MULTI-MODAL DEEP LEARNING WITH SENTINEL-3 OBSERVATIONS FOR THE DETECTION OF OCEANIC INTERNAL WAVES

L. Drees[1,*], J. Kusche[1], R. Roscher[1,2]

[1] IGG, University of Bonn, Germany - ldrees@uni-bonn.de, kusche@geod.uni-bonn.de, ribana.roscher@uni-bonn.de
[2] Institute of Computer Science, University of Osnabrueck, Germany

KEY WORDS: multi-modal, deep learning, internal waves, sentinel-3, multi-stream, late fusion, neural network

ABSTRACT:

The observation of waves that propagate along density interfaces inside the ocean poses a significant challenge, as their visible surface signatures are much lower compared to their internal amplitudes. However, monitoring internal waves is important as they redistribute large amounts of energy, play a role in mixing and vertical heat transfer, and modify water and nutrient transports. Although satellite observations would allow global monitoring of internal waves at constant time intervals, their automatic detection is challenging: In optical images, internal waves are hardly visible and can be obscured by clouds, whereas radar data have limitations in coastal regions and their spatial coverage is not perfect. Furthermore, the occurrence of internal waves can be confused with other ocean phenomena. In this work, we present an automated detection framework for internal waves based on multiple data sources in order to compensate for the shortcoming of single data sources. In our application, we use Ocean and Land Color Imager and Synthetic Aperture Radar Altimeter data. Our contributions are (1) we develop a multi-modal deep neural network `SONet` with multi-streams and late fusion, which performs a classification on the basis of training with both modalities, and (2) we establish a method to deal with missing modalities. Experiments in the Amazon Shelf region show `SONet` achieves adequate results when both modalities are available, but also when only a single modality is available. By exploiting correlations between the modalities, `SONet` classifies OLCI images off the SRAL ground track better than uni-modal network `ONet`, which describes a great advantage of our multi-modal network.

1. MOTIVATION

We witness a growth in the number of satellites with integrated sensors characterized by various spatial, spectral and temporal resolutions. It is therefore common in remote sensing that the same scene is observed simultaneously with different sensors and therefore multi-modal data is available for a joint analysis. Compared to data from individual sensors, the different modalities usually have certain properties and characteristics that can be exploited for a better understanding of the scene. Especially for satellite missions, where many research questions from different scientific fields are addressed simultaneously, multi-modal data is common. For the Sentinel-3 mission, for example, Ocean and Land Color Imager (OLCI) and Synthetic Aperture Radar Altimeter (SRAL) are mounted on the same satellite such that radar signals and optical images are acquired simultaneously in intersecting observation areas. We present a framework which combines SRAL and OLCI observations for an automatic detection of oceanic internal waves (IWs), as illustrated in Figure 1. Oceanic IWs are gravity waves at internal density layers of the water. Compared to surface waves, they are significantly larger, both in amplitude (up to 200 meters) and in wavelength (up to multiple kilometers). For example, IWs play a key role in understanding the interaction of large-scale tides and smaller scale turbulences (Jackson et al., 2012). However, the detection of IWs is a challenge because they are hardly visible optically on the sea surface and active sensors like SRAL with a better detection rate do not observe the whole area. With this work we show that it is worthwhile to combine both modalities SRAL and OLCI in one model for the detection of IWs. Although the data set we have created is currently still specific to our study site, our experiments can already show that the use of both modalities leads to an increase in accuracy compared to uni-modal OLCI-based methods.

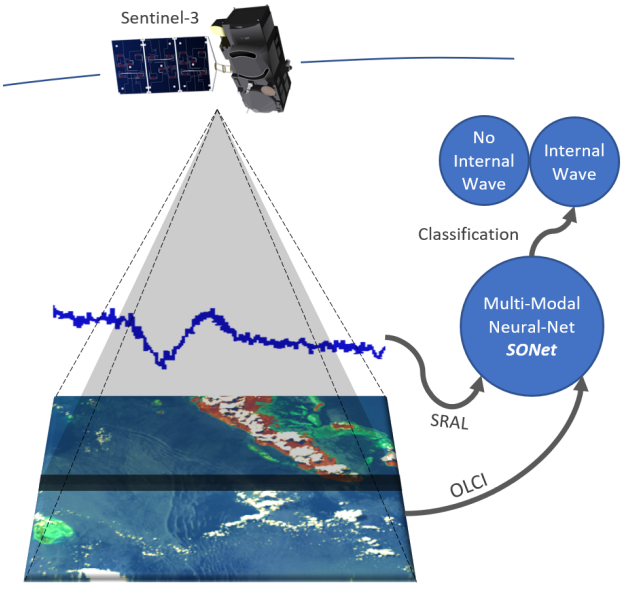


Figure 1. Overview of the approach: Multi-modal data SRAL and OLCI simultaneously acquired by Sentinel-3 are classified by the multi-modal neural network `SONet`.

The paper is structured as follows: In Sec. 2 we present current research on multi-modal learning in earth sciences and explain the connections to this work. Besides a brief insight into the investigation of IWs is given. Sec. 3 introduces the study site and our used multi-modal data set. The multi-modal deep learn-

* Corresponding author

ing framework and how it process the data set is presented in Sec. 4. In Sec. 5 we show a concrete implementation of the multi-stream procedure with late fusion and discuss the results afterwards.

2. STATE OF THE ART

In this section we first discuss multi-modal approaches using earth observation data. Afterwards we present the current state of research on IWs.

2.1 Multi-modal deep learning in earth sciences

In general, the potential of machine learning methods and especially of approaches of deep learning in remote sensing is large and includes classification and regression tasks as well as state prediction tasks such as now-casting of precipitation, seasonal forecasts and modelling of global mass transport (Reichstein et al., 2019, Ma et al., 2019). Related to this is the promising research area of multi-modal learning. Apart from conventional data fusion methods (Lahat et al., 2015, Gupta, Cheng, 2006) in the field of Earth sciences, so far, only a few multi-modal deep learning approaches exist. One multi-modal multi-source approach is used for underwater mapping of the seabed. In order to analyze habitats for marine ecology, few visual images of autonomous underwater vehicles and a multitude of bathymetric data from ships are used to perform classification tasks (Rao et al., 2014). Another related approach is the multi-source classification of cloud, shadow and land cover scenes with data from different satellite missions (Shendryk et al., 2019). But in contrast to our method, all input data (PlanetScope and Sentinel-2 imagery) are of optical nature. There are also temporal multi-modal networks with the aim to learn a common representation of the data, which have both different modalities and are variable in time (Yang et al., 2017). Similar to our architecture is the multi-stream approach for temporal Sentinel-2 data to generate land cover classes from VHSR images and time series with high spatial resolution (Benedetti et al., 2018). Apart from Earth Sciences, multi-modal deep learning is already more widespread, including audio-visual speech recognition (Mroueh et al., 2015), scene alignment (Aytar et al., 2017b, Aytar et al., 2017a), image captioning (Srivastava, Salakhutdinov, 2012) and video hyperlinking (Vukotić et al., 2016).

