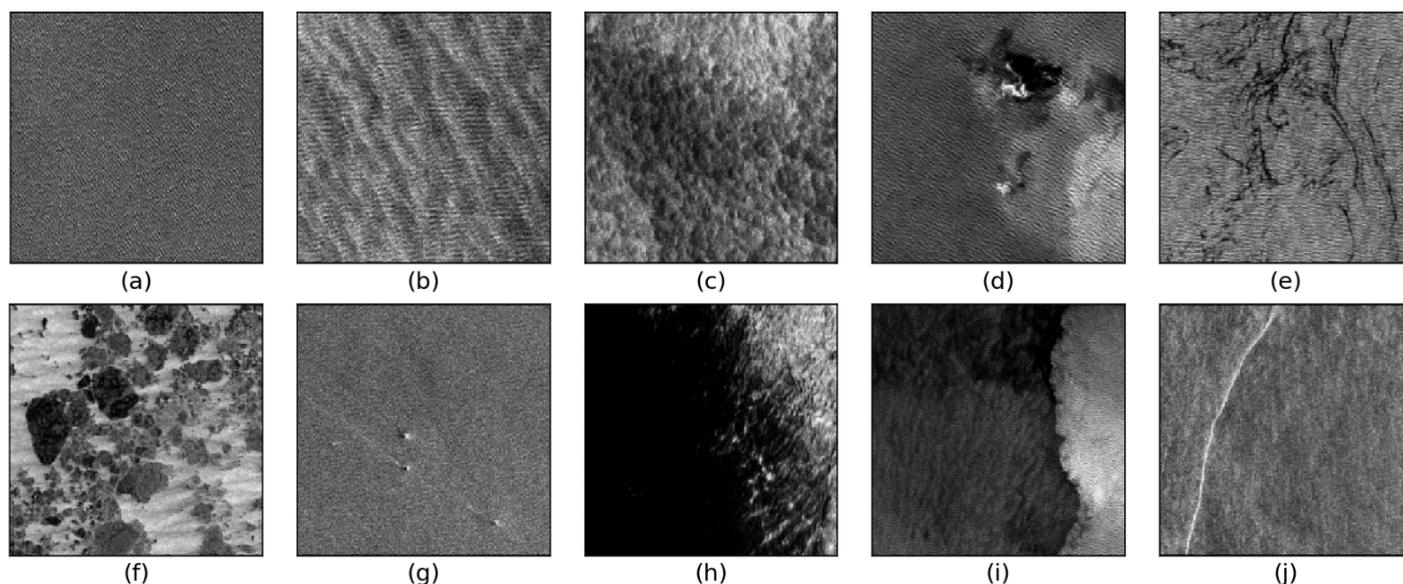
Figure 1 : From (a) to (j) are image examples of pure ocean waves, wind streaks, micro convective cells, rain cells, biological slicks, sea ice, icebergs, low wind area, atmospheric front and oceanic front.

The phenomena selected by TenGeoP are commonly observed by the S-1 SAR, but the list is not exhaustive; the signatures of other phenomena could also be observed by the SAR such as internal waves (Jia, et al., 2018). Note that, at present, there is no WV acquisition in the Artic Ocean and closed seas like the Mediterranean Sea.

FORM-CR-GB-7-1

As provided by Wang (Wang, et al., 2019), all TenGeoP images were transformed into a quick look image from the S-1 WV Sigle Look Complex (SLC) products (Torres, et al., 2012). The raw S-1 WV images have a resolution of 5 m (more than 4000 pixels in both range and azimuth directions), which is not necessary for large-scale geophysical phenomena detection (~100 m to 5 km). For this reason, TenGeoP SAR images were downscaled from 5 to 50 m. Such processing may also reduce the speckle noise. The final processing step applied to the TenGeoP database was machine learning normalization to set a common minimum and maximum values for each image of the entire database. Then, the images were provided in PNG format. More details about each step of processing are sufficiently presented by Wang (Wang, et al., 2019) and further information are also available from: http://www.seanoe.org/data/00456/56796/ such as the data distribution tree:

