Deep Sea Research Part I: Oceanographic Research Papers
June 2023, Volume 196, Pages 104024 (26p.)
https://doi.org/10.1016/j.dsr.2023.104024
https://archimer.ifremer.fr/doc/00828/93980/

K nearest neighbors classification of water masses in the western Alboran Sea using the sigma-pi diagram

Belattmania Ayoub [1], El Arrim Abdelkrim [1], Ayouche Adam [2],, * Charria Guillaume [5], Hilmi Karim [3], El Moumni Bouchta [4]

[1] Faculté des Sciences et Techniques, Tanger, Morocco
[2] Laboratory for Ocean Physics and Satellite Remote Sensing (LOPS), UMR6523, Ifremer, Univ. Brest, CNRS, IRD, Brest, France
[3] Institut National de Recherche Halieutique, Casablanca, Morocco
[4] Université Abdelmalek Essaâdi, Tétouan, Morocco
[5] Laboratory for Ocean Physics and Satellite Remote Sensing (LOPS), UMR6523, Ifremer, Univ. Brest, CNRS, IRD, Brest, France

* Corresponding author : Adam Ayouche, email address : adam.ayouche@ensta-bretagne.org

Abstract :

Different classification techniques of water masses have been developped using the potential temperature-salinity (θ-S) diagram and its volumetric analysis. In this study, we propose a new method to automatically classify water masses via a supervised machine learning algorithm based on the K nearest neighbors (Knn), in the potential density and potential spicity (σ-π) coordinates. This method is applied to temperature and salinity data collected in the western side of the Alboran Sea during a glider mission, dedicated to sample the Western Alboran Gyre (WAG) in late winter 2021. The water masses in the studied region were classified into five different categories following a supervised learning process, based on ocean profile databases available on the region of interest. The results corroborate previous studies of the spatial distribution of water masses in the Alboran Sea, inferred from traditional method based on the expert analysis of the (θ-S) diagram, and suggest that this methodology is efficient and reliable for water masses classification. Compared to a classical clustering computation (herein k-means), this method is more appropriate in a region where the characteristics of the water masses change considerably in both space and time.

Highlights

► High spatial resolution glider profiles of θ-S in the western Alboran sea. ► Water masses derived on a (σ-π) diagram using Knn algorithm. ► Classification results confirm earlier derived circulation schemes. ► The proposed method outperforms classical clustering analysis in delineating water mass boundaries.

Keywords : Alboran Sea, Western Alboran Gyre, water masses, (σ-π) diagram, K nearest neighbor classification.

1. Introduction

A water mass is a volume of oceanic water with horizontal and vertical extensions, and having specific physical characteristics. In general, most of the water masses are formed by atmosphere-ocean exchanges, however some others acquire their characteristics (e.g minimum salinity) through biochemical or physical processes (e.g convection). The signature of such characteristics are represented by tracers such as the potential temperature and salinity. These tracers are important to understand the oceanic circulation at different global and regional scales, as the thermohaline circulation (Broecker, 1991). The thermohaline circulation plays a key role in the climate regulation by the transport of heat, carbon and oxygen across the different basins around the world (Clark et al., 2002).

In this context, Pantiulin (2002) sketches a brief history about the genesis of the concept of water masses, depending on the evolution of the in-situ observations of temperature and salinity. Indeed, the definition, classification and first principles of water masses appeared for the first time in the monograph called the Norwegian Sea in 1909 (Hansen and Nansen). The latter was followed by the introduction of the potential temperature-salinity (θ-S) diagram as a tool to analyze water masses properties, in a Norwegian study after the first world war (Hansen, 1916). He showed on a wide area of the eastern Atlantic ocean that the variations in the (θ-S) diagram can be attributed to the intrusion of offshore water masses.

Since then, the (θ-S) diagram has been used widely in physical oceanography and by numerous authors across different fields. Major progress in the water mass analysis was the introduction of the volumetric (θ-S) diagram which was used in different studies that includes the Pacific ocean, the Indian ocean, the Atlantic ocean and the Global ocean (Cochrane, 1958; Pollak, 1958; Montgomery, 1958). In these studies, the quantity of volumetric units for standard levels of depths were estimated statistically. This estimation was based on a division of the oceans in bi-variate classes defined by their temperature and salinity.

Other studies followed the previous ones based on the volumetric (θ-S) statistically analysis methodology. They improved and reworked this methodology for the sake of understanding the water masses distribution in a volumetric (θ-S) diagram (Miller and Stanley, 1961; Wright and Worthington, 1970; Worthington, 1981).

Besides the volumetric (θ-S) diagram analysis, other techniques of water mass classification have been used such as the cluster analysis where the data are grouped on the basis of a set of measured parameters. The objective of this method is to find an optimal data distribution which minimizes a certain metric that define the similarity within the clusters. For example, Kim et al. (1991) applied a clustering analysis based on the average linkage between groups for the temperature and salinity to identify the water masses in the Yellow sea and the East China sea. The metric used for their clustering analysis is the squared Euclidean distance defined as the normalized temperature and salinity differences between points.

Hur et al. (1999) studied the yellow and east china seas for over 40 years (1950-1992) using historical data of temperature and salinity. They included in their study the geographical distance and the depth separation in computing the distance for the clustering method. Naranjo et al. (2015) examined the distribution and spatio-temporal evolution of water masses in the strait of Gibraltar using clustering analysis. These authors used historical values of potential temperature, salinity and potential density for each water mass as initial centroids for the classification. Roseli et al. (2015) applied the k-means algorithm on temperature and salinity data from CTD casts in two different seasons (fall and summer), to classify water masses at the Shallow Sunda Shelf of Southern South China Sea. Recently, Gao et al. (2020) proposed a novel and robust method to identify the frontiers between water masses in the Northern South China sea. Their identification of the water masses center is based on ranges and standard deviations of the potential spicity π in different potential density layers, and water volumetric distributions in the bi-dimensional plan (σ-π).

The (θ-S) diagram and its different techniques for analyzing water masses have been developed for several oceans; but it is also interesting to conduct such studies for regions where several water masses from different oceans can interact, such as the Alboran Sea : the westernmost Mediterranean sub-basin where Atlantic and Mediterranean waters interact through the strait of Gibraltar. In our knowledge, only traditional water masses analysis based on (θ-S) diagram, have been previously used in this region (Bryden et al., 1982; Pistek et al., 1985; Gascard and Richez, 1985; Parrilla et al., 1986; Parrilla and Kinder, 1987; Millot et al., 2006; Millot, 2009; Renault et al., 2012; Millot, 2014).

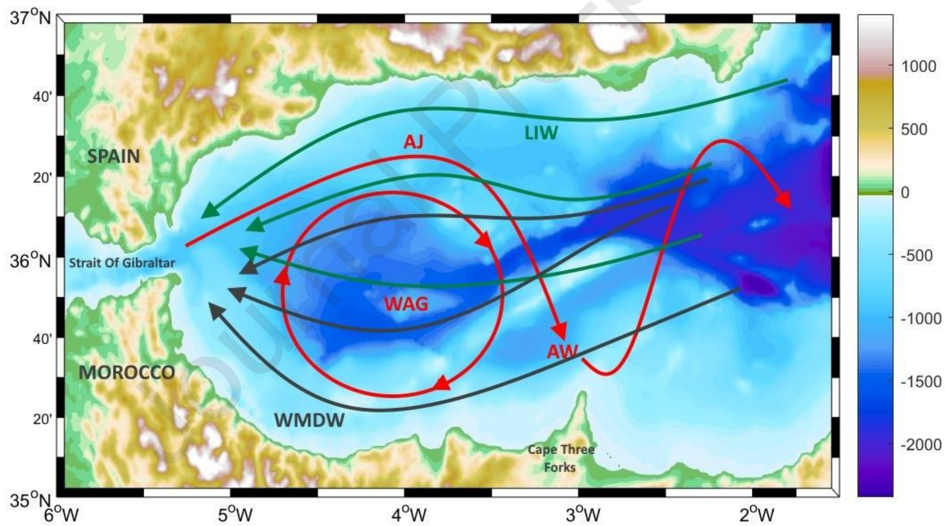


Figure 1: Map of the Western Alboran Sea sketching the bathymetric depths and topographic elevations in meters (m) relative to the mean sea level. Red arrows show the general surface circulation of Atlantic Water (AW), showing the Western Alboran Gyre (WAG) as well as the Atlantic Jet (AJ). The green and black arrows represent respectively the intermediate and deep circulation of Mediterranean waters (LIW and WMDW).

The related circulation schemes have been sketched for each water mass of the Atlantic ocean and the Mediterranean sea (figure 1) where their properties and distributions are summarized in table 1 and can be described as follows:

The Atlantic Water (AW), located in the western side of the strait of Gibraltar, is injected in the Alboran Sea (top 200 m depth). It is subject to different variations through its cyclonic path (Coriolis effect) at the surface due to its interaction with the atmosphere and the surface mixing with older Atlantic water. This water becomes saltier ($\sim$ 38 psu) and progressively cooler in winter ($\sim$ 13 $^{\circ}$C) and therefore this results in increased density. Then, this water is called the Modified Atlantic Water (MAW). At the surface layer, quasi-homogeneous light waters are observed with a salinity S = 36.6 psu: the Surface Atlantic Water (SAW). A second layer, the North Atlantic Central Waters (NACW), is characterized by a minimum of salinity separating the SAW from the MAW. This separation is progressively dissipated through mixing in the strait of Gibraltar and the Alboran Sea (T=11-17 $^{\circ}$C, S=35.6-36.5 psu).

Previous studies about the Mediterranean Waters (MWs) in the Alboran Sea suggest the presence of Winter Intermediate Water (WIW), Levantine Intermediate Water (LIW), Western Mediterranean Deep Waters (WMDW) and the Tyrrhenian Deep Water (TDW). The LIW and WMDW were considered as the main contributors for the outflow. The WIW results from AW cooling along the continental shelf of the Liguro-Provencal sub-basin and is generated periodically in the Alboran Sea near the Spanish continental shelf. The WIW can be identified by its minimum potential temperature (12.9-13 $^{\circ}$C) between 100 and 350 m depth and between 28 and 29 kg.m^{-3} isopycnals. The LIW from the Western Mediterranean sea generated by winter convection is the most salty and warmest water mass encountered at mid depth (200-600 m) in the Alboran Sea. The LIW is mostly concentrated in the north and center sides of the Alboran Sea and absent along the African coast. The LIW is characterized by temperature and salinity maximum (13.1-13.3 $^{\circ}$C, 38.47-38.52 psu). The WMDW is generated in the gulf of Lion by deep convection and is cold ($<$ 12.9 $^{\circ}$C) and relatively salty ($>$ 38.4 psu) water. The WMDW is considered as the most dense water in the Mediterranean sea (at 800 m depth in the central part of the Alboran Sea). The TDW is the result of mixing between ancient WDMW in the Tyrrhenian sea and the LIW coming from the Western Mediterranean sea through the strait of Sicily. The TDW is slightly denser than the LIW and lighter than the WMDW and lies between these two water masses. In the Alboran Sea, the temperature and salinity values of the TDW are respectively within the range 13-13.1 ˚C and 38.41-38.51 psu.