2.2 Investigation of internal waves

For many years, it has been an effort to detect oceanic IWs using remote sensing methods in order to determine their parameters and energy, to get information about the stratification of the water and the mixed layer depth, or to investigate their influence on tidal (or baroclinic) currents (Klemas, 2012). IWs are caused by external forces acting on stratified water levels, e.g. wind stress or tides over a region of bottom topography (Alpers et al., 2008, Robinson, 2010, Magalhães et al., 2016, Zhao et al., 2004). Although there are maps showing IWs hotspots in all areas of the oceans (Jackson et al., 2012), a global automatic detection procedure does not exist yet. This would be beneficial also for other satellite missions, like SWOT, which are interested in mesoscale processes like geostrophic velocities. These have an even greater amplitude than IWs, but since the phenomena overlap, it will be difficult to determine geostrophic velocities at scales smaller than about 70 km without knowledge of IWs locations (Qiu et al., 2017). In many cases, the sea surface elevations of IWs are too small to be visible, but rather

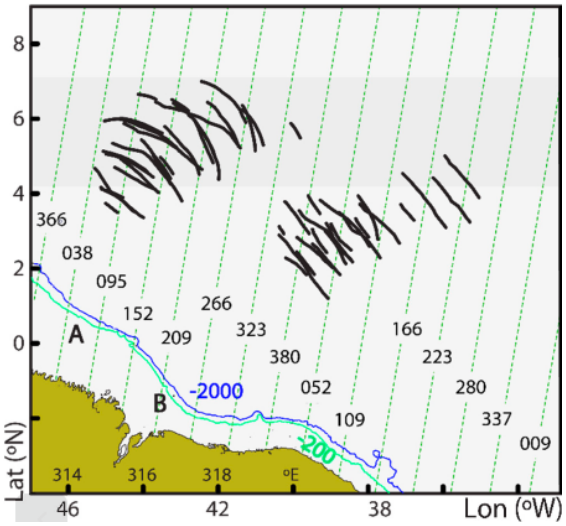


Figure 2. Map of the study region off the Amazon Shelf taken from (Santos-Ferreira et al., 2019) with ground tracks of 15 RO of Sentinel-3A (green dashed lines). Black bold curves indicate location of IWs crest observations by (Magalhães et al., 2016). The presumed IWs-origins are marked by labels A and B.

the currents on the surface. In Synthetic Aperture Radar (SAR) images as well as in optical images like OLCI and Moderate-resolution Imaging Spectroradiometer (MODIS), IWs can be recognized as alternating stripes on the water surface (Fig. 5a). These stripes go back to a sharp change from rough to smooth sea surface. They appear as wave packets or solitary events with large amplitude - then called internal solitary wave (ISW) or internal soliton (Jackson, 2007, Alpers et al., 2008, Ikeda, 1995). IWs are also observable in altimeter signals (Fig. 5b). SRAL of Sentinel-3 provides parameters in which a certain pattern of peaks indicates IWs. While the change of significant wave height (SWH) and sea level anomaly (SLA) is often low, peaks in the radar backscatter coefficient (σ_0^{Ku}) as well as in the differenced-mean-square slope ($\delta\bar{s}_n^2$) is crucial, since radar backscatter allows a good estimate in sea surface roughness (Santos-Ferreira et al., 2018).

3. DATA

3.1 Study site

We focus on an area in the Atlantic Ocean off the Amazon Shelf, which is known for large amplitude ISW (Magalhães et al., 2016, Santos-Ferreira et al., 2019). Spatially, we concentrate on certain relative orbits (RO) of the Sentinel-3 mission, namely RO 38, 95, 152 and 209 in the western part and RO 380, 52, 109 and 166 in the eastern part (Fig. 2). In the period from April 2017 to August 2019 we collected and annotated a total amount of 2373 data samples on these orbits. Also Sentinel-3B, which is equipped with the same sensors, has been contributing data since January 2019. However, Sentinel-3B flies 140 degree out of phase compared to Sentinel-3A, so the ground tracks are offset in such a way that the spatial coverage is approximately doubled.

3.2 Multi-modal dataset with lack of modalities

We use OLCI Level-1b-EFR top-of-atmosphere (TOA) radiometric full resolution image data with 21 bands, which

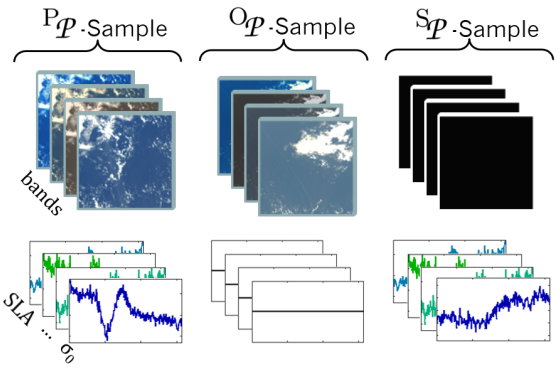


Figure 3. The dataset consists of three subsets. One sample from each subset is illustrated. The sample of subset $^{P}\mathcal{P}$ contains both modalities. There is a lack of modality in subsets $^{O}\mathcal{P}$ and $^{S}\mathcal{P}$ because SRAL resp. OLCI data is missing. They are replaced by zero matrices.

stem from the ESA-Copernicus Open Access Hub (https://scihub.copernicus.eu/dhus/#/home). As complementary modality, we use the Level 2 Sentinel-3 SRAL Water data product "SRAL Altimetry Global in NTC", which is provided by EUMETSAT (https://archive.eumetsat.int/usc/). An original OLCI image has a size of 4091 px (along-track) and 4865 px (cross-track; corresponds to a swath width of 1270 km), the original SRAL track includes the entire RO.