> /F: Pure Ocean Waves
> /G: Wind Streaks
> /H: Micro Convective Cells
> /I: Rain Cells
> /J: Biological Slicks
> /K: Sea Ice
> /L: Iceberg
> /M: Low Wind Area
> /N: Atmospheric Front
> /O: Oceanic Front

The TenGeoP database is not perfectly balanced because, as described by Wang (Wang, et al., 2019), the labeling of the database was done manually by visual inspection with two essential criteria: either only one phenomenon is visible in the image, or its patterns are clearly visible by the human eye. Similarly, seasonality plays a crucial role in the appearance of phenomena and their spatial distribution. For example, a few IB images were found in May to October due to the iceberg seasonality.

4. Methodology of labelling

Since the annotation is performed manually, it must be done by SAR imaging expert analysts with such guidelines in order to homogenize their sensitivity to the perception of the studied phenomena. Therefore, a set of criteria on the definition of phenomena has been established and taken from the categories analysis discussed by Wang (Wang, et al., 2019). In the present note, details are added on the contouring of those phenomena. To better ensure the coherence of the annotations, a cross-review of the annotations was carried out in order to discuss and correct possible outliers or difficult cases, because depending on the texture complexity of the SAR images, the selection of a label is sometimes difficult.

The flow chart on Figure 2 is repeated over the TenGeoP database until 50 images per acquisition mode (WV1/WV2) have been successfully segmented by label among F to O. Empirically, a slightly different procedure is applied in the case of sea-ice TenGeoP labels.
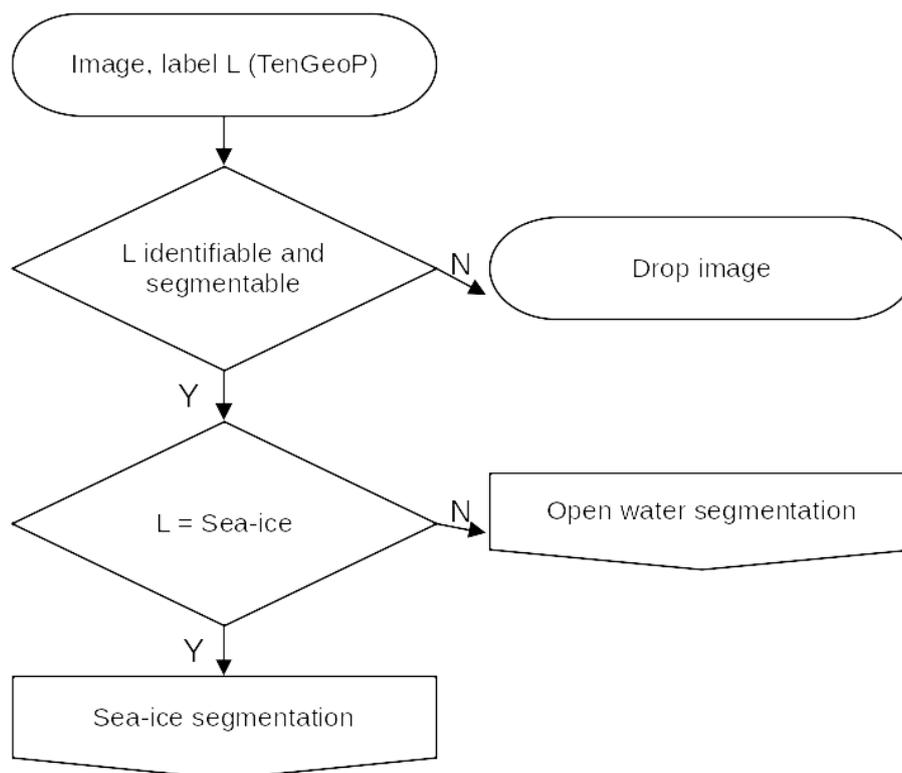
FORM-CR-GB-7-1

Figure 2 : Segmentation process flowchart.

4.1. General principles

The segmentation is provided as polygons in the original image pixel coordinate system. Each pixel was affected by at most one label among the 10 available from TenGeoP. Therefore, particular attention was paid to the fact that those polygons do not intersect. Particularly, this requires that some phenomena included in larger ones (for example icebergs or rain cells) are extruded.