Water mass	Description	Reference
SAW	Quasi homogenous salinity layer ($S \approx 36.6$) and a constant temperature gradient.	(Gascard and Richez, 1985; Parrilla et al., 1986; Vélez-Belchı et al., 2005)
NACW	The Seperation layer between the SAW and MAW. It's characterized by a salinity minimum (35.5-36.6) that attenuated quite rapidly after entering the Mediterranean Sea.	(Gascard and Richez, 1985; Parrilla et al., 1986; Vélez-Belchı et al., 2005)
MAW	A mixture layer of Atlantic (16˚C-36.5) and Mediterranean waters (12.9˚C-38.45)	(Gascard and Richez, 1985; Parrilla et al., 1986; Vélez-Belchı et al., 2005)
LIW	The warmest and saltiest Mediterranean waters, easily recognised anywhere in the sea. Concerning the western Alboran, it is characterised by (T=13.1-13.2˚C and S=38.5).	(Gascard and Richez, 1985; Parrilla and Kinder, 1987; Millot et al., 2006; Millot, 2009, 2014)
WIW	Results from the AW wintertime cooling in the northern part of the western basin and characterised by a Temperature minimum (12.9-13˚C).	(Millot, 2009, 2014)
TDW	Results from mixing between ancient WDMW in the Tyrrhenian sea and the LIW coming from the Western Mediterranean sea. Its core characteristics are in ranges ($T = 13.0$-13.1˚C and $S = 38.48$-38.51).	(Millot et al., 2006; Millot, 2009, 2014)
WMDW	Formed in the Liguro-Provencal mainly from an AW-LIW mixture by wintertime convection processes. It is Cold ($<$ 12.9˚C) and relatively salty ($>$ 38.4).	(Gascard and Richez, 1985; Parrilla and Kinder, 1987; Millot et al., 2006; Millot, 2009, 2014)

Table 1: Summary of water masses definitions with their respective references.

The application of clustering analysis methods for the purpose of automatically classify water masses, has yielded encouraging results in many regions. Nevertheless, these techniques have revealed many shortcomings in region with a high spatio-temporal variability and could not exactly identify the water mass boundary (Gao et al. (2020)). Clustering analysis is particularly relevant to distinguish water masses with similar salinity and temperature variance (Naranjo et al. (2015)). This is not the case in the Alboran Sea, where SAW is widely variable in temperature and the MWs range much more in temperature than in salinity.

Within this context of challenges to be solved in automatic water masses classification notably in region where intense mixing occurs, in this paper, we propose a novel methodology that classify automatically water masses in the Alboran Sea, based on machine learning supervised algorithm, applied on curvilinear potential density and potential spicity (σ-π) diagram. Two datasets, described in section 2, have been used for the study. The first concerns the global database of temperature and salinity vertical profiles used as a training dataset of the algorithm. The second is relative to the glider in-situ observations collected in the Western Alboran Sea, on which the classification algorithm is applied. Section 2 also describes the classification methodology, starting with the labeling process and ending with the sensitivity test of the employed method. Water mass classification results in the glider transects are provided in section 3 and are discussed in section 4. Finally, conclusions are drawn in section 5.

2. Materials and Methods

2.1. Database

To build the training water mass classes, we assemble the available in-situ observations from oceanographic databases such as World Ocean Database 2018 ’WOD18’ (Boyer et al., 2019) and the Global Data Assembly Centers ’GDACs’ (Argo, 2021) in a given geographic domain (Figure 2).

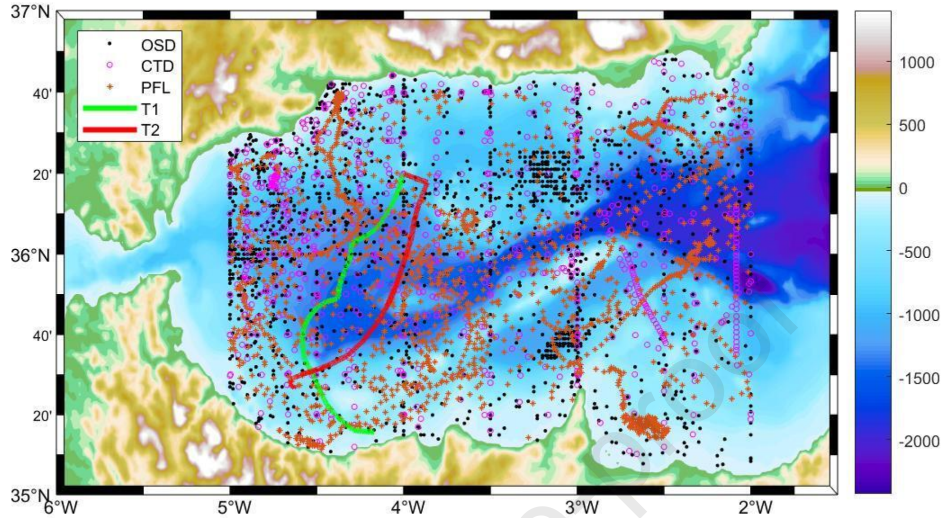


Figure 2: Map of the Western Alboran Sea sketching the bathymetric depths and topographic elevations in meters (m) relative to the mean sea level. The black dots and magenta circles indicate the localization of the vertical profiles of WOD18 related to Ocean Station Data (OSD) dataset and Conductivity Temperature Depth (CTD) dataset respectively (Table 2). Brown asterisks represent Argo Profiling Floats (PFL) trajectories from GDACs (Table 2). The first glider transect (T1) is sketched in green and the second transect (T2) in red.

The resultant product is based on 5068 vertical profiles of temperature and salinity including 1759 sampling cycles of Argo floats. The in-situ data are gathered over a broad range of temporal scales between 1951 and 2020. Table 2 summarizes the key informations about the mentioned databases.

2.2. Glider data

In this study, our analysis is focused on the second mission performed by the Moroccan association AGIR (Leader of the Marine Observatory of Al Hoceima) in the Western Alboran Sea, as part of the European project ODYSSEA (`https://odysseaplatform.eu/fr/home-fr/`). This mission was conducted after a first one in late fall 2020 (from 10 November to 11 December) in the

Dataset	Description	Temporal range	# of casts	# of TS observations
OSD	measurements made from a stationary vessels using reversing thermometers mounted on special bottles including LVR CTD rosette system, LVR STD and LVR XCTD (XCTD is collected from moving vessels).	1951-2011	2344 stations	26568
CTD	data from a stationary vessels using HVR CTD rosette system, STD (The salinity S is computed from the conductivity) data measured at high frequency with respect to depth as well as HVR XCTD (XCTD is collected from moving vessels).	1975-2018	965 stations	344881
PFL	contains temperature and salinity data collected from drifing profiling floats of the Argo project.	2006-2020	1759 cycles	249788

Table 2: Database information used for the supervised learning. The acronyms XCTD, STD, LVR and HVR stand respectively for: eXpandable Conductivity Temperature Depth, Salinity Temperature Depth, Low Vertical Resolution and High Vertical Resolution. All casts with a depth increment less than two meters are considered High Resolution otherwise, the casts are considered as Low Resolution.

same region (Nibani et al., 2021). The second mission occurred in late winter – early spring (from 11 February to 23 March). During this mission, a Sea-Explorer glider (manufactured and commercialized by ALSEAMAR in France), equipped by a Seabird CTD, performed a total of 873 cycles from the surface to approximatevely 500 m depth with a sampling rate of 4 seconds. Only a part of these cycles (during the 11 first days of the mission) was dedicated to sample the WAG and have been studied herein (Figure 2). In this paper, only the classification of water masses in the WAG and the ambient environment will be discussed.

Isotherms and isohalines sketched hereafter in all vertical sections as continuous lines, represent the interpolated temperature and salinity on a grid of (horizontal and vertical) resolution dx=1.1km and dz=1m. The interpolation is performed using the optimal spatial kriging. In order to remove high frequencies, the interpolated data were smoothed using a gaussian filter with a width corresponding to the radius of deformation in the studied region (Bosse et al., 2015). The parameter $p \in \{T,S\}$ of each transect is transformed in a smoothed parameter $\tilde{p} \in \{\tilde{T}, \tilde{S}\}$ by a convolution product:

$$(1) \qquad \tilde{p}(x,z) = \int_{x_{min}}^{x_{max}} p(x,z) \times \exp \frac{-x^2}{2L^2} dx$$

Where x is the distance along the section, x_{min} and x_{max} the section limits, z the depth and L the standard deviation. Taking $L = 15$ km is sufficient to conserve the signal linked to the WAG.

In addition to the glider data described previously, and in order to further test the performance of our method, more examples of data and their related classification results are represented in Appendix A.

2.3. Data single-labeling

To build a training dataset with a unique labeling, each sample of the database has been attributed to a water mass from those described in the introduction {SAW, NACW, MAW, WIW, LIW, TDW, WMDW}. To keep a clear physical sense, the approximate boundaries between the water masses, have been defined manually by specifying polygons in the θ-S plane (Figure 12a). the separation interface has been characterized in such a way to present the water masses as objectively as possible on the basis of the values of θ, defined in the various studies cited in the introduction. The large seasonal variability of SAW, the intermittency of NACW as well as the occasional direct mixing of dense MWs with AW were taken into account during this process. Then each sample labeled on the θ-S plane is projected into the coordinate system, potential density and potential spicity (σ-π) (Figure 12b). The reason why the labeling was not directly done on the (σ-π) diagram is explained by the fact that the characteristics of the water masses in the study area are defined in previous studies via the θ-S diagram and that the equivalent potential spicity properties will only be deduced after the projection of the labeled samples into the (σ-π) plane.

The constructed training dataset is therefore $A = \{(\sigma_s, \pi_s, \lambda_s)\}_{s=1}^{N}$ where σ_s, π_s and λ_s are respectively the potential density anomaly, the potential spicity and the water mass label of a sample s at a given longitude, latitude and depth. The choice of the (σ-π) diagram for this classification study is justified in the next part of this section (2.4.3). It's worth mentioning that the terminologies of spicity and spiciness are used by several authors with different definitions to describe a 'spice' type variable in physical oceanography. Some authors have chosen to derive such a variable, called potential spicity, so that its contours are orthogonal to those of potential density (Veronis, 1972; Huang et al., 2018). Other studies are based on the non-orthogonal functions, called spiciness (Jackett and McDougall, 1985; Flament, 2002; McDougall and Krzysik, 2015). In our study, the potential spicity (π) is calculated on the basis of its definition as a function whose contours are orthogonal to those of the potential density (Huang et al., 2018) via the MATLAB subroutine gsw_pspi(SA, CT, pr), also provided by (Huang et al., 2018), where (SA, CT, pr) is the absolute salinity ($g.8kg^{-1}$), conservative temperature (°C) and reference pressure (db) (`https://github.com/lanlankai/Spicity-JGR`). The pressure value of pr = 0 (the sea surface pressure) was taken as a reference level. Another remark concerns the WIW and WMDW : No traces of these two water masses were detected in the glider transects. In these cases, they will be excluded from the training dataset to avoid the distortion of the classification results. The choice to eliminate WMDW from the study was based on the 12.85 °C potential temperature isoline used by (Millot, 2014) as an unambiguous definition of WMDW. Thus $\lambda_s \in \{SAW, NACW, MAW, LIW, TDW\}$. In analogy with Millot (2009) and Millot (2014) we differentiate hereafter, for convenience, a lower-TDW from an upper-TDW that will behave more like WMDW and LIW, respectively.