Since the originally provided data are too large to be processed by our multi-modal network, we extract image patches $\boldsymbol{O} = [351 \times 351 \times 21]$ as well as SRAL tracks $\boldsymbol{S} = [313 \times 4]$ (due to 313 observations per parameter SWH, SLA, σ_0^{Ku}, $\delta\bar{s}_n^2$) which are still georeferenced (Fig. 4). Thus, our dataset consists of three subsets $^{O}\mathcal{P}$, $^{S}\mathcal{P}$, and $^{P}\mathcal{P}$. Subset $^{O}\mathcal{P}$ consists only of optical data, subset $^{S}\mathcal{P}$ only of radar signals, and $^{P}\mathcal{P}$ contains both, that means multi-modal data (Fig. 3). In case that not both modalities are available for a sample we call this "lack of modality". Therefore the subsets $^{S}\mathcal{P}$ and $^{O}\mathcal{P}$ suffer from a lack of modality because the OLCI resp. SRAL modality is missing there. Just subset $^{P}\mathcal{P}$ is a multi-modal data set without lack of modalities. All samples of the subsets are referenced in classes `IW` and `NoIW`.

3.2.1 Subset $^{O}\mathcal{P}$ It consists of ^{O}N OLCI samples $^{O}\boldsymbol{O} = [^{O}\boldsymbol{O}_1,\ldots,^{O}\boldsymbol{O}_{^{O}N}]$ with labels $^{O}\boldsymbol{y} = [^{O}\boldsymbol{y}_1,\ldots,^{O}\boldsymbol{y}_{^{O}N}]^{\mathsf{T}}$. Images in this subset are mostly taken off the SRAL ground track. Multi-modal recorded data, where the SRAL signal is disturbed by land influences and therefore unusable, are also in this subset. The same applies to images where waves appear at the lateral edge of the image and are therefore not in the field of view of the SRAL sensor. The images show considerable variation in wave forms and sizes, brightness and cloud cover, which makes classification challenging.

3.2.2 Subset $^{S}\mathcal{P}$ Radar data lies in this subset if the underlying OLCI image is covered by clouds and therefore not applicable. A total number of ^{S}N SRAL samples $^{S}\boldsymbol{S} = [^{S}\boldsymbol{S}_1,\ldots,^{S}\boldsymbol{S}_{^{S}N}]$ with reference vector $^{S}\boldsymbol{y} = [^{S}\boldsymbol{y}_1,\ldots,^{S}\boldsymbol{y}_{^{S}N}]^{\mathsf{T}}$ is available. We focus on the parameters SWH, SLA and σ_0^{Ku} from Ku-band in 20 Hz resolution and use them in their original state. Additionally we compute the $\delta\bar{s}_n^2$ from the σ_0^{Ku} and σ_0^{C} as presented in (Santos-Ferreira et al., 2019).

3.2.3 Subset $^{P}\mathcal{P}$ This subset contains both modalities SRAL $^{P}\boldsymbol{S} = [^{P}\boldsymbol{S}_1,\ldots,^{P}\boldsymbol{S}_{^{P}N}]$ and OLCI $^{P}\boldsymbol{O} = [^{P}\boldsymbol{O}_1,\ldots,^{P}\boldsymbol{O}_{^{P}N}]$ with common label vector $^{P}\boldsymbol{y} = [^{P}\boldsymbol{y}_1,\ldots,^{P}\boldsymbol{y}_{^{P}N}]^{\mathsf{T}}$, where ^{P}N is the total number of samples in this subset. It concerns image patches where the SRAL signal is not corrupted by coastal topography and the OLCI image is covered by less than 25 % of clouds. In any case, the point where SRAL indicates an IW must be visible in the OLCI image. Spatial and temporal synchronization of the modalities is achieved by ensuring SRAL ground track is in all samples exactly centered over the OLCI patch and ends at the boundaries of the image. This ensures a strong alignment between the data.

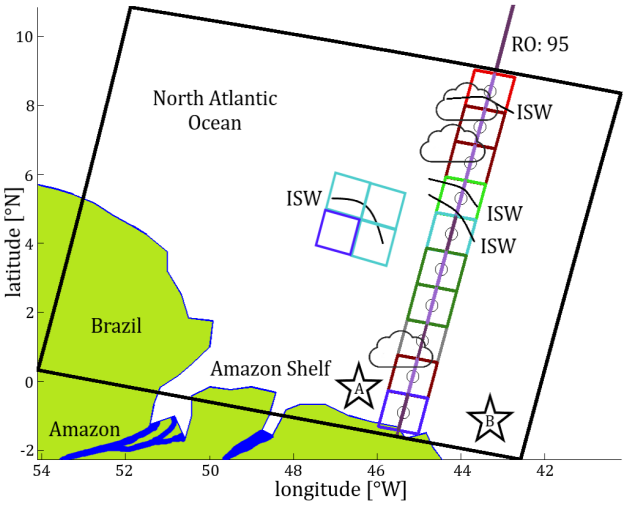
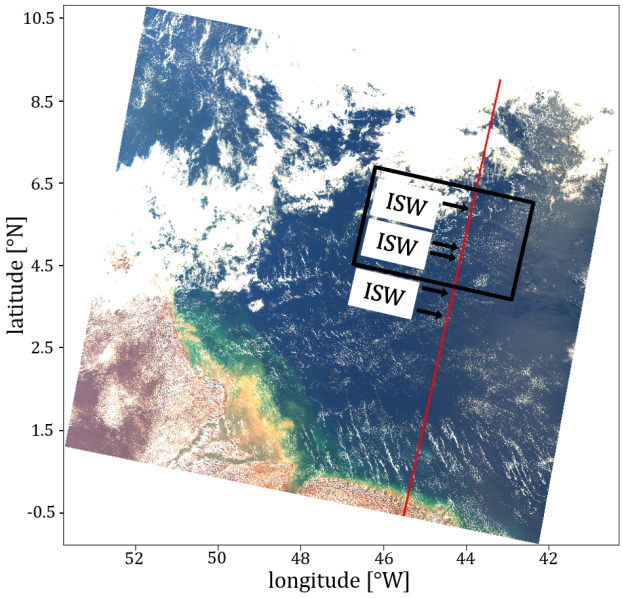


Figure 4. Schematics of data extraction from original OLCI image (black frame) and SRAL ground track (violet track) illustrate the different ground coverage of the modalities. Green patches indicate $^{P}\mathcal{P}$ (dark green: `NoIW`, light green: `IW`), blue patches $^{O}\mathcal{P}$ (dark blue: `NoIW`, light blue: `IW`), and in the case of red ones OLCI is not usable, so they belong to $^{S}\mathcal{P}$ (dark red: `NoIW`, light red: `IW`). Bright violet on the track indicates that the SRAL modality is used. For gray patches, both modalities are discarded. Stars A and B show the presumed ISW origins off the Amazon Shelf.