Segmentations do not contradict the labels in TenGeoP. In each provided segmented image, the TenGeoP label results in at least 1 polygon.

Particular attention was dedicated to the quality of the segmentations: polygons are provided only when the analyst is confident in his results. For this reason, some parts of the images are not segmented. In addition, only images with good readability and well-defined demarcations are selected, although this could result in a selection bias.

A distinction can be made between the various categories based on their typical spatial extent. The so-called global phenomena have dimensions of the order are larger than the SAR image (20 by 20 kms approximately). This is particularly the case for Pure Ocean Waves, Wind Streaks, Micro Convective Cells and Sea Ice. On the other hand, local phenomena cannot in general fill the whole SAR image: Rain Cells, Icebergs, Atmospheric and Oceanic Fronts. Biological Slicks and Low Wind Areas, although closer to global phenomena often coexist with another background. In the TenGeoP database, a vast majority of images labelled as global phenomena exhibit this single phenomenon over the whole image. These segmentations bring lower improvements compared to

the TenGeoP single label. For this reason, as much as possible, multiple label images were favored in our segmentation database.

Another specificity of local phenomena lies in their contouring. Phenomena such as fronts and icebergs might exhibit more defined borders in comparison to global phenomena for example, but for practical reasons, a margin is kept around those borders in the segmentation polygons to ensure a smooth transition with other coexisting classes

4.2. Open water segmentation

The procedure is presented in Figure 3, and sub-processes represented as rectangle with double struck vertical edges are detailed below.

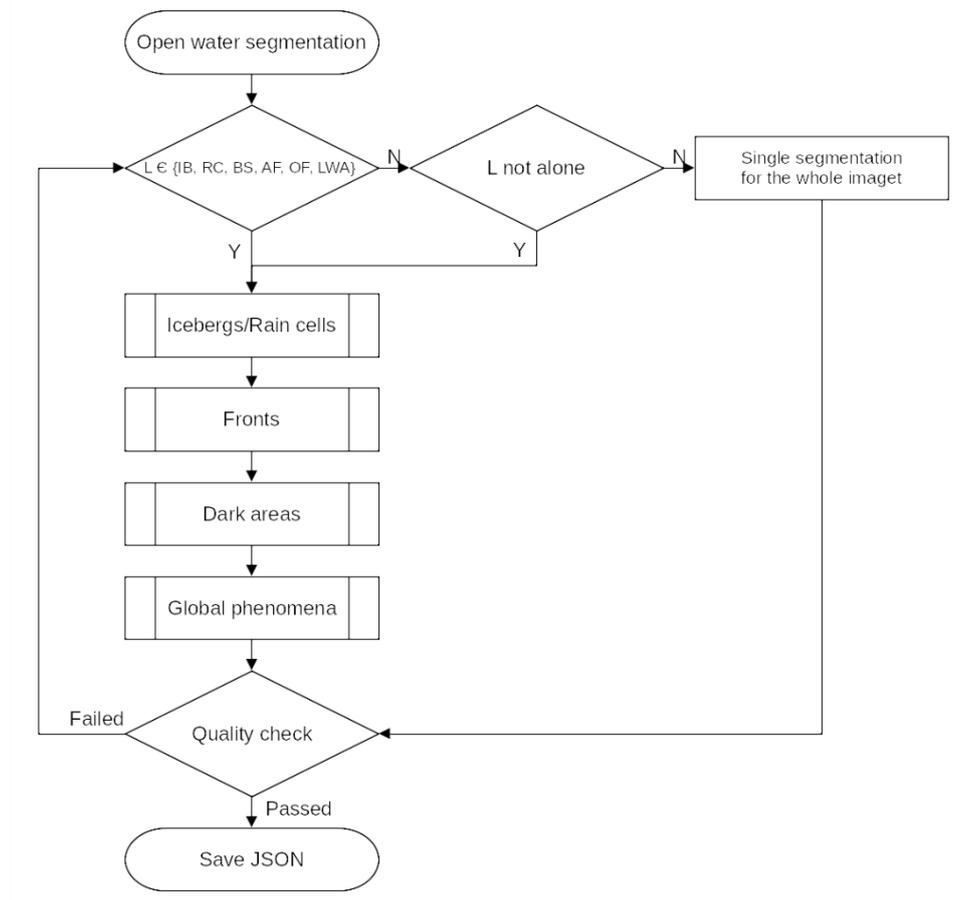Figure 3 : Segmentation process flowchart over all TenGeoP labels except sea-ice.