2.4. K nearest neighbors classification

2.4.1. Problem statement

The classification using the nearest neighbor search (Cover and Hart, 1967; Fix and Hodges, 1989) is a well known decision procedure, non parametric for automatic learning. It is used in this study to evaluate the presence and prevalence of each water mass sampled by the glider in the different transects. This method has been considered as one of the widely used classification algorithms owing to its simplicity and straightforward implementation. However, it has few shortcomings affecting its accuracy of classification (Gallego et al., 2022; Gou et al., 2022) which are discussed in sections 2.4.2 and 2.4.3. This classification technique has an objective of classification and attribution to a request point q belonging to a sample of observations Q, the class of the instance of training of the nearest neighbor based on a metric that define the similarity between observations and classes of a training dataset A. Moreover, it is useful to consider more than one neighbor, so the technique is more commonly referred to as K nearest neighbors (Knn) classification where the K nearest neighbors are used to determine the class (Cunningham and Delany, 2007). Figure 3 visualizes the overview scheme for the proposed K nearest neighbors classification of water masses.

We suppose a supervised learning set of data A $= \{(\sigma_s, \pi_s, \lambda_s)\}_{s=1}^{N}$ as described previously. In the training step, the dataset A is simply stored with any explicit learning. In the inference step, for each request instance q belonging to the dataset Q $= \{(\sigma_j, \pi_j)\}_{j=1}^{M}$, a Knn search is done to get the K closest instances $N(\sigma_s, \pi_s) = \{(\sigma_s^{(i)}, \pi_s^{(i)}, \lambda_s^{(i)})\}_{i=1}^{k}$ which are the nearest to q on the basis of the metric d. Therefore, the predicted water mass label λ_p is obtained using a weighted combination of labels $(\lambda^{(i)}|^{(i=1..k)})$ based on the d metric as follows:

$$\text{(2)} \qquad \lambda_p = f(q, A) = \frac{\sum_{i=1}^{k} d^{-1}(q,(\sigma^{(i)},\pi^{(i)})).\lambda^{(i)}}{\sum_{i=1}^{k} d^{-1}(q,(\sigma^{(i)},\pi^{(i)}))}$$

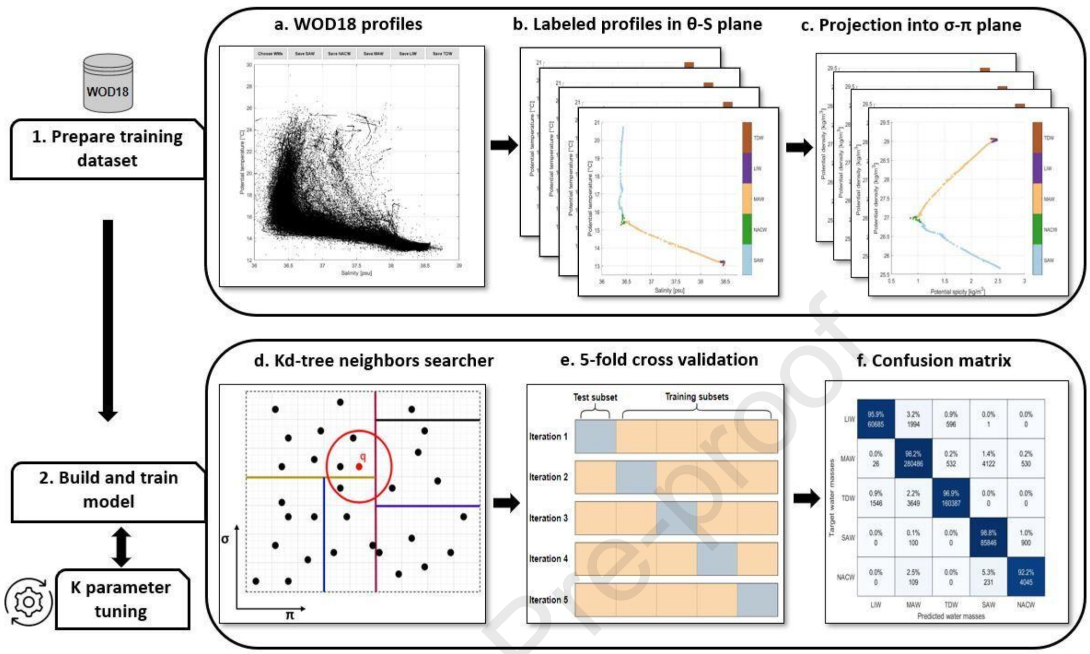


Figure 3: Flowchart of the proposed K nearest neighbors classification of water masses.

Thus, a Knn instance with a smaller distance will contribute more to the prediction for the instance.

In addition to classifying water masses into different categories, we can quantify the fraction of a given water mass λ_q in a request sample q belonging to Q as follows :

$$F_{q,\lambda_q} = \frac{\sum_{i=1}^{m} d^{-1}(q,(\sigma^{(i)},\pi^{(i)}))}{\sum_{i=1}^{k} d^{-1}(q,(\sigma^{(i)},\pi^{(i)}))} \quad (3)$$

where m is the number of samples representing the water mass λ_q among the K nearest neighbors on the basis of the metric d. Distances have been normalized to have all them lying between 0 and 1.

Such a quantification is helpful to supplement the information displayed in figures like 15b or 16b.

2.4.2. K nearest neighbors search

The simplest solution to the problem stated before remains on computing the d metric between the request point q and each point of the dataset and to return the k nearest points on the basis of d. The computing complexity is O(N × L), where N is the size of the data ensemble and L is its dimension (herein L=2). This method can be costly due to the huge amount of data. Within this context, numerous studies have been concerned with finding new approaches that are efficient with computations through employing fast search algorithms or using a training dataset size reduction scheme (Ougiaroglou and Evangelidis, 2016; Hou et al., 2018; Gallego et al., 2022). In our study, this drawback was overcome by searching for the K nearest neighbors using the spatial K dimensional tree subdivision structure (Kd-tree) (Chen et al., 2019). The latter is a well known optimisation for the Knn algorithm convenient for reduced dimensional spaces. The points ensemble N is divided recursively in the 2D space (σ-π) into a binary tree with N levels and log(N) depths.

This division continues until reaching at least a well defined number of points for each node. Therefore, the K nearest neighbors search for a point with a given request is done following these steps:

1. The determination of the node to which the query point belongs.
2. The search of the closest K points within that node on the basis of the metric d.
3. The determination of all other nodes having any area that is within the same metric d, in any direction, from the query point to the K^{th} closest point on the basis of the metric d.
4. The search of the closest K points within those nodes on the basis of the metric d.

2.4.3. Parameter definition and performances analysis

The Knn performances are known to be sensitive to choices of the metric and the parameter K which depend on the data characteristics (Jiang et al., 2007). Therefore, they must be chosen appropriately to improve the classification performances. The metric selection can affect the form, the volume and the orientation of classes because some data points can be close for a metric and distant for another one. A small parameter K can capture a local structure in the data and therefore the result can be sensitive to noise, however a larger K permits to capture the global structure of data and suppress the noise effect but consumes more memory (Ghosh, 2006; Kang, 2021).

In this study, the chosen distance metric for the query points categorization is the Euclidean distance. Therefore, the metric d in equations 2 and 3 has the following form:

$$d(q,(\sigma^{(i)},\pi^{(i)})) = \sqrt{(q-(\sigma^{(i)},\pi^{(i)}))^2} \tag{4}$$

As mentioned in section 2.3, we choose the definition of potential spicity proposed by(Huang et al., 2018) who attempted to rehabilitate in the least square sense, the (Veronis, 1972) form of orthogonality between this variable and potential density. Thus, the choice of σ-π coordinates system instead of the traditional θ-S diagram is justified by the orthogonality and the dimensional homogeneity of these two pairs (σ and π). This allows a precise and concise measure of the distance d compared with θ-S diagram (Huang et al., 2018; Gao et al., 2020). Also, despite the existence of numerous techniques of data scaling, many authors have shown the impact of these techniques on the stability of machine learning algorithms performances as it is the case for the Knn (Ambarwari et al., 2020; Shahriyari, 2017). Furthermore, one of the primary challenges is selecting the most suitable method for scaling. The latter problem is avoided here since σ and π share almost the same range.

Indeed, the major difference between the use of the two aforementioned diagrams is to determine boundaries between water masses. In our case, the previously described labeling method makes it possible to reduce this difference to 1.2%. However, to show the advantage of σ-π diagram for the computation of distance on which our method is based; the samples of the training dataset forming the labeled boundaries of water masses have been eliminated in order to construct separate water masses in the two spaces σ-π and θ-S (figure 4). This situation represents the case of non-continuity of the training dataset or the case of difficulty to determine the boundaries characteristics between the water masses in a subjective way.

Taking as reference the classification results of the two transects (section 3.2), we computed the total percentage of samples that changed membership from one water mass to another for the two diagrams. The results sketched in figures 5 and 6 represent a total difference of 10.38% for σ-π diagram versus 17.7% for θ -S diagram. Therefore, using our methodology of classification, σ-π diagram is more appropriate for water masses frontiers determination. Also, the difference in distance calculation between the two spaces σ-π and θ-S is clearly visible in the computation of the fraction of a given water mass in a given sample (figures 7 and 8).

Concerning neighborhood size K selection, several methods have been developed with a view to predict its optimal value and to overcome its sensitivity (Zhongguo et al., 2017; Zhang et al., 2018; Gou et al., 2019, 2022). In our case, the choice of the parameter K was based on the traditional L-Fold Cross Validation method (Paik and Yang, 2004; Ghosh, 2006; Kang, 2021). This validation technique is based on estimating an accuracy rate for different values of K and select the one that induces the smallest classification error rate. The latter have been illustrated using the confusion matrix (Provost and Kohavi, 1998).

Indeed, the L-Fold Cross Validation consists in splitting the dataset $A = \{(\sigma_s, \pi_s, \lambda_s)\}_{s=1}^{N}$ in L independent subsets randomly selected with quasi-constant sizes. A subset is used to validate the produced model with the help of the L-1 remaining subsets. This process is applied L times so that each subset is used exactly one time for the validation. The classification error rate for all the partitions L is defined by a set $A' = \{(\sigma_s, \pi_s, \lambda_s)\}_{s=1}^{N'} \in A$, as:

$$\tau = \frac{1}{N'} \sum_{i=1}^{N'} I_i \left(\lambda^{(i)} \neq \lambda_p^{(i)} \right) \qquad (5)$$

where

$$I_i\left(\lambda^{(i)} \neq \lambda_p^{(i)}\right) = 0, \text{if } \ \lambda^{(i)} = \lambda_p^{(i)}$$
$$= 1, \text{otherwise}$$

Two popular choices of L are 5 and 10. In this study, we fix L to 5.

In practice, the total classification error rate for the set $A = (\sigma_s, \pi_s, \lambda_s)^{N_{s=1}}$ is deduced from the confusion matrix. This matrix illustrates not only the algorithm errors but also how the classification algorithm works for each class (Markoulidakis et al., 2021). Indeed, the confusion matrix is a cross table where each column represents the predicted class instances, and each row represents the real class instances. The classes $\lambda_s \in \{SAW, NACW, MAW, LIW, TDW\}$, are listed in the same order in the rows and columns, so the correctly classified elements are located on the main diagonal. During the cross validation L-fold, if the predicted class of the test sample is correct, then the diagonal element of the confusion matrix is incremented by 1. However, if the predicted class is incorrect, then the element off diagonal is incremented by 1. Once, all the training samplings are classified, the classification error rate is based on the ratio of the number of sampling incorrectly classified and the total number of classified samplings.