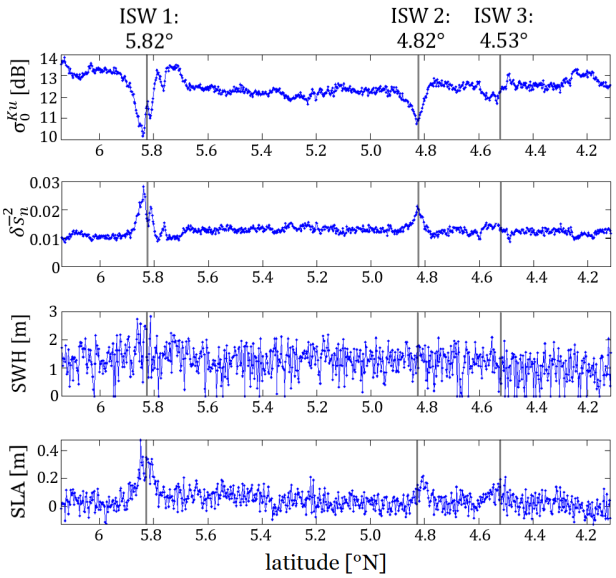
4. MULTI-MODAL DEEP LEARNING NETWORK

In our work, we design a multi-modal neural network called `SONet`, which is jointly trained on both modalities OLCI and SRAL. This is useful when a sensor fails, the image is obscured by clouds, or the radar signal is corrupted by coastal topography, as illustrated in Fig. 5. For instance, radar signals may indicate an IW, but in the OLCI image it is clearly visible that it is actually a rainstorm. Furthermore, correlations between the modalities can be utilized to control, discard or support results obtained from the other sensor. Besides, the spatial coverage of the OLCI image (swath width of 1270 km) is much larger than that of the SRAL signal (no swath, across width footprint diameter about 2km) on the ground.

Generally, processing of multi-modal data in machine learning is not trivial due to different characteristics, dimensions, units, scales and resolutions of input modalities. (Baltrušaitis et al., 2018) summarizes the core challenges of multi-modal learning as representation learning, alignment, fusion, translation and co-learning. In this work, the first three play an important role, while translation and co-learning are not required. Representation learning is about generalizing the data to exploit complementary and redundancy, for which we use multi-streams in `SONet`. The multi-stream technology is widely used



Georeferenced OLCI image with satellite ground track and marked ISWs.



SRAL parameter from inside the black box marked in Fig. 5a.

Figure 5. (a) True color OLCI image (visualized with the bands R=11, G=6, B=3) from the Amazon Shelf region from 25.05.2017 (Relative Orbit: 95, Cycle: 18, Frame Along Track: 2880). Several ISWs are visible crossing vertically the red ground track of the satellite on which SRAL measures. The flight direction is north to south. (b) 20 Hz SRAL records from inside the black box in Fig. 5a. From top to bottom: Radar backscatter (σ_0^{Ku}), Differenced-Mean-Square Slope ($\delta\bar{s}_n^2$), Significant Wave Height (SWH), Sea Level Anomaly (SLA).

in multi-modal deep learning because the modalities have different properties, and therefore require different operations for feature extraction (Wu et al., 2016, Huang, Kingsbury, 2013). Alignment describes the task to extract the connection between the modalities. In our dataset it is helpful that a spatial and temporal correlation already exist. Via fusion, both modalities are merged in the network to obtain a joint feature representation (Ngiam et al., 2011). At `SONet`, we opt for late fusion in terms of the depth of the network in which we are fusing. Although the interconnections in a late fusion are significantly weaker than in an early one, it is more suitable for modalities that have very different semantics. Furthermore it is easier to compensate a lack of modality (Liu et al., 2018). Other possibilities are early fusion, which requires similar semantics, much preprocessing and a high level of knowledge about the modality alignment, or to fuse multiple times and compute a weighted sum each (Vielzeuf et al., 2018).

4.1 Our multi-modal architecture

We have developed the neural network `SONet`, which handles the challenge of two-modality samples (Fig. 6). In the following we describe the complete structure starting with the required input form, the streams for each modality, the fusion of these streams and ending with the classification head.

For the input all OLCI data of the subsets form a matrix ${}^{O}\boldsymbol{X}$ and all SRAL data form a matrix ${}^{S}\boldsymbol{X}$. ${}^{O}\boldsymbol{X}$ has the dimension $[N \times M \times M \times B]$, where N is the number of samples, M the side length of the OLCI patch and B the number of used bands. ${}^{S}\boldsymbol{X}$ has the dimension $[N \times L \times T]$, where L is the number of SRAL observations that fall into a patch and T is the number of actually used parameter. So ${}^{O}\boldsymbol{X}$ represents the input for `O-Stream` and ${}^{S}\boldsymbol{X}$ for `S-Stream`. Note that both input matrices have the same number of samples N and correspond to the same reference label vector $\boldsymbol{y}$.

The modalite-specific streams consist of recurring sequences of convolutional blocks. Each convolution block contain (in the order given) CONV_{2D} or CONV_{1D}, ReLU (activation), BN (batch normalization), Pool_{Max} and DO (dropout), whereas regularization techniques BN and DO are optional. The output of the streams are flat layers, which can be considered as compact representation of the input modalities.

One key element of multi-modal learning is the fusion of the modalities. We decide for late fusion and merge both compact modality-specific representations at the end of the streams. There are several methods to fuse layers, like the operators addition, multiply, average, maximum, minimum and concatenation. Please note, that depending on the concrete fuse option the layers must have an identical size in at least one dimension. As result a joint representation of both modalities is obtained.

In deeper layers, the joint representation is subsequently compressed with fully-connected (FC) layers, until an output layer returns the classification result (classification head). The classification head consists of FC layers with ReLU activation. While in all hidden layers including the streams, ReLU is used as activation function, the output layer consisting of two neurons uses Softmax activation. Therefore, the value of the last two neurons can be interpreted as probability for the class assignment of the samples. One neuron stores the probability that the sample maps `NoIW`, the other neuron that it maps `IW`. We store the probability of being class `IW` in $\hat{\boldsymbol{y}}$.