4.2.1. Icebergs and Rain cells

These two phenomena are processed together although almost never present simultaneously as they both can exhibit bright pixels. On top of bright signatures, icebergs often exhibit shadows and wakes. Their shadows are considered when present as part of the iceberg contour while their

FORM-CR-GB-7-1

wakes are not. Rain cells signature can also consist of dark areas. Rain cells polygons include those dark areas to the appreciation of the analyst.

4.2.2. Fronts

Atmospheric fronts and oceanic fronts are characterized by segments of local NRCS discontinuity. An atmospheric front appears as a strong NRCS gradient between two areas with heterogeneous NRCS, one bright and the other dark. On the other hand, oceanic fronts appear either as bright or dark segments over relatively homogeneous backgrounds.

4.2.3. Dark areas

Under this term fall often coexisting phenomena which are low wind areas and biological slicks. Indeed, slicks appear only in low wind conditions. A low wind area is considered when the NRCS background becomes so low that potential ocean waves modulations disappear.

Biological slicks segmented in this work appear exclusively as filaments. Particular attention was paid here to obtain accurate contouring as possible of those filaments. Those contouring may have applications in determining their dominant direction. These filaments may well mix and blend, which can lead to confusion with areas of low wind. Discrimination here is again left to the discretion of the analyst.

4.2.4. Global phenomena

Under this category fall MCC, WS and POW. All are global phenomena with poorly defined borders, which is the reason why they are segmented at last. All of them result from the small-scale interaction of the surface roughness with the wind. MCC results in circular cells, while wind streaks result in parallel rolls, with wave lengths much greater than the typical swell NRCS modulations. There exists a continuum of real-world situations in between those two situations, which can make it quite complicated to discriminate. For application purposes, wind streaks are segmented when a dominant direction can be identified without ambiguity. This information can then be useful in a SAR wind processor to estimate the wind direction. If neither of these two structures is detected and if swell NRCS modulations are visible, the Pure Ocean Waves label is used. A variable NRCS background is allowed in the POW category if the modulations remain homogeneous.

4.2.5. Quality check

After segmentations, this process consists in a series of checks:

- Polygons do not intersect.
- SAR images edges are not segmented. The edges of SAR images often show black bands. This is related to the calibration and processing of SAR products. Therefore, these areas have not been annotated in order not to pollute the annotations with these edge effects (Figure 4).
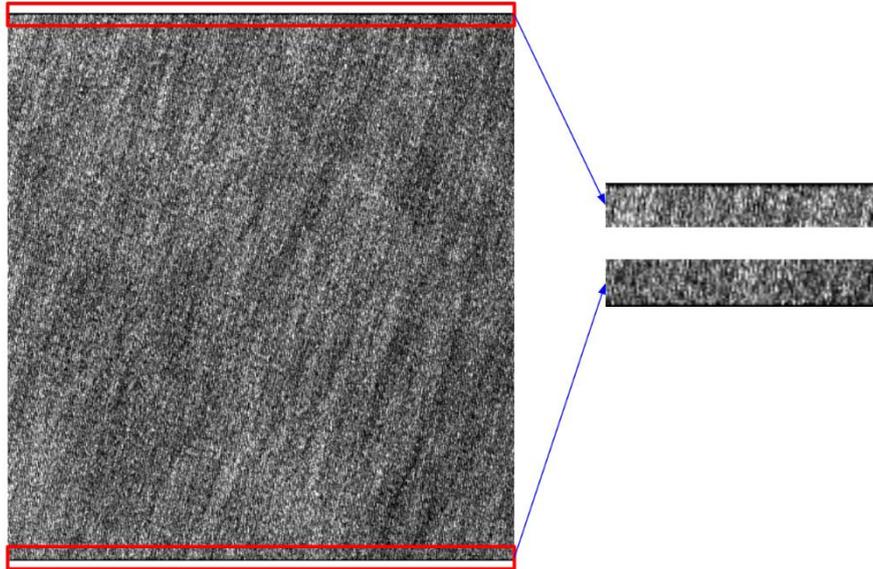- Review by another analyst.