Numerous evaluations of K between 10 and 100 recorded classification error rates between 2% and 2.2%. The parameter k=51 seems to be a good compromise between the complexity and precision of computation. The multi-classes confusion matrix, a matrix of 5×5 dimension, relative to this value of K is sketched in Figure 9. This matrix is build from the cross validation 5-fold applied to a total number of sampling N=604855. The classes SAW; MAW; TDW record an accuracy beyond the total accuracy of 98%. For the case of the NACW, the algorithm classifies incorrectly almost 4% of the training samplings between the surface SAW and subsurface MAW layers. The LIW record the highest classification error rate. Indeed, 8% of the sampling that are supposed to belong to this class were confused with the classes MAW; TDW. The reason behind this is the relatively tight ans sinuous relationship between the LIW, MAW and TDW classes in regards to the θ-S and σ-π diagrams.

It is mentioned that the experimental environment of model building was performed on a computer with an Intel i7-1165G7 CPU @2.80 GHz with 8 GB memory. For a total number of sampling N=604855 forming the training dataset, prediction speed was 44000 observations per second and the total training time was 72.03 seconds.

2.5. Sensitivity of the method

A sensitivity analysis was performed to assess the impact of the spatio-temporal distribution of the hydrological profiles forming the training dataset. This is achieved by computing the percentage of samples that move from one water mass to another, when spatio-temporal variability is reduced in the training dataset. The classification results of the two transects, using all profiles of the database are taken as reference.

Regarding the spatial sensivity and as the distribution of MWs in the Alboran Sea mainly depends on latitude (e.g the presence of LIW in the northern 2/3 of the basin), the area has been divided into two regions separated by latitude 35˚45'N. Hydrological profiles of each region were used separately as a training dataset to examine the impact of database spatial distribution on the classification results of the proposed method. The results of this analysis is presented in the following section.

The temporal sensitivity was examined to evaluate the impact of the temporal ranges of the training dataset. This is achieved by dividing the training data into profiles acquired during four periods, from 1950 to 1980, from 1950 to 1990, from 1950 to 2000 and from 1950 to 2010. Ocean profiles related to each period were used separately as a training dataset. The confusion matrices computed for these four cases (figure 10) showed that no significant changes occur. Therefore, the classification results are not altered by the temporal ranges of the training data.

Also, temporal sensitivity of the seasonal variability of SAW was performed. The training dataset was divided into profiles collected during fall, winter, spring and summer seasons. The confusion matrices computed for these four cases showed less than 2% of difference between predicted SAW samples using the whole dataset and those predicted by using the separate seasonal data. Thus, seasonal variability of SAW does not influence the classification results.

3. Results

3.1. Water masses labeling in the σ-π plane

All the observation of the potential temperature and salinity obtained from the database used to build the training dataset and reaching a maximum depth of 700 m are sketched in figure 11. The seven water masses previously described can be distinguished as follows: the SAW are the lightest and characterized by a salinity layer quasi-homogeneous subject to intense seasonal variability and a constant temperature vertical gradient. The NACW is below the SAW and characterized by a salinity minimum with $\theta - S$ between 14°C-36 psu and 16°C-36.4 psu. Under the Atlantic Waters, the $\theta - S$ diagram shows a linear stripe limited by the isopycnals $\sigma \simeq 27.2$ kg.m^{-3} and $\sigma \simeq 28.8$ kg.m^{-3}. These values characterize the MAW resulting from the mixing between the Atlantic and Mediterranean Waters.

Beyond a salinity of 38 psu, the $\theta - S$ diagram is characterized by a tight and sinuous relationship, representing more than 80% of the total water volume. During its presence, the WIW is clearly noticed by its local temperature minimum (13°C-13.1°C, 38.25-38.35) linking the AW and the LIW. This water mass is characterized by a temperature and salinity local maximum shown by the $\theta - S$ diagram, with salinity values up to 38.58. The TDW is represented by a curved line linking the LIW and WMDW. This water mass is clearly indicated by its low temperature ($<$ 12.9°C), its relatively low salinity ($\simeq$38.4) and its high density ($\simeq$29.09 kg.m^{-3}).

Figure 12a sketches a part from the labelled training data in the coordinates system $\theta - S$. The labels are the water masses SAW ; NACW ;MAW ;LIW ;TDW. The equivalent result is projected on the $\sigma - \pi$ plan as shown in Figure 12b.

The latter shows the water masses characteristics which are clearly identified through the analysis of potential spicity.

The general aspect of the water masses in the $\sigma - \pi$ plan are perceived as a rotation transformation of the $\theta - S$ plan around the origin with an angle $\alpha = 45^{\circ}$ (Figure 12b). Indeed, the Atlantic waters (SAW, NACW and MAW) keep a geometric aspect of a curve as an elbow. The inflexion point of this curve represents the interface between the surface waters (SAW) and those of the subsurface (MAW). These waters are characterized by a linear relation defined by positive and negative coefficients respectively. The NACW reveals a potential spicity minimum $\pi = 0.45$ kg.m^{-3}. The Mediterranean waters (LIW and TDW) keep the aspect of broad relationship where the LIW is characterized by a local maximum of potential spicity $\pi = 2.58$ kg.m^{-3}, equivalent to a local maximum of salinity $S = 38.52$ psu.

3.2. Glider transects classification

As mentioned in section 2.2, the first days of the glider profiling were dedicated to the survey of Moroccan Mediterranean waters offshore and more precisely of the Western Alboran Gyre (WAG), located between the strait of Gibraltar and the Tres Forcas cape. This quasi-steady anticyclonic gyre has a typical diameter of approximately 100 km and a depth of 200 m and represents the most intense dynamical structure of the mean circulation in the western Mediterranean sea, with surface currents reaching 1.5 m s^{-1} (Álvaro Viúdez et al., 1996; Vélez-Belchı et al., 2005; Flexas et al., 2006).

The vertical profiles of temperature and salinity acquired between 11 and 22 February 2021 during the first and second transects are represented in Figures 13 and 14 respectively.

The warm and fresh anomalies characterizing the WAG appear noticeably in the temperature and salinity fields. Globally, the vertical distribution of T is characterized by decreasing values with depth and by a sharp vertical gradient. The WAG core highlights temperature values higher than 15°C and positive anomaly compared to the ambient environment. The latter results in a deepening of the isothermal layers by several tens of meters inside the WAG and an upwelling of these layers outside the WAG.

Despite the relatively long time period sampling of both transects ($\sim$ 7 days for the first transect and $\sim$ 5 days for the second transect), we consider a quasi-synoptic situation, highlighting the water mass composition during this period of time in the studied region. Therefore, we only consider the spatial mixing variability of the water mass.

The classification methodology applied on both glider transects (Figures 15b and 16b) shows that the AW engulfs the top layer (from surface to 200-250m) of the Alboran Sea, just below the isopycnal $\sigma = 28.9$ kg.m^{-3}. This layer is characterized by a density anomaly which is the result of temperature and salinity anomalies. The isohaline light layer (S $<$ 36.6 psu, sigma $<$ 27 kg.m^{-3}) is classified as a SAW in both transects. However, only the second transect outlines the presence of the NACW in its southern side near the Moroccan coast as highlighted by a spicity minimum (Figure 16b).

Beyond the isopycnal $\sigma = 28.9$ kg.m^{-3}, we found the MWs adjoining the WAG and containing the LIW and the TDW. The LIW layer is absent in the south, near the Moroccan coasts, and is principally concentrated in the center and the north of the Alboran Sea; where it thickens. The TDW is mainly present along the two transects from the south to the north. Neverthless, we distinguish an upper TDW which is found just below the LIW and a lower TDW which is located in the southern side below the AW. The $\sigma - \pi$ diagrams (Figures 15a and 16a) outline the distribution of the TDW : In the southern part of the transect, the dense MWs (herein the lower TDW) are individually mixed with the AW leading to a relatively straight shape with some bending in the deep part of the profiles. However, far away from the African coasts the MWs are overlapped and slightly mixed leading to a sinuous shape in which the upper TDW tends to connect the LIW with dense MWs.

Thus, the classification of glider transects shows: (i) the formation of the WAG by the newly flushed AW, (ii) The presence of the LIW in the 2/3 North of the region, (iii) the presence of an upper TDW below the LIW and (iv) the presence of a lower TDW in the southern side below the AW.

3.3. Training dataset sensitivity

In the case where only the profiles gathered below 35°45'N are used as training dataset, the confusion matrix (Figure 17a) shows that the TDW is very well classified while almost 30% of LIW samples move to TDW (26.2%) and MAW (2.5%). The results of the new classification applied to the first glider transect (Figure 18a) show that the algorithm captures the uplift of dense MWs in the southern part of the basin and that the LIW layer is still present in the northern 2/3 of the transect but the latter becomes less thick. In the case where only the profiles gathered beyond 35˚45'N are used as training data, the confusion matrix (Figure 17b) shows that the LIW regains about 10% of its samples compared to in the previous case, while almost 30% of TDW samples move toward MAW (20.2%) and LIW (8.7%). The results of the new classification applied to the first transect of the glider are shown in Figure 18b. The spatial distribution of the LIW is close to that relating to the use of the entire database. However, the uplift of dense MWs in the south is not well represented. Similar tests were carried out on the data from the second transect with similar results.

3.4. Clustering analysis

To show the advantage of our method compared to those of unsupervised classification, a cluster analysis based on the iterative algorithm k-means (Appendix B), classically used to specify water masses characteristics (Roseli et al., 2015; Molleri et al., 2010), was applied on the $\sigma-\pi$ diagram to classify the water mass in the both transects. As the k-means is also based on distance computation, we choose the $\sigma - \pi$ coordinate system to allow a concise measure of this distance. The similarity between the samples and the centroids of the clusters (selected randomly in the first step) is indicated by the euclidean distance defined as in equation 3. The results obtained and the related analyzes being the same for the two transects, we limit ourselves to the presentation of those relating to the first transect. To ensure that the chosen number of clusters, k, is representative of the system, different values of k between 2 and 5 were tested (Figure 19). The silhouette method (Appendix B) is used as the tool to validate the clustering quality and to see how well each sample lies within its cluster. In this test, the positive silhouette value nearest to one, indicate cases where a sample is well clustered. Samples with negative silhouette values are considered as poorly classified.

The classification results show that the clustering analysis performs well to distinguish the Atlantic and Mediterranean Waters for a k value greater than or equal to 3 (Figures 19b, 19c and 19d). The separating interface between these waters is formed by the 28.8 kg.m^{-3} isopycnal, which is approximately equal to the value found by our method ($\sigma \sim 28.9$ kg.m^{-3}). However, when different AWs are considered, only the SAW can be distinguished by the algorithm. The MWs (LIW, upper and lower TDW) are inherents and this results in a single layer above the 28.8 kg.m^{-3} isopycnal.