4.2 Loss functions

In order to train the network, two loss functions are introduced. The first $\mathcal{CE}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ includes the measure of binary cross entropy

$$\mathcal{CE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -[\boldsymbol{y} \cdot \log(\hat{\boldsymbol{y}}) + (1 - \boldsymbol{y}) \cdot \log(1 - \hat{\boldsymbol{y}})] \quad (1)$$

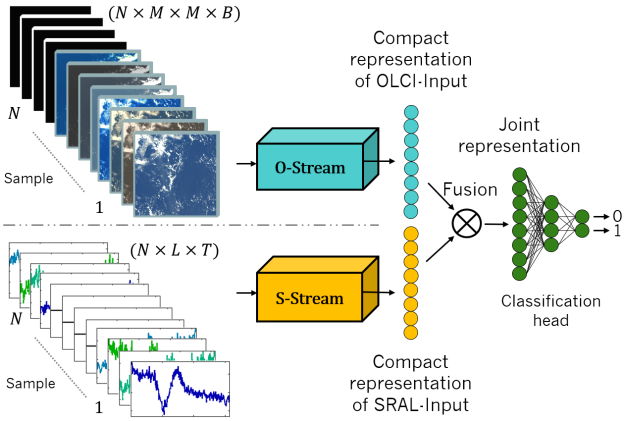


Figure 6. Architecture of `SONet`. Data is divided at the beginning in such a way that each modality first passes its own stream. As output the streams deliver flat layers, which are compact representations of the respective input modalities. Late fusion of these modality specific representations results in a joint representation layer. A classification head, which connects to this, compresses this representation further up to the classification output layer, which consists of two neurons representing probabilities of a sample being `NoIW` or `IW`.

where $\boldsymbol{y}$ represent the reference label and $\hat{\boldsymbol{y}}$ describes the probability of being class `IW`. Further the focal loss $\mathcal{FL}(\boldsymbol{p_t})$ is used which is suitable to compensate for an imbalance in the data set (Lin et al., 2017).

$$\mathcal{FL}(\boldsymbol{p}_t) = -\alpha_t \cdot (1-\boldsymbol{p}_t)^{\gamma} \cdot \log(\boldsymbol{p}_t) \tag{2}$$

Here γ is a previously defined integer scalar, $\boldsymbol{\alpha}_t$ the weighting factor

$$\boldsymbol{\alpha}_t = \begin{cases} \alpha & \text{if } \boldsymbol{y} = 1 \\ 1-\alpha & \text{otherwise} \end{cases} \tag{3}$$

where α is a predefined value between 0 and 1 and

$$\boldsymbol{p}_t = \begin{cases} \hat{\boldsymbol{y}} & \text{if } \boldsymbol{y} = 1 \\ 1-\hat{\boldsymbol{y}} & \text{otherwise} \end{cases}. \tag{4}$$

5. EXPERIMENTAL SETUP

First, information on data preprocessing, including brightness enhancement, normalization and data augmentation is given. This is followed by details on network architecture from input, streams and fusion to modifications for uni-modal baseline models. Finally the training procedure is explained with cross validation, hyperparameter settings and evaluation.

5.1 Data preprocessing

5.1.1 Brightness correction and normalization Since the brightness of OLCI images is very different from each other, it has also proved to be useful to correct each image individually. The 75% quantile values are calculated for all pixels of a band. The bands are then divided by their respective quantiles. After this operation all pixels that are larger than 1 are set to 1. Subsequently, a normalization is performed over each subset by calculating a z-transform over all samples of the subset. We also reduce the input image size to $[128 \times 128]$ for reasons of runtime. For SRAL only z-transform is performed.

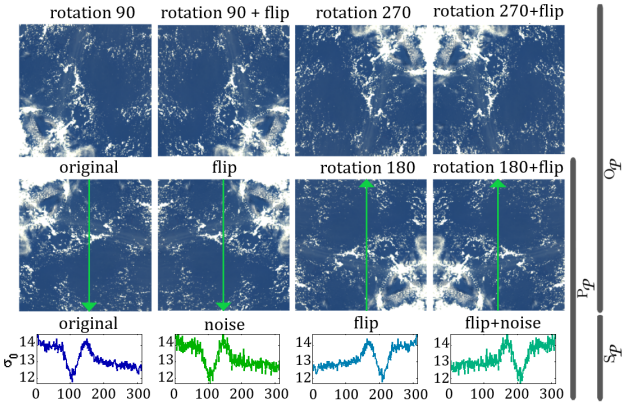


Figure 7. Exemplary illustration of an augmentation by rotation and flipping of the different subsets. The number of possible augmentations for Set ${}^{\mathrm{P}}\mathcal{P}$ is considerably limited in order to keep the alignments between modalities.

5.1.2 Augmentation of aligned modalities While commonly used data augmentation techniques such as flipping and rotating can be applied to the subsets ${}^{\mathrm{O}}\mathcal{P}$ and ${}^{\mathrm{S}}\mathcal{P}$, the operations are restricted for subset ${}^{\mathrm{P}}\mathcal{P}$. With each augmentation of the OLCI input, without an appropriate augmentation of the corresponding SRAL input, the alignment between the modalities would be lost, which would no longer be beneficial for multi-modal training. Fig. 7 shows which augmentations are performed in this work. ${}^{\mathrm{O}}\mathcal{P}$ is augmented 7 times with rotations and flippings. For ${}^{\mathrm{P}}\mathcal{P}$ the upper 4 augmentations are not suitable, so only 3 are left. ${}^{\mathrm{S}}\mathcal{P}$ also reaches 3 more augmentations with noise of the signal. After the augmentations, it is ensured that there are the same number of samples from each class by adding additional noise.

5.2 Detailed network structure

5.2.1 Stream-Input As already mentioned in Sec. 4.1, the input to `O-Stream` is ${}^{\mathrm{O}}\boldsymbol{X} = [{}^{\mathrm{P}}\boldsymbol{O}, {}^{\mathrm{O}}\boldsymbol{O}, {}^{\mathrm{S}}\boldsymbol{O}]$ and the input to `S-Stream` is ${}^{\mathrm{S}}\boldsymbol{X} = [{}^{\mathrm{P}}\boldsymbol{S}, {}^{\mathrm{O}}\boldsymbol{S}, {}^{\mathrm{S}}\boldsymbol{S}]$. Since the matrices ${}^{\mathrm{S}}\boldsymbol{O}$ and ${}^{\mathrm{S}}\boldsymbol{S}$ do not exist in their respective subsets, they are zeroed placeholder in the appropriate size so that the matrices can be concatenated together. So ${}^{\mathrm{S}}\boldsymbol{O}$ is a zero matrix with the size $[{}^{\mathrm{S}}N \times M \times M \times B]$ and ${}^{\mathrm{O}}\boldsymbol{S}$ with dimension $[{}^{\mathrm{O}}N \times L \times T]$. Both streams have a common reference vector $\boldsymbol{y} = [{}^{\mathrm{P}}\boldsymbol{y}, {}^{\mathrm{O}}\boldsymbol{y}, {}^{\mathrm{S}}\boldsymbol{y}]$ with length $N = [{}^{\mathrm{P}}N + {}^{\mathrm{O}}N + {}^{\mathrm{S}}N]$, which is the total number of all samples in all subsets.