FORM-CR-GB-7-1

Figure 4 : SAR image (s1a-wv1-slc-vv-20160104t013058-20160104t013101-009337-00d7fd-075) illustrating the presence of the black bands on the edges.

Only once these checks are all passed, the image JSON segmentation is added to the present database.

4.3. Segmentation sub-process over sea-ice TenGeoP labels

Sea-ice labeled images appear as a particular category, as only a limited set of phenomena may coexist on them. Most of the sea-ice TenGeoP images only exhibit sea-ice over the whole image. Also, sea-ice areas are never found on other labels TenGeoP images, contrary to all other labels. This might be a problem in providing a coherent SAR images segmentation over both open water and sea-ice as such a model would benefit from the coexistence of sea-ice and open water in the learning database. For this reason, particularly over the sea-ice TenGeoP labelled images, a particular emphasis was put on including multi-label segmented images. Such a process is described on the flowchart of Figure 5.
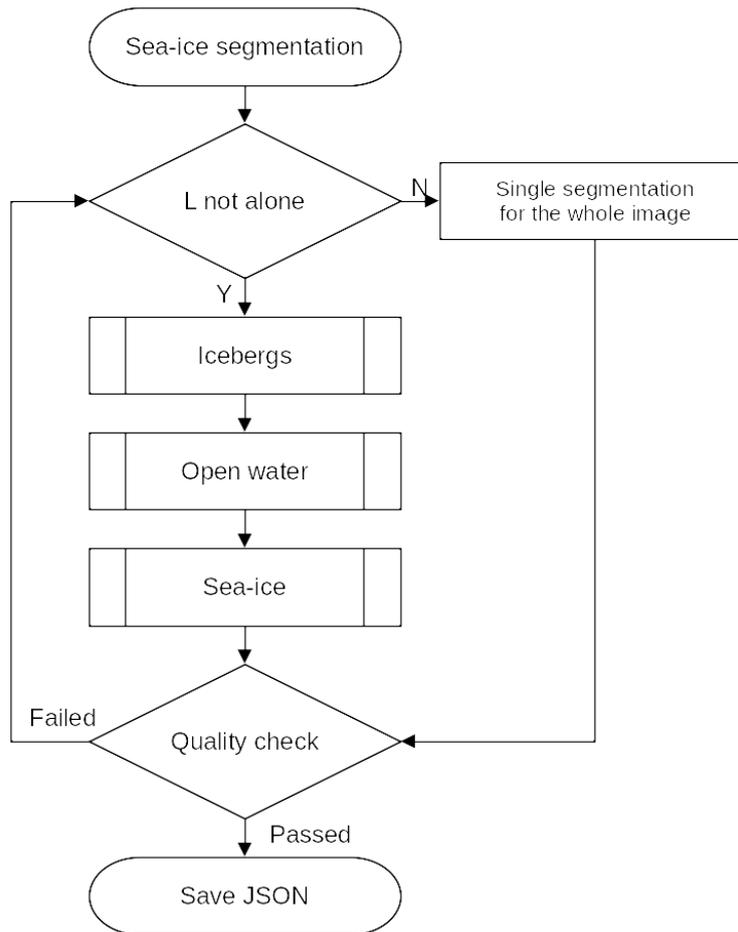
Figure 5 : Sea-ice segmentation sub-process flowchart.

4.3.1. Icebergs

Icebergs are segmented similarly as in the open water case. In the case of breaking sea-ice, icebergs may be segmented if they drifted a typically larger distance than neighboring ice floes. Also, some icebergs trapped in sea-ice could be identified and were segmented as such, although surrounded by sea-ice and not open water.