The silhouette values show that when water masses are divided into 2 clusters (Figure not shown) all the samples are correctly clustered, showing a positive and significant silhouette values greater than the mean value (0.96 in this case). Nevertheless, for the other number of clusters (k=3, 4 and 5), several samples are wrongly grouped with a negative silhouette values (Figures not shown). however in all this cases, clusters number 1 and 2 still well classified and this explain the fact that (i) the k-means performs well for k $\geq$ 3 in distinguishing the Atlantic and Mediterranean Waters and that (ii) in the AWs, only the SAW is well defined.

4. Discussion

4.1. The hydrographic structure of the WAG

The hydrographic structure of the WAG sampled through the glider shows that the gyre vertical extent (∼ 180 m) is characterized by a large homogeneous layer in salinity with values lower than 36.6 psu. These results are in agreement with other cruises that sampled the WAG at its usual location (as in our study), in this case (Álvaro Viúdez et al., 1996; Nibani et al., 2021). Using data from an intensive field experiments, these authors recorded the same salinity characteristics of the isohaline layer, occupying the upper part of the gyre and reported a typical vertical extension of 180-200m. However, in comparison with oceanographic cruises coinciding with the eastward migration event of the WAG (Vélez-Belchı et al., 2005; Flexas et al., 2006), the salinity within the gyre is higher than that found in our study (up to 0.2 psu). This difference is explained by the fact that in its usual location, the WAG is more exposed to inputs of fresh AW than when it is located further east. Moreover, these authors recorded a reduced vertical extensions of the WAG of 130-150m.

4.2. The classification results

As shown above, the water in the study region is classified into 5 types via a Knn classification method based on σ-π diagram. The labeling process result shows that the characteristics of the different water masses evolving in the Alboran Sea can be clearly identified through the σ-π coordinate system. Indeed, each water masses represents a physical property and a geometric aspect that correlates with that of the traditional $\theta - S$ diagram but which can be studied from a different angle.

The classification results obtained for the AW show that the latter is not sensitive to the spatio-temporal variability of the training dataset. The core of the WAG marked with a large vertical thickness of homogeneous salinity layer (<36.6 psu), is principally generated by the SAW. This result is in good agreement with previous studies (Álvaro Viúdez et al., 1996; Vélez-Belchı et al., 2005;

Flexas et al., 2006). These authors show through a three dimensional description of he Western Alboran Sea that the WAG is characterized by recent AW transported from the Strait of Gibraltar into the core of the gyre and occupying a considerable part of it. The 28.9 kg.m^{-3} isopycnal, found as separating interface between AW and MWs, corresponds to that deduced from the $\theta - S$ diagram analysis by Gascard and Richez (1985) in their study of water masses and circulation in the western Alboran Sea. The no significant NACW samples detected during the second transect can be interpreted as points being closer to NACW than to any other water mass (in this case SAW and MAW) and not as samples marking the pure NACW.

The vertical distributions result of the MWs in both transects is sensitive to the spatial variability of the training dataset. By using the whole labeling data, the obtained result is in agreement with those inferred from the expert analysis of the $\theta - S$ diagram. The spatial distribution of the LIW layer that thickens from the north to the south corroborates with the works of Parrilla et al. (1986); Millot (2014). These authors found that the properties of the LIW is quite recognizable in most of the Alboran Sea, except in the southernmost part near the Moroccan coasts. They showed that the path of the LIW in the Alboran basin did not cross south of 35˚30'N. The LIW limits obtained by our classification method are 35˚30'N and 35˚45'N for the first and second transect respectively. The spatial distribution of the TDW (the upper TDW and the lower TDW) is in total concordance with the studies of Millot (2009, 2014) and highlights the direct link between the deep MWs and the AW in the southern side of the two transects. Indeed Millot (2009, 2014), shows through a $\theta - S$ diagram analysis of zonal hydrographic transects that, in southern part of the Alboran Sea, dense MWs mixes directly with AW.

4.3. Comparison with clustering analysis

The comparison between the method adopted in this paper and the k-means algorithm shows that this latter can not distinguish water masses when several AWs and MWs are considered. In fact, the uplift of the dense MWs in the southern part and the presence of the LIW in the $\frac{2}{3}$ parts of the northern basin can not be outlined by the clustering analysis. This comparison corroborates the performed analysis by Cheng et al. (2014); Millot (2019). Indeed, Cheng et al. (2014) shows that in a well-defined range of potential density, water masses having similarities in temperature and salinity are inseparable by the clustering analysis. Millot (2019) shows that the method proposed by Naranjo et al. (2015) is rather a computation of euclidean distances between the samples and a set of centroids representing the water masses than a clustering analysis. He concludes that, in regions of relatively moderate mixing processes such as in the Strait of Gibraltar, a subjective (θ-S) diagram analysis based on a traditional method where boundaries of water masses are defined by experts experience, is much more robust than clustering analysis.

Thus, conventional cluster analysis are not always appropriate to discriminate water masses and there is no clear physical meanings of the water masses boundaries. As the labeling process guides the decision of the algorithm towards the choice of the water mass representing each sample, our methodology retains a part of this physical meaning through a labeling approach based on a traditional method for defining boundaries between water masses.

5. Conclusions

The objective of this study was to identify the spatial distribution of water masses in the Western basin of the Alboran Sea. To do this, a novel methodology based on water masses automatic classification using the Knn search was applied to the T-S data acquired by a glider. These data have been projected on the orthogonal and dimensional homogeneous coordinates system: potential density anomaly-potential spicity (σ-π).The parameters used in this algorithm have been selected in order to get the most accurate classification. The latter have been insured by a supervised machine learning process based on available data from the World Ocean Database 18 and the Global Data Assembly Center (Figure 2). From all the water masses described in section 1, the WIW and the WMDW were not successfully detected by the glider and therefore were excluded from the training dataset. Thus, the water masses in the glider transects were classified in 5 categories: SAW; NACW; NAW; LIW and TDW.

In comparison to the classic method of classification based on clustering analysis (herein the k-means), the proposed method in this paper permits to ascertain the water masses frontiers with a reasonable and robust approach. In the studied region, the classification results are in good agreement with the circulation schemes established in previous studies and inferred from the traditional method based on the subjective expert analysis of the (θ-S) diagram, showing:

- The formation of the WAG by the recently advected Atlantic Water (Álvaro Viúdez et al., 1996; Flexas et al., 2006) ;
- The uplift of the dense Mediterranean Waters (the lower-TDW) near the Moroccan coasts (Millot, 2009, 2014) ;
- The presence of the LIW in the 2/3 North of the Western basin of the Alboran Sea (Parrilla et al., 1986; Millot, 2009).

The application of our approach for ocean water masses classification has many advantages. By combining traditional method based on expert analysis and Machine learning technique, this methodology is useful and appropriate to automatically classify water masses in regions where intense mixing occurs such as the Western Alboran Sea. Although the labeling process requires the knowledge of the water masses characteristic in the study area, the adaptation of this technique to other regions is easy and straightforward. Indeed, this methodology can be applied easily to other sub-basins or marginal seas as long as a sufficient number of in-situ observations describing the whole spatio-temporal variability of the area can be provided as a training dataset.

The speed of the proposed method will make it possible on the basis of basic hydrographic data collected during typical research cruises or autonomous systems, to provide classification results in real time. Remarkably, Using the proposed methodology, researchers non-particularly specialists in oceanography, can take advantage of previous knowledge of water masses characteristics validated by experts to solve the problem of water masses classification. Within this context, a Graphical User Interface (GUI) is under development in order to enable users performing the entire process described in this manuscript (figure 3), within all ocean basins.

Acknowledgment

The authors gratefully acknowledge Houssini Nibani, President of AGIR Association who has performed this glider mission in the Western Aboran Sea in the framework of the European project ODYSSEA. We are also grateful to Laurent Beguery and Orens Fommervault from Alseamar France For their cooperation and their answers to the questions concerning the glider data collected in this region. Thanks to the anonymous reviewers for their constructive comments and helpful suggestions.

Appendix A. Examples of using K Nearest Neighbors Classification to study the spatial distribution of water mass in the Western Alboran Sea.

Other examples of data acquired over different time period have been used to study the spatial distribution of water mass in the Western Alboran Sea (figure A.20). It's about:

- A CTD transect of a field experiment acquired in September 1992 on board of the R/V Garcia del Cid (Álvaro Viúdez et al., 1996). Being available on WOD18, these data have been removed from the training dataset to assess the results of the classification in a more objective way.
- A hydrographic (CTD) cast of an intensive oceanographic survey (BIOMEGA) collected on board of the Spanish R/V Garcia del Cid during October 2003 (Flexas et al., 2006). Data were provided through SeaDataNet Pan-European infrastructure for ocean and marine data management (`https://www.seadatanet.org`);
- A glider transect (from 11 to 17 November 2020) of the first mission performed by the Moroccan association AGIR (Nibani et al., 2021), as part of the European project ODYSSEA (`https://odysseaplatform.eu/fr/home-fr/`).

The temperature and salinity fields, and the classification results obtained for the three aforementioned oceanographic cruises, are sketched in figures A.21, A.22 and A.23. This leads to the same interpretation of glider data previously described in section 2.2.

Appendix B. k-means clustering and silhouette method.

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem (Kaufman and Rousseeuw, 1990). It is an iterative, data-partitioning algorithm which aims to partition n observations into k groups, called clusters. The algorithm proceeds as follows :

1. Select k initial centroids at random after indicating the desired k number of clusters ;
2. Compute sample-to-cluster-centroid distances of all observations to each centroid and then assign each observation to the cluster with the closest centroid ;
3. Compute the average of the observations in each cluster to obtain k new centroid locations;
4. Repeat steps 2 and 3 until cluster assignments do not change, or the maximum number of iterations is reached.

k-means aims at minimizing an objective function that depends on the distance of the data points to the cluster centroids. Suppose $D = \{x_1, ..., x_n\}$ is the dataset to be clustered. K-means problem can be expressed as follows :

$$\min \sum_{k=1}^{K} \sum_{x \in C_k} f(x, c_k) \tag{B.1}$$

The function 'f' computes the distance between object x and centroid c_k which is defined by:

$$c_k = \sum_{x \in C_k} \frac{x}{n_k} \tag{B.2}$$

where n_k is the number of data objects assigned to cluster C_k.

To evaluate the clustering analysis quality, (Rousseeuw, 1987) introduced the so-called silhouette method. This technique provides a graphical representation which helps the user to select the number of clusters and to see how well each sample lies within its cluster. The silhouette value for each sample is a measure of how similar that sample is to other samples in the same cluster, compared to samples in other clusters. The silhouette value s_i for the i^{th} sample is defined as :

$$s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \tag{B.3}$$

where a_i is the average dissimilarity of the i^{th} sample with all other data within the same cluster and b_i is the minimum average dissimilarity of the i^{th} sample to samples in a different cluster. Distance metric is employed to calculate the dissimilarity between samples. When a cluster contains only a single sample, it is unclear how a_i should be defined and then s_i is set to 1.

Indeed, from the preceding definition, it is clear that $-1 \leq s_i \leq 1$ for each sample i. A high and positive value indicates that the sample is well matched to its own cluster, and distant from neighboring clusters. A low or negative silhouette value, correspond to cases in which samples are assigned to wrong clusters.