It has been evaluated in advance that accuracy cannot be increased by combining various OLCI bands. In preliminary experiments, we observed that with band 16 the IWs are most visible. Therefore this band is used as the single input. Compared to the use of all bands, this offers a considerable reduction of runtime. For SRAL, we use information from several parameters, namely radar backscatter σ_0^{Ku}, differenced-mean square slope $\delta\bar{s}_n^2$, SWH, and SLA. After preprocessing, we first perform data augmentations for all subsets, resulting in ${}^{\mathrm{P}}N \rightarrow {}^{\mathrm{P}}N^*$, ${}^{\mathrm{O}}N \rightarrow {}^{\mathrm{O}}N^*$, and ${}^{\mathrm{S}}N \rightarrow {}^{\mathrm{S}}N^*$, with N^* indicating the size after augmentation. After band selection, the `O-Stream` input ${}^{\mathrm{O}}\boldsymbol{X}$ has a size of $[N^* \times 128 \times 128 \times 1]$, and the `S-Stream` input ${}^{\mathrm{S}}\boldsymbol{X}$ has a size of $[N^* \times 313 \times 4]$.

5.2.2 `O-Stream` Each of four convolutional block consists of three different layers in the order of $\mathrm{CONV}_{2\mathrm{D}}$, ReLU and $\mathrm{Pool}_{\mathrm{Max}}$. The kernel size in all $\mathrm{CONV}_{2\mathrm{D}}$ of all convolutional blocks is (3×3) at a stride of (1) in each direction. Besides, in

CONV_{2D} layers of the O stream, L2 kernel regularization with the parameter 0.01 is also applied to reduce overfitting. The number of filters increases continuously from 16 in the first convolutional block over 32 and 64 to 128 in the last one. Pool_{Max} uses a kernel of size (2×2) and a stride of (2) in each direction. So the layer size is exactly halved by each pooling operation.

5.2.3 `S-Stream` 3 convolutional blocks consisting of CONV_{1D}, ReLU and Pool_{Max} are used in this stream. However, the main difference in the convolutional blocks, is the layer CONV_{1D}, so the parameters are folded in the direction of L only. The number of filters is 16 in the first convolution block and 32 in the second and 64 in the last one. The kernel size of CONV_{1D} is (3) with stride (1). Accordingly, the POOL_{Max} is one-dimensional with a pooling kernel of size (2) and stride of (2).

5.2.4 Fusion After flattening both streams, we use a fusion of width FC_{128} and fusion type addition to merge the streams outputs. Therefore, we need both flat layers to have the same size. To get both flattened stream outputs to the size of 128 neurons, a dense layer is used (Flat $\rightarrow \text{FC}_{128}$). The joint representation also has the size of 128 neurons. Subsequently, the classification head is connected to the joint representation. The number of neurons in the classification head is slowly reduced by connecting two fully connected layers with ReLU activation of width 32 and 8. The softmax output layer with 2 neurons is the last layer.

5.2.5 Uni-modal baseline networks For comparison, uni-modal networks `SNet` and `ONet` are created, in which the classification head is directly attached to the respective stream. So the fusion is omitted, but stream and classification head are identical to `SONet`. The input matrices are slightly modified for the training of uni-modal networks, since it is not efficient to use the zero placeholder matrices for training. Therefore ${}^{\text{O}}\boldsymbol{X}$ is reduced for `ONet` to ${}^{\text{ONet}}\boldsymbol{X} = [{}^{\text{P}}\boldsymbol{O}, {}^{\text{O}}\boldsymbol{O}]$ with shortened reference label ${}^{\text{ONet}}\boldsymbol{y} = [{}^{\text{P}}\boldsymbol{y}, {}^{\text{O}}\boldsymbol{y}]$, and ${}^{\text{S}}\boldsymbol{X}$ is reduced for `SNet` to ${}^{\text{SNet}}\boldsymbol{X} = [{}^{\text{P}}\boldsymbol{S}, {}^{\text{S}}\boldsymbol{S}]$ with ${}^{\text{SNet}}\boldsymbol{y} = [{}^{\text{P}}\boldsymbol{y}, {}^{\text{S}}\boldsymbol{y}]$.

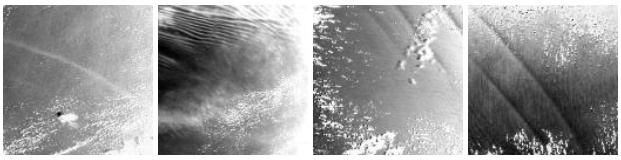
5.3 Training procedure

5.3.1 Cross Validation For the purpose of cross-validation, the data set (see Fig. 1) is quartered by combining the data from two RO in each case (1: 38+152, 2: 95+209, 3: 380+109, 4: 52+166). The split of the total 2973 sample is done in this way to find a similar amount of data in each quarter (1: 663, 2: 653, 3: 456, 4: 601). Moreover this ensures the samples in the different quarters are geographically separated from each other. Consciously we decided against a temporal split, because IWs at the same location at different times can look very similar, which would falsify the cross validation. In the experiments, training is performed sequentially on three quarters and testing is performed on the fourth one. This results in a total of 4 cross validation runs.

5.3.2 Pre-training and fine-tuning As a baseline, we first train models from the uni-modal networks. For `SNet` we use 50 epochs, with a learning rate of $1e-4$, linear decay and a batchsize of 64. $\mathcal{CE}$ with Adam Optimizer has proven to be the most suitable loss-function. ONet is trained for a longer period of 200 epochs, but with a lower learning rate of $1e-5$. Decay and batchsize are identical to `SNet`. However, the loss function is $\mathcal{FL}\,(\alpha = 0.5, \gamma = 3)$ with Adam Optimizer.
Since uni-modal networks and `SONet` have the same streams, it has been proven useful for training `SONet` to use as initial stream-weights those from uni-modal networks. Thus the essential feature extraction is already given, so that only the connection between the modalities has to be learned. For this, `SONet` is trained with 100 epochs with a learning rate of $1e-5$, a batch size of 64 and a linear decay. As with `ONet`, $\mathcal{FL}\,(\alpha = 0.5, \gamma = 3)$ is used with Adam optimizer.