4.3.2. Open water

Open water often appears as dark homogeneous areas in breaking sea-ice areas. They are often segmented as Low Wind Areas, due to the manifest absence of NRCS modulations, and more rarely as Pure Ocean Waves. Sea-ice leads were not considered as open water in this work.

FORM-CR-GB-7-1

4.3.3. Sea ice

Sea-ice is segmented at last. Its signature on NRCS is very diverse.

4.4. Annotations tool and exploitation

The technical tool that was used to make the annotation is Labelme. Labelme is an open-source annotation tool, available on https://github.com/wkentaro/labelme. It was written in Python to support manual image polygonal annotation for object detection, classification, and segmentation. The tool allows to create various shapes, including polygons, circles, rectangles, lines, line strips, and points. The resulting label can be saved as JavaScript Object Notation (JSON) files directly from the application. Each JSON file includes the labeled pixels, the SAR quicklook used for annotation and quality flag if filled. By default, the name of the JSON file corresponds to the source image of TenGeoP.

For each TenGeoP class phenomenon, 50 SAR images were labeled by acquisition mode (WV1 and WV2), which gives us 100 images annotated by class, all incidences combined. knowing that the accuracy of the annotation is in the order of 2 to 3 pixels (at TenGeoP resolution). Since the accuracy of the annotation is not at the pixel scale, the boundaries of the segmented classes on the same image may not match perfectly. In other words, there may be unsegmented pixels between classes, and they will be systematically labeled as background by Labelme.

A Jupyter notebook in support of this document allows to process JSON files and extract labels in array format. They are mainly based on the Labelme resources, more details are inserted in the Jupyter notebook. The processing of JSON files will generate the following tree (Figure 6) for each annotated JSON file, where ***.png** is the input TenGeoP image and ***.npy** is a three-dimensional labeled array with a shape of 3 (*number of pixels in y axis, number of pixels in x axis, 10*); 10 represents the labeled layers of 10 phenomena.

```
|--<sar_wv_product_name>
         |
         |----<sar_wv_product_name>.npy
         |----<sar_wv_product_name>.png
         |----labels_vizualisation.png
```

Figure 6 : Products generated when processing a JSON annotated file.