References

Ambarwari, A., Jafar Adrian, Q., Herdiyeni, Y., 2020. Analysis of the effect of data scaling on the performance of the machine learning algorithm for plant identification. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi) 4, 117 – 122. URL: https://jurnal.iaii.or.id/index.php/RESTI/article/view/1517, doi:10.29207/resti.v4i1.1517.

Argo, 2021. Argo float data and metadata from global data assembly centre (argo gdac). URL: https://www.seanoe.org/data/00311/42182/, doi:https://doi.org/10.17882/42182.

Bosse, A., Testor, P., Mortier, L., Prieur, L., Taillandier, V., d'Ortenzio, F., Coppola, L., 2015. Spreading of levantine intermediate waters by submesoscale coherent vortices in the northwestern mediterranean sea as observed with gliders. Journal of Geophysical Research: Oceans 120, 1599–1622. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014JC010263, doi:https://doi.org/10.1002/2014JC010263, arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2014JC010263.

Boyer, T., Baranova, O., Locarnini, R., Mishonov, A., Grodsky, A., Paver, C., Weathers, K., Smolyar, I., Reagan, J., Seidov, D., Zweng, M., 2019. World ocean atlas 2018 product documentation ocean climate laboratory ncei / nesdis / noaa noaa national centers for environmental information. doi:10.13140/RG.2.2.34758.01602.

Broecker, W., 1991. The great ocean conveyor. Oceanography URL: https://doi.org/10.5670/oceanog.1991.07.

Bryden, H.L., HL, B., HM, S., 1982. Origin of the mediterranean outflow .

Chen, Y., Zhou, L., Tang, Y., Singh, J.P., Bouguila, N., Wang, C., Wang, H., Du, J., 2019. Fast neighbor search by using revised k-d tree. Information Sciences 472, 145–162. URL: https://www.sciencedirect.com/science/

article/pii/S0020025518307126, doi:https://doi.org/10.1016/j.ins.2018.09.012.

Cheng, G.S., Sun, J.D., Zu, T.T., Chen, J., Wang, D.X., 2014. Analysis of water masses in the northern south china sea in summer 2011. Journal of Tropical Oceanography 33, 10–16.

Clark, P., Pisias, N., Stocker, T., Weaver, A., 2002. The role of the thermohaline circulation in abrupt climate change. Nature 415, 863–9. doi:10.1038/415863a.

Cochrane, J.D., 1958. The frequency distribution of water characteristics in the pacific ocean. Deep Sea Research (1953) 5, 111–127. URL: https://www.sciencedirect.com/science/article/pii/0146631358900029, doi:https://doi.org/10.1016/0146-6313(58)90002-9.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13, 21–27. doi:10.1109/TIT.1967.1053964.

Cunningham, P., Delany, S.J., 2007. k-nearest neighbour classifiers.

Fix, E., Hodges, J.L., 1989. Discriminatory analysis. nonparametric discrimination: Consistency properties. International Statistical Review / Revue Internationale de Statistique 57, 238–247. URL: http://www.jstor.org/stable/1403797.

Flament, P., 2002. A state variable for characterizing water masses and their diffusive stability: Spiciness. Progress in Oceanography 54, 493–501.

Flexas, M., Gomis, D., Ruiz, S., Pascual, A., Leon, P., 2006. In situ and satellite observations of the eastward migration of the western alboran sea gyre. Progress in Oceanography 70, 486–509. URL: https://www.sciencedirect.com/science/article/pii/S0079661106000590, doi:https://doi.org/10.1016/j.pocean.2006.03.017. gabriel T. Csanady: Understanding the Physics of the Ocean.

Gallego, A.J., Rico-Juan, J.R., Valero-Mas, J.J., 2022. Efficient k-nearest neighbor search based on clustering and adaptive k values. Pattern Recognition 122, 108356. URL: https://www.sciencedirect.com/science/article/pii/S0031320321005367, doi:https://doi.org/10.1016/j.patcog.2021.108356.

Gao, Y., Huang, R.X., Zhu, J., Huang, Y., Hu, J., 2020. Using the sigma-pi diagram to analyze water masses in the northern south china sea in spring. Journal of Geophysical Research: Oceans 125, e2019JC015676. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JC015676, doi:https://doi.org/10.1029/2019JC015676, arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019JC015676. e2019JC015676 2019JC015676.

Gascard, J., Richez, C., 1985. Water masses and circulation in the western alboran sea and in the straits of gibraltar. Progress in Oceanography 15, 157–216. URL: https://www.sciencedirect.com/science/article/pii/007966118590031X, doi:https://doi.org/10.1016/0079-6611(85)90031-X.

Ghosh, A., 2006. On optimum choice of k in nearest neighbor classification. Computational Statistics Data Analysis 50, 3113–3123. doi:10.1016/j.csda.2005.06.007.

Gou, J., Ma, H., Ou, W., Zeng, S., Rao, Y., Yang, H., 2019. A generalized mean distance-based k-nearest neighbor classifier. Expert Systems with Applications 115, 356–372. URL: https://www.sciencedirect.com/science/article/pii/S0957417418305293, doi:https://doi.org/10.1016/j.eswa.2018.08.021.

Gou, J., Sun, L., Du, L., Ma, H., Xiong, T., Ou, W., Zhan, Y., 2022. A representation coefficient-based k-nearest centroid neighbor classifier. Expert Systems with Applications 194, 116529. URL: https:

//www.sciencedirect.com/science/article/pii/S0957417422000288, doi:https://doi.org/10.1016/j.eswa.2022.116529.

Hansen, B., 1916. Nogen hydrografiske metoder. Forh Skand Naturf Mote 16, 357–359.

Hansen, B.H., Nansen, F., . The norwegian sea : its physical oceanography based upon the norwegian researches 1900-1904.

Hou, W., Li, D., Xu, C., Zhang, H., Li, T., 2018. An advanced k nearest neighbor classification algorithm based on kd-tree, in: 2018 IEEE International Conference of Safety Produce Informatization (IICSPI), pp. 902–905. doi:10.1109/IICSPI.2018.8690508.

Huang, R.X., Yu, L.S., Zhou, S.Q., 2018. New definition of potential spicity by the least square method. Journal of Geophysical Research: Oceans 123, 7351–7365. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JC014306, doi:https://doi.org/10.1029/2018JC014306, arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018JC014306.

Hur, H.B., Jacobs, G.A., Teague, W.J., 1999. Monthly variations of water masses in the yellow and east china seas, november 6, 1998. Journal of Oceanography 55, 171–184.

Jackett, D.R., McDougall, T.J., 1985. An oceanographic variable for the characterization of intrusions and water masses. Deep Sea Research Part A. Oceanographic Research Papers 32, 1195–1207.

Jiang, L., Cai, Z., Wang, D., Jiang, S., 2007. Survey of improving k-nearest-neighbor for classification. Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007) 1, 679–683.

Kang, S., 2021. k-nearest neighbor learning with graph neural networks. Mathematics 9. URL: https://www.mdpi.com/2227-7390/9/8/830, doi:10.3390/math9080830.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley.

Kim, K., Kim, K.R., Rhee, T., Rho, H., Limeburner, R., Beardsley, R., 1991. Identification of water masses in the yellow sea and the east china sea by cluster analysis, in: Takano, K. (Ed.), Oceanography of Asian Marginal Seas. Elsevier. volume 54 of Elsevier Oceanography Series, pp. 253–267. URL: https://www.sciencedirect.com/science/article/pii/S0422989408701004, doi:https://doi.org/10.1016/S0422-9894(08)70100-4.

Markoulidakis, I., Rallis, I., Georgoulas, I., Kopsiaftis, G., Doulamis, A., Doulamis, N., 2021. Multiclass confusion matrix reduction method and its application on net promoter score classification problem. Technologies 9. URL: https://www.mdpi.com/2227-7080/9/4/81, doi:10.3390/technologies9040081.

McDougall, T.J., Krzysik, O.A., 2015. Spiciness. Journal of Marine Research 73, 141–152.

Miller, A., Stanley, R., 1961. Volumetric ts diagrams for the mediterranean sea. ATLANTIS 263, 6009–6146.

Millot, C., 2009. Another description of the mediterranean sea outflow. Progress in Oceanography - PROG OCEANOGR 82, 101–124. doi:10.1016/j.pocean.2009.04.016.

Millot, C., 2014. Heterogeneities of in- and out-flows in the mediterranean sea. Progress in Oceanography 120, 254–278. URL: https://www.sciencedirect.com/science/article/pii/S0079661113001857, doi:https://doi.org/10.1016/j.pocean.2013.09.007.

Millot, C., 2019. Comments on computations about the mediterranean outflow composition. Bollettino di Geofisica Teorica ed Applicata 60.

Millot, C., Candela, J., Fuda, J.L., Tber, Y., 2006. Large warming and salinification of the mediterranean outflow due to changes in its composition. Deep Sea Research Part I: Oceanographic Research Papers 53, 656–666. URL: https://www.sciencedirect.com/science/article/pii/S0967063706000112, doi:https://doi.org/10.1016/j.dsr.2005.12.017.

Molleri, G.S.F., Kampel, M., de Moraes Novo, E.M.L., 2010. Spectral classification of water masses under the influence of the amazon river plume. Acta Oceanologica Sinica , 1URL: http://www.aosocean.com/en/article/doi/10.1007/s13131-010-0031-1, doi:10.1007/s13131-010-0031-1.

Montgomery, R., 1958. Water characteristics of atlantic ocean and of world ocean. Deep Sea Research (1953) 5, 134–148. URL: https://www.sciencedirect.com/science/article/pii/0146631358900042, doi:https://doi.org/10.1016/0146-6313(58)90004-2.

Naranjo, C., Sammartino, S., García-Lafuente, J., Bellanco, M.J., Taupier-Letage, I., 2015. Mediterranean waters along and across the strait of gibraltar, characterization and zonal modification. Deep Sea Research Part I: Oceanographic Research Papers 105, 41–52. URL: https://www.sciencedirect.com/science/article/pii/S0967063715001399, doi:https://doi.org/10.1016/j.dsr.2015.08.003.

Nibani, H., Hilmi, K., Damghi, A., Beguery, L., Fommervault, O., Amhaoauch, Z., 2021. AL HOCEIMA LAUNCHES ITS FIRST FUNCTIONAL MARINE OBSERVATORY IN NORTH AFRICA, in: 9th EuroGOOS International conference, Shom and Ifremer and EuroGOOS AISBL, Brest, France. URL: https://hal.archives-ouvertes.fr/hal-03331067.

Ougiaroglou, S., Evangelidis, G., 2016. RHC: a non-parametric cluster-based data reduction for efficient k-nn classification. Pattern Anal. Appl. 19, 93–109. URL: http://dx.doi.org/10.1007/s10044-014-0393-7, doi:10.1007/s10044-014-0393-7.

Paik, M., Yang, Y., 2004. Combining nearest neighbor classifiers versus cross-validation selection. Statistical Applications in Genetics and Molecular Biology 3. URL: https://doi.org/10.2202/1544-6115.1054, doi:doi:10.2202/1544-6115.1054.

Pantiulin, A.N., 2002. Water masses: birth of the idea URL: https://ices-library.figshare.com/articles/report/Theme_Session_on_ACFM_and_Assessment_Working_Group_Reports/19258325, doi:10.17895/ices.pub.8512.

Parrilla, G., Kinder, T., 1987. The physical oceanography of the alboran sea 184, 31.