	${}^{\text{P}}\mathcal{P}$		${}^{\text{O}}\mathcal{P}$		${}^{\text{S}}\mathcal{P}$	
RO	`NoIW`	`IW`	`NoIW`	`IW`	`NoIW`	`IW`
38	35	15	15	6	30	5
95	126	60	106	48	29	24
152	127	30	173	96	61	70
209	94	6	6	7	121	26
380	46	17	51	45	20	12
52	80	6	132	44	141	15
109	40	2	126	14	63	20
166	45	4	71	1	39	23

Table 1. Total amount of 2973 referenced samples in the Amazon shelf area divided into RO (sorted from west to east), subset and class.



IW from ${}^{\text{O}}\mathcal{P}$ correctly classified by `SONet`.

IW from ${}^{\text{O}}\mathcal{P}$ incorrectly classified by `SONet`.

Figure 8. Examples of predictions with `SONet` from subset ${}^{\text{O}}\mathcal{P}$ of positive (`IW`) referenced data. The upper row shows predicted `IW` (true positives), the lower row predictions for `NoIW` (false negatives). Visualized is band 16, since this band is used for training and testing.

5.3.3 Evaluation metrics To evaluate the classifiers, a confusion matrix is calculated from the reference class labels $\boldsymbol{y}$ and the predicted class labels $\bar{\boldsymbol{y}}$ ($\hat{\boldsymbol{y}}$ rounded to 0 and 1). The evaluation metrics overall accuracy (OA), average accuracy (AA) and F_1-Score are derived from this. As a further metric the MSE is introduced. This is calculated from the probabilities $\hat{\boldsymbol{y}}$ the network returns in the output layer and the referenced labels $\boldsymbol{y}$.

$$\text{MSE} = \frac{1}{N} \cdot \sum_{n=1}^{N} (\boldsymbol{y}_n - \hat{\boldsymbol{y}}_n)^2 \tag{5}$$

It can be understood as a measure of how reliably the classifier can decide in the prediction for one of the classes. The smaller MSE, the higher the reliability of the prediction.

6. RESULTS

We intend to investigate which classification result is achievable on the specific subsets. For this, we compare the performance of the uni-modal networks `ONet` and `SNet` with the multi-modal network `SONet`. The results are summarized in Tab. 2.

For subset ${}^{\text{S}}\mathcal{P}$, it is noticeable that `SNet` and `SONet` perform about equally well. Both networks achieve an OA of approxim-

	${}^{\mathrm{P}}\mathcal{P}$				${}^{\mathrm{O}}\mathcal{P}$				${}^{\mathrm{S}}\mathcal{P}$			
	OA [%]	AA [%]	F_1	MSE	OA [%]	AA [%]	F_1	MSE	OA [%]	AA [%]	F_1	MSE
SONet	91.86 ±1.00	84.60 ±3.13	0.75 ±0.04	**0.06** ±0.01	**77.96** ±3.41	**71.02** ±3.13	**0.54** ±0.07	**0.18** ±0.01	**88.14** ±1.38	**86.28** ±1.08	**0.79** ±0.02	**0.09** ±0.01
ONet	71.47 ±5.72	52.52 ±3.79	0.21 ±0.06	0.21 ±0.02	70.31 ±5.58	63.88 ±4.58	0.44 ±0.05	0.20 ±0.02				
SNet	**92.19** ±1.06	**87.50** ±1.94	**0.76** ±0.03	**0.06** ±0.01					87.79 ±0.88	86.02 ±1.07	**0.79** ±0.02	0.10 ±0.01

Table 2. Results achieved with SONet compared with ONet and SNet after cross validation. The mean values of the cross validation runs and the standard deviations (indicated by ±) are given. The results are divided into tests with multi-modal data (subset ${}^{\mathrm{P}}\mathcal{P}$) and tests with uni-modal data (${}^{\mathrm{O}}\mathcal{P}$: OLCI, ${}^{\mathrm{S}}\mathcal{P}$: SRAL).

ately 88%, an AA of 86%, and perform equally well considering F_1 and MSE. No results are given for ONet as it is designed for OLCI data as input only. Overall, this subset shows that IWs can be detected well with SRAL data, as already shown by (Santos-Ferreira et al., 2019), and that our deep learning framework is a suitable method to detect them. Besides SONet and SNet have the ability to use all four parameters σ_0^{Ku}, $\delta\bar{s}_n^2$, SWH, and SLA as joint input and to weight them according to their information content. Furthermore, apart from normalization, no preprocessing of the SRAL data is necessary.

Focusing on subset ${}^{\mathrm{O}}\mathcal{P}$, it is evident for all parameters that SONet performs significantly better than ONet. OA increases from 70.31% to 77.96%, AA from 63.88% to 71.02%, and also F_1 (0.54 instead of 0.44) and MSE (0.18 instead of 0.20) are significantly improved. While ONet is trained just on OLCI data, SONet uses multi-modal training to increase the accuracy. Thus, SONet succeeds in exploiting correlations and alignments between the modalities. The direct comparison between ${}^{\mathrm{S}}\mathcal{P}$ and ${}^{\mathrm{O}}\mathcal{P}$ shows that the classification based on the optical data does not achieve comparable high accuracies as the ones obtained by radar data. This is caused by the diversity of OLCI images due to different brightness, wave characteristics, and cloud loading (Fig. 8). The radar signal is less sensitive to these influences - in addition, the amount of training data in our data set for ${}^{\mathrm{S}}\mathcal{P}$ is larger.

We would like to point out that when testing uni-modal data with SONet the other modality is set to 0 due to the lack of modality. The approach of zeroing to fix the lack of modality works well for training and testing SONet, which is underlined by a similar or higher accuracy of the network in comparison to the uni-modal networks.

${}^{\mathrm{P}}\mathcal{P}$ is the multi-modal subset, which is directly used as input in SONet. However, in ONet and SNet, only those modalities can be included that are designed for the corresponding network. Thus, although both modalities are available, tests with ONet discard the SRAL modality and tests with SNet discard the OLCI modality. SNet reaches the highest values for all parameters OA (92.16%), AA (87.50%), F1(0.76%), and MSE (0.06). SONet is either as good (MSE) or slightly worse (OA: 91.86%, AA: 84.60%, F1:0.75%) but with standard deviation similar to SNet. ONet has a significantly lower accuracy in all categories, where OA is about 21% and AA about 35% below best performance. As already seen in set ${}^{\mathrm{O}}\mathcal{P}$, the classification based just on OLCI images is much more difficult, which means that ONet performs worse than SNet. We observe that SONet is able to utilize the strongest modality with SRAL and suppress the weaker one, which is a strength of a multi-modal network. Furthermore, reliability of SONet can be considered higher as it uses complementary information, and therefore utilizes a more comprehensive view on the phenomenon.