A quick look labels visualization (**labels_vizualisation.png**) is generated in the same directory (the legend corresponds to the definition of categories as defined in the previous section). Some examples of "labels_vizualisations.png" are illustrated in Figure 7.
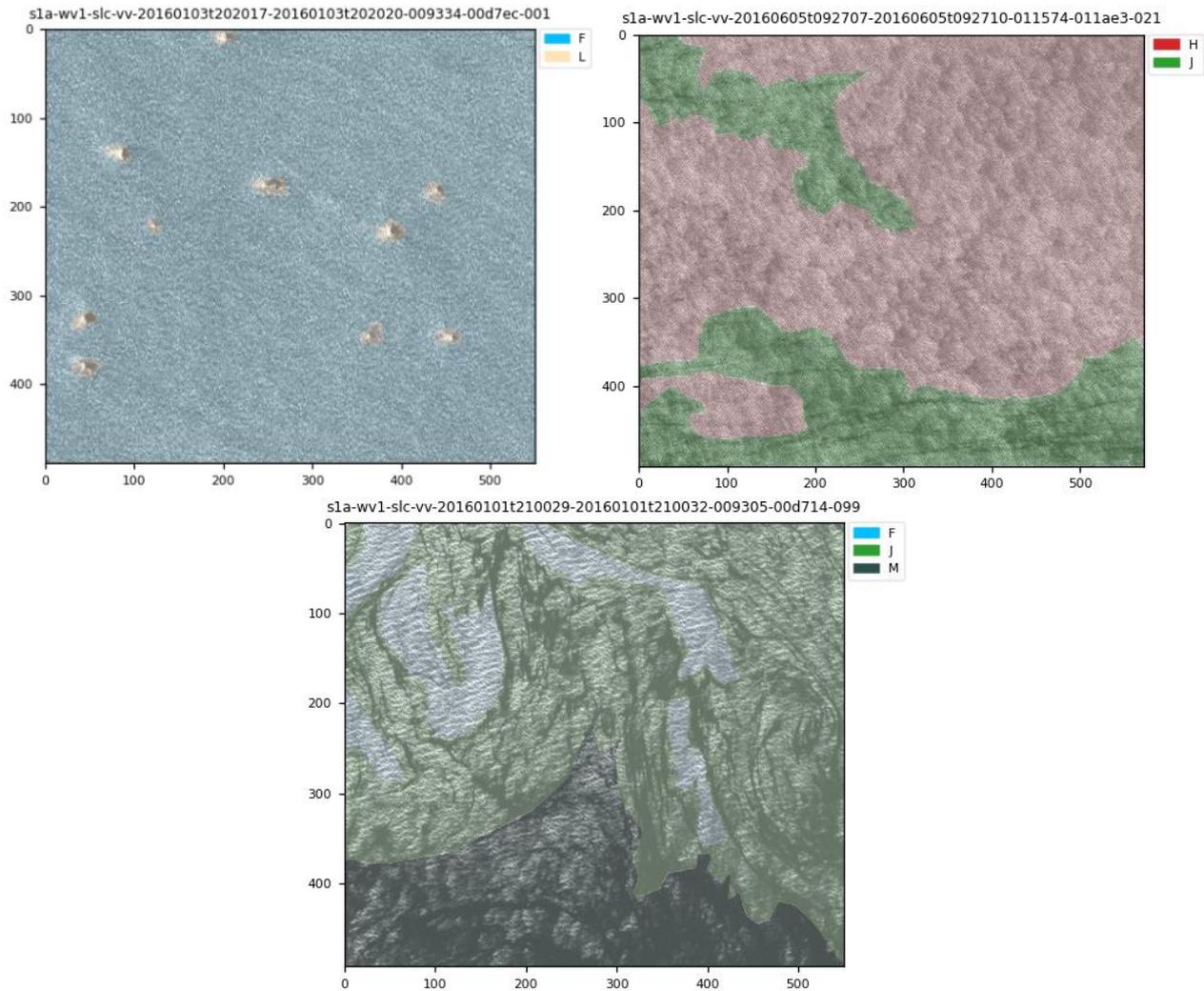
Figure 7 : Manually annotated SAR images from TenGeoP database. In the background, the surface roughness is displayed, and the colors correspond to the different classes. To bear in mind that image in the left side was classified as Iceberg, the middle was classified as Micro convective cells and the one in the right side as biological slicks in TenGeoP database. The titles correspond to the image attributes.

One should know that the size of the SAR images in the TenGeoP database is not constant. Depending on the database exploitation, a cropping of the images and their labels could be applied to homogenize the data. When using this dataset in a learning algorithm, it is strongly recommended not to resize the images and their labels and to prefer cropping. Such a processing avoids the modification of labels even if the precision of the annotations is not at the pixel level.

5. Conclusion

The presented database allows for the training of Sentinel-1 SAR image segmentation. Up to now, state-of-the-art learning-based inferences were based on classifications, i.e., a single label per image. An illustration of this data base is presented on Figure 8   Manual annotation is time consuming, especially when an emphasis is put, as is intended here, on its quality. For this reason,