Parrilla, G., Kinder, T.H., Preller, R.H., 1986. Deep and intermediate mediterranean water in the western alboran sea. Deep Sea Research Part A. Oceanographic Research Papers 33, 55–88. URL: https://www.sciencedirect.com/science/article/pii/0198014986901081, doi:https://doi.org/10.1016/0198-0149(86)90108-1.

Pistek, P., De Strobel, F., Montanari, C., 1985. Deep-sea circulation in the alboran sea. Journal of Geophysical Research: Oceans 90, 4969–4976.

Pollak, M., 1958. Frequency distribution of potential temperatures and salinities in the indian ocean. Deep Sea Research (1953) 5, 128–133. URL: https://www.sciencedirect.com/science/article/pii/0146631358900030, doi:https://doi.org/10.1016/0146-6313(58)90003-0.

Provost, F., Kohavi, R., 1998. Guest editors' introduction: On applied research in machine learning. Machine learning 30, 127–132.

Renault, L., Oguz, T., Pascual, A., Vizoso, G., Tintore, J., 2012. Surface circulation in the alborán sea (western mediterranean) inferred from remotely sensed data. Journal of Geophysical Research: Oceans 117. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/

10.1029/2011JC007659, doi:https://doi.org/10.1029/2011JC007659, arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2011JC007659.

Roseli, N.H., Akhir, M.F., Husain, M., Tangang, F., Ali, A., 2015. Water mass characteristics and stratification at the shallow sunda shelf of southern south china sea. Open Journal of Marine Science 05, 455–467. doi:10.4236/ojms.2015.54036.

Rousseeuw, P., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. URL: http://portal.acm.org/citation.cfm?id=38772, doi:http://dx.doi.org/10.1016/0377-0427(87)90125-7.

Shahriyari, L., 2017. Effect of normalization methods on the performance of supervised learning algorithms applied to htseq-fpkm-uq data sets: 7sk rna expression as a predictor of survival in patients with colon adenocarcinoma. Briefings in bioinformatics 20. doi:10.1093/bib/bbx153.

Veronis, G., 1972. On properties of seawater defined by temperature, salinity, and pressure. Journal of Marine Research 30, 227–255.

Álvaro Viúdez, Tintoré, J., Haney, R.L., 1996. Circulation in the alboran sea as determined by quasi-synoptic hydrographic observations. part i: Three-dimensional structure of the two anticyclonic gyres. Journal of Physical Oceanography 26, 684 – 705. URL: https://journals.ametsoc.org/view/journals/phoc/26/5/1520-0485_1996_026_0684_citasa_2_0_co_2.xml, doi:10.1175/1520-0485(1996)026<0684:CITASA>2.0.CO;2.

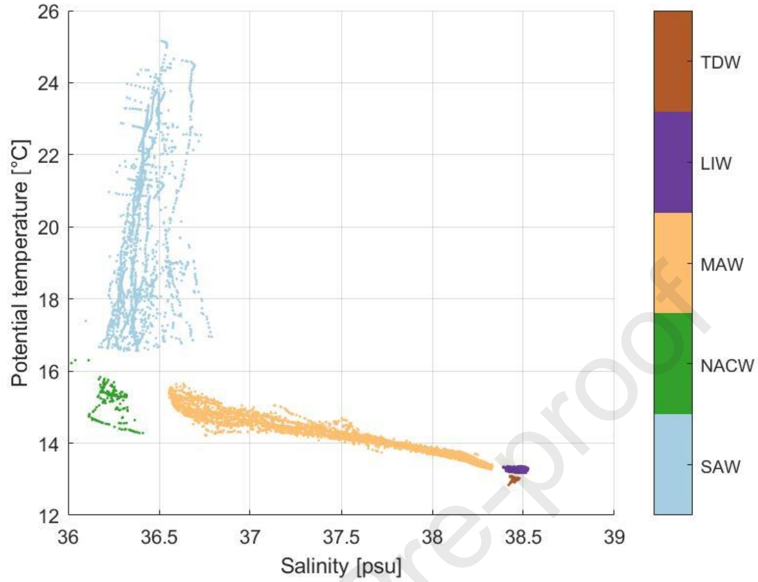
Vélez-Belchı, P., Vargas-Yáñez, M., Tintoré, J., 2005. Observation of a western alboran gyre migration event. Progress in Oceanography 66, 190–210. doi:10.1016/j.pocean.2004.09.006.

Worthington, L.V., 1981. The water masses of the world ocean : Some results of a fine-scale census.

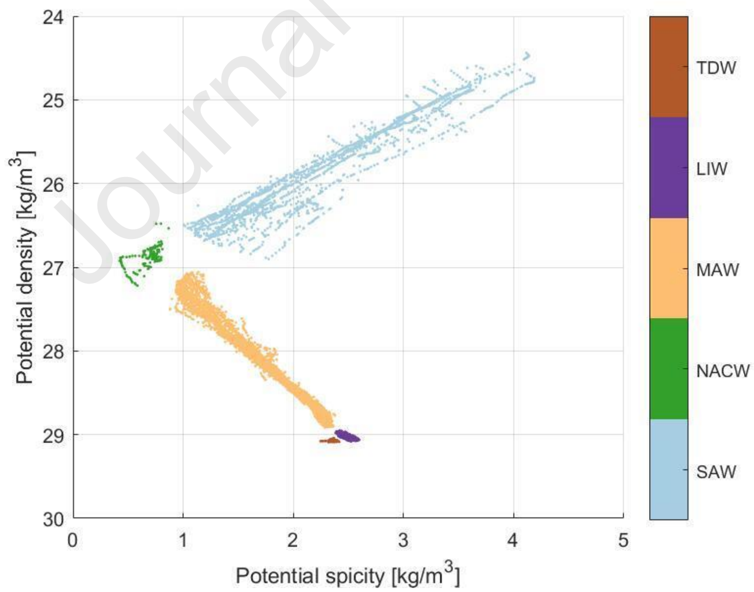
Wright, W., Worthington, L., 1970. Serial atlas of the marine environment. folio 19. the water masses of the north atlantic ocean: a volumetric census of temperature and salinity. Technical Report. Woods hole oceanographic institution ma.

Zhang, S., Li, X., Zong, M., Zhu, X., Wang, R., 2018. Efficient knn classification with different numbers of nearest neighbors. IEEE Transactions on Neural Networks and Learning Systems 29, 1774–1785. doi:10.1109/TNNLS.2017.2673241.

Zhongguo, Y., Hongqi, L., Liping, Z., Qiang, L., Ali, S., 2017. A case based method to predict optimal k value for k-nn algorithm. J. Intell. Fuzzy Syst. 33, 55–65. URL: https://doi.org/10.3233/JIFS-161062, doi:10.3233/JIFS-161062.

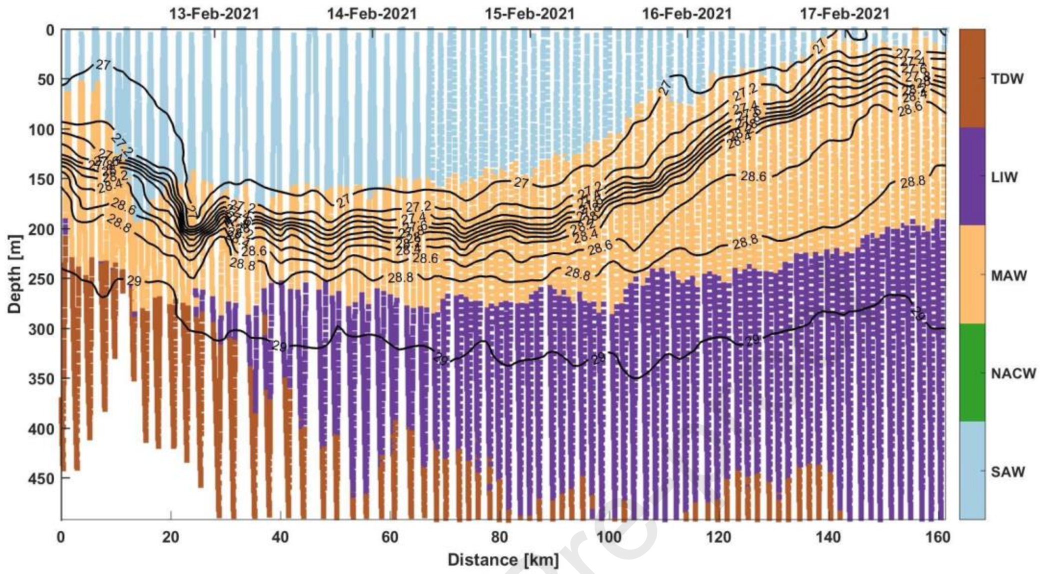


(a)

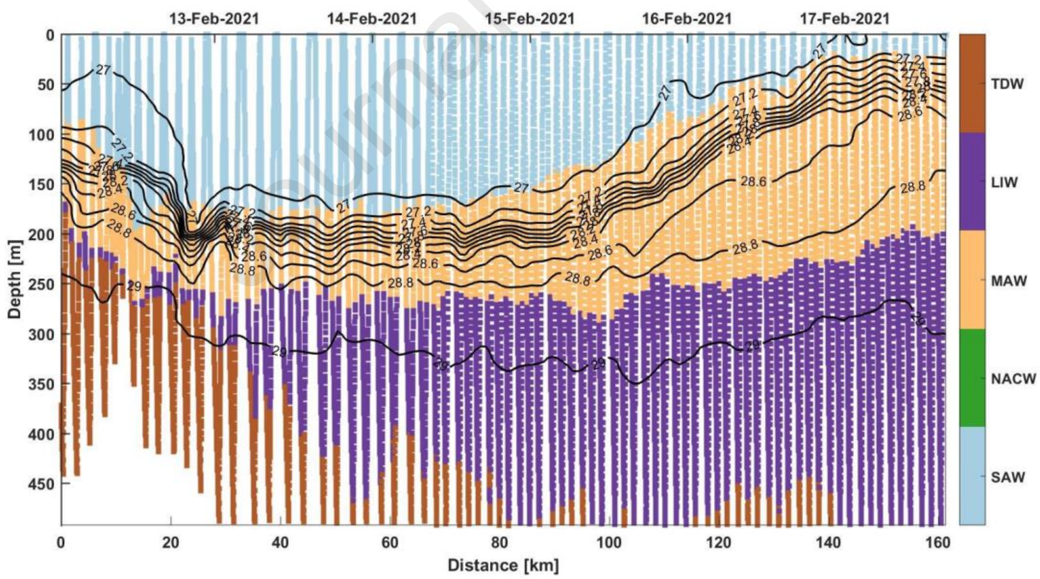


(b)

Figure 4: Example of a $\theta - S$ diagram (a) labelled by the Atlantic (SAW, NACW, MAW) and Mediterranean (LIW, TDW) water masses, with frontiers removed, and its equivalent $\sigma - \pi$ diagram (b).