We have conducted additional experiments with a Random Forest (RF) (Breiman, 2001). Unlike SONet, a RF has the weakness that it is poorly suited for multi-modal input that has different dimensions, which requires prior manual embedding or the application of a dimensionality reduction algorithm. We use PCA-obtained feature vectors of the same size for each modality independently, to avoid a potential weighting between both modalities. With uni-modal input, RF achieves maximum accuracies of ONet and SNet. Nevertheless, the accuracy with multi-modal input does not increase over uni-modal input.

7. CONCLUSION

In this work, we demonstrated that our multi-modal deep learning framework is able to detect oceanic internal waves. We thus feel confident to suggest that such networks are a promising research direction in the earth sciences. Overall, the multi-stream technique with late fusion is well suited to exploit correlations and alignments between modalities. If both modalities are available, strong results (overall accuracy: 92%) are already achieved based just on Sentinel-3 SRAL data. However, the ground coverage of Sentinel-3 OLCI is much larger, which is essential for a continuous global observation of internal waves. Hence, areas which are not covered by SRAL tracks, a multi-modal network significantly increases the overall accuracy (78% instead of 70%) and the average accuracy (71% instead of 64%), when compared to an uni-modal ONet network. Meanwhile, SONet also performs as well as SNet in areas where SRAL is present. Due to the higher reliability through different input data types, we recommended to use SONet for classification for all subsets. We have also shown that zeroing the missing modality does not negatively affect the training of a multi-modal network. This allows a multi-modal data set to be extended very easily by uni-modal data. Future work will concern applications from satellite remote sensing and the integration of further modalities, but also the joint use of close-range data from multi-sensor systems, as is the case in the field of precision agriculture.

ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2070 – 390732324.

REFERENCES

Alpers, W., Brandt, P., Rubino, A., 2008. Internal waves generated in the straits of gibraltar and messina: Observations from space. *Remote sensing of the European seas*, Springer, 319–330.

Aytar, Y., Castrejon, L., Vondrick, C., Pirsiavash, H., Torralba, A., 2017a. Cross-modal scene networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(10), 2303–2314.

Aytar, Y., Vondrick, C., Torralba, A., 2017b. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*.

Baltrušaitis, T., Ahuja, C., Morency, L.-P., 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443.

Benedetti, P., Ienco, D., Gaetano, R., Ose, K., Pensa, R. G., Dupuy, S., 2018. $M^3Fusion$: A Deep Learning Architecture for Multiscale Multimodal Multitemporal Satellite Data Fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(12), 4939–4949.

Breiman, L., 2001. Random forests. *Machine learning*, 45(1), 5–32.

Gupta, A. K., Cheng, Y. C., 2006. Method, system and module for multi-modal data fusion. US Patent 7,152,033.

Huang, J., Kingsbury, B., 2013. Audio-visual deep learning for noise robust speech recognition. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 7596–7599.

Ikeda, M., 1995. *Oceanographic applications of remote sensing*. CRC press.

Jackson, C., 2007. Internal wave detection using the moderate resolution imaging spectroradiometer (MODIS). *Journal of Geophysical Research: Oceans*, 112(C11).

Jackson, C. R., Da Silva, J. C., Jeans, G., 2012. The generation of nonlinear internal waves. *Oceanography*, 25(2), 108–123.

Klemas, V., 2012. Remote sensing of ocean internal waves: An overview. *Journal of Coastal Research*, 28(3), 540–546.

Lahat, D., Adali, T., Jutten, C., 2015. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9), 1449–1477.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, K., Li, Y., Xu, N., Natarajan, P., 2018. Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730*.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B. A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152, 166–177.

Magalhães, J. M., Da Silva, J., Buijsman, M. C., Garcia, C., 2016. Effect of the North Equatorial Counter Current on the generation and propagation of internal solitary waves off the Amazon shelf (SAR observations). *Ocean Science*, 12(1), 243–255.

Mroueh, Y., Marcheret, E., Goel, V., 2015. Deep multimodal learning for audio-visual speech recognition. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2130–2134.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A. Y., 2011. Multimodal deep learning. *Proceedings of the 28th international conference on machine learning (ICML-11)*, 689–696.

Qiu, B., Nakano, T., Chen, S., Klein, P., 2017. Submesoscale transition from geostrophic flows to internal waves in the northwestern Pacific upper ocean. *Nature communications*, 8(1), 1–10.

Rao, D., De Deuge, M., Nourani-Vatani, N., Douillard, B., Williams, S. B., Pizarro, O., 2014. Multimodal learning for autonomous underwater vehicles from visual and bathymetric data. *2014 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 3819–3825.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. et al., 2019. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195.

Robinson, I. S., 2010. *Discovering the Ocean from Space: The unique applications of satellite oceanography*. Springer Science & Business Media.

Santos-Ferreira, A., da Silva, J., Magalhaes, J., 2018. SAR Mode Altimetry Observations of Internal Solitary Waves in the Tropical Ocean Part 1: Case Studies. *Remote Sensing*, 10(4), 644.

Santos-Ferreira, A., da Silva, J., Srokosz, M., 2019. SAR Mode Altimetry Observations of Internal Solitary Waves in the Tropical Ocean Part 2: A Method of Detection. *Remote Sensing*.

Shendryk, Y., Rist, Y., Ticehurst, C., Thorburn, P., 2019. 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.

Srivastava, N., Salakhutdinov, R. R., 2012. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 2222–2230.

Vielzeuf, V., Lechervy, A., Pateux, S., Jurie, F., 2018. Centralnet: a multilayer approach for multimodal fusion. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Vukotić, V., Raymond, C., Gravier, G., 2016. Multimodal and crossmodal representation learning from textual and visual features with bidirectional deep neural networks for video hyperlinking. *Proceedings of the 2016 ACM workshop on Vision and Language Integration Meets Multimedia Fusion*, ACM, 37–44.

Wu, Z., Jiang, Y.-G., Wang, X., Ye, H., Xue, X., 2016. Multi-stream multi-class fusion of deep networks for video classification. *Proceedings of the 24th ACM international conference on Multimedia*, ACM, 791–800.

Yang, X., Ramesh, P., Chitta, R., Madhvanath, S., Bernal, E. A., Luo, J., 2017. Deep multimodal representation learning from temporal data. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5447–5455.

Zhao, Z., Klemas, V., Zheng, Q., Yan, X.-H., 2004. Remote sensing evidence for baroclinic tide origin of internal solitary waves in the northeastern South China Sea. *Geophysical Research Letters*, 31(6).