such an effort should be pursued in the future. This work represents about a month of full-time analysis work, providing a baseline for future effort.
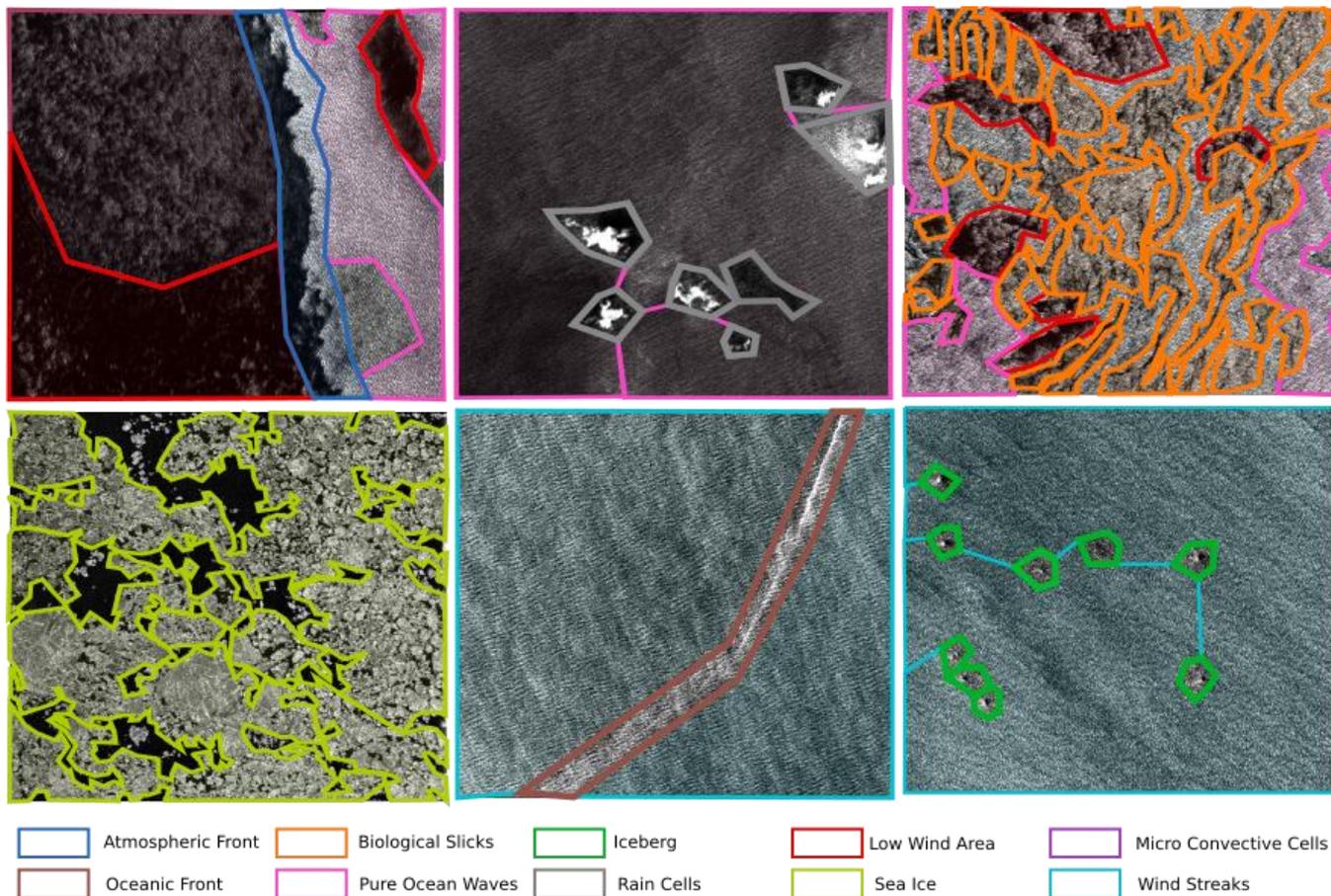

Figure 8 : Samples taken from manually annotated SAR data set. First row, from let to right: Atmospheric Front, Rain Cells and Biological Slicks Sentinel 1 Wave Mode images TenGeoP labels. Second row: Sea Ice, Oceanic Front and Icebergs.

6. References

1. Colin, A. et al., 2022. Semantic Segmentation of Metoceanic Processes Using SAR observations and Deep Learning. *Remote sensing*.
2. Jia, T., Liang, J., Li, X. & Sha, J., 2018. SAR observation and numerical simulation of internal solitray wave refraction and reconnection behind the Dongsha Atoll. *Journal of Geophysical Research*.
3. Kerbaol, V., Chapron, B. & Vachon, P., 1998. Analysis of ERS-1/2 synthetic aperture radar wave mode imagattes. *Journal of Geophysical Research*.
4. Torres, R. et al., 2012. GMES Sentinel-1 mission. *Remote Sensing of Environment*.
5. Wang, C., Mouche, A. & Tandeo, P., 2019. A labelled ocean SAR imagery dataset of ten geophysical phenomena from Sentinel-1 wave mode. *Geosci Data J*.

FORM-CR-GB-7-1