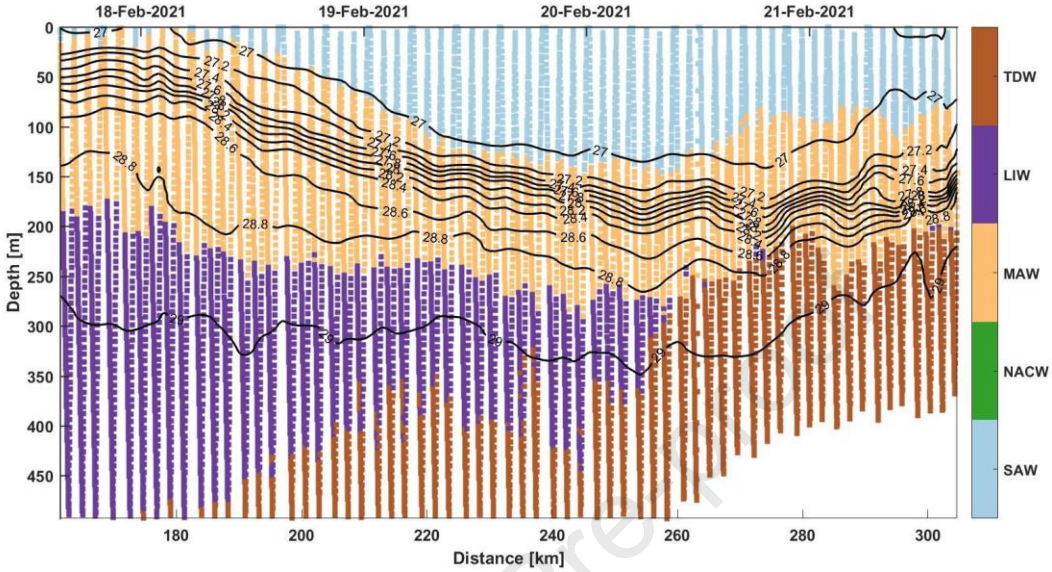


(a)

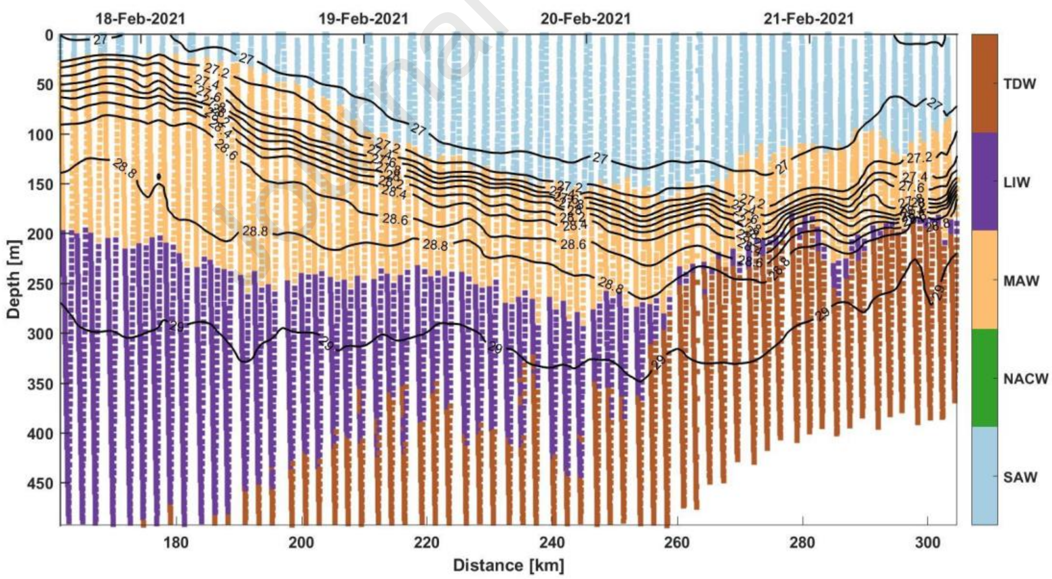


(b)

Figure 5: Classification of the water masses in the first transect using training dataset with frontiers removed for $\sigma - \pi$ (a) and $\theta - S$ (b) diagrams.

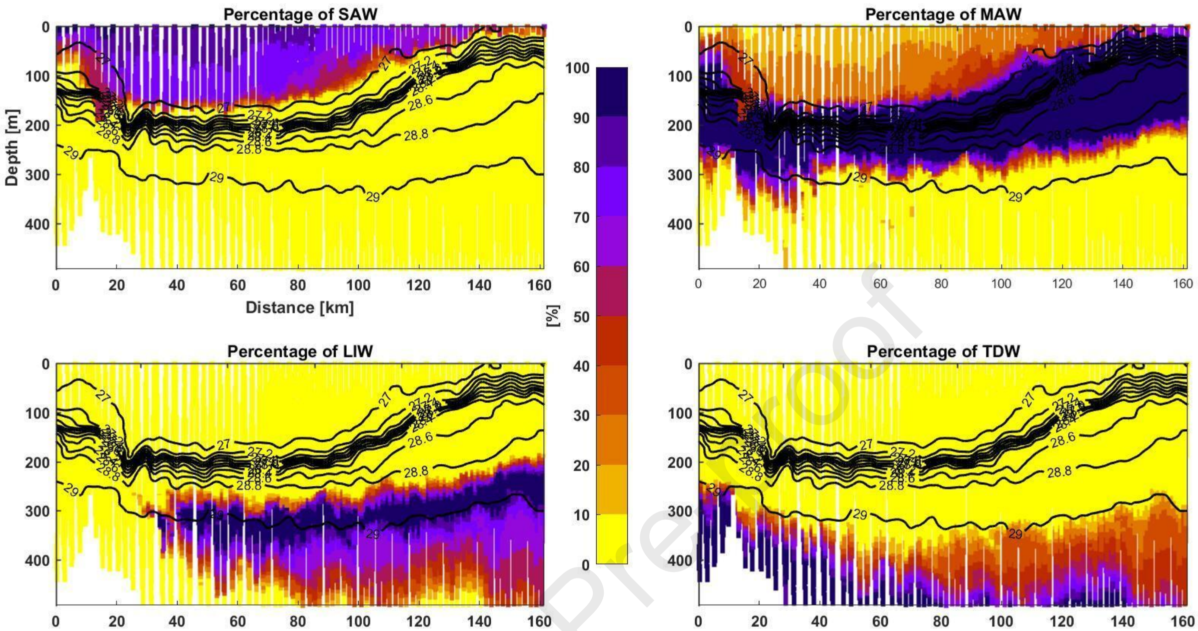


(a)

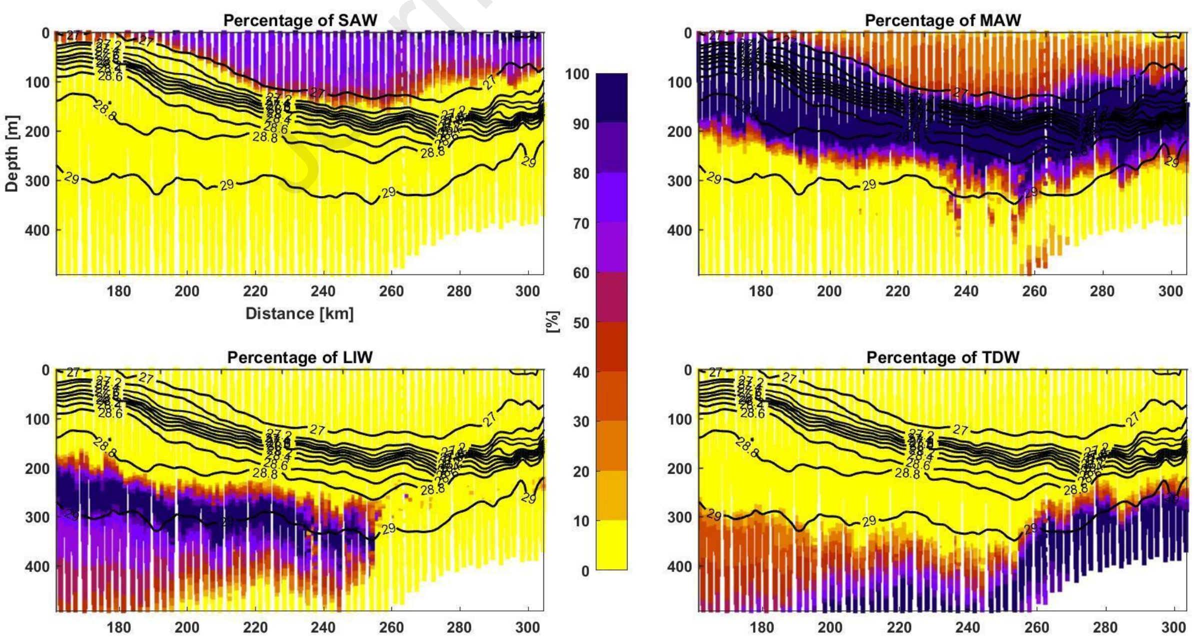


(b)

Figure 6: Classification of the water masses in the second transect using training dataset with frontiers removed for $\sigma - \pi$ (a) and $\theta - S$ (b) diagrams.

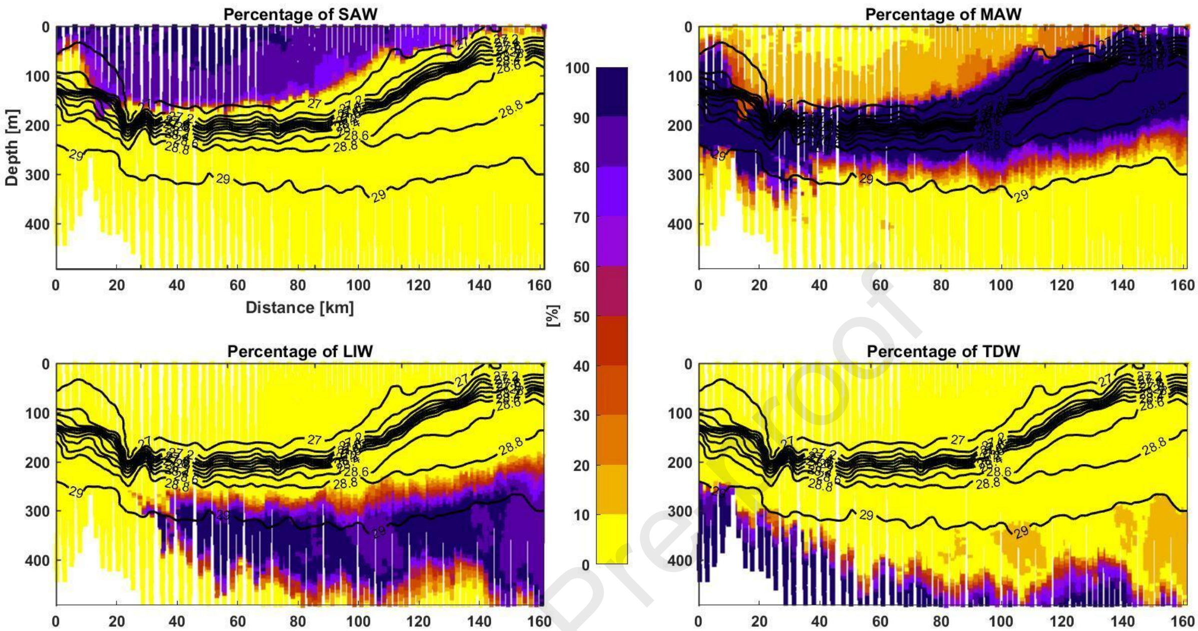


(a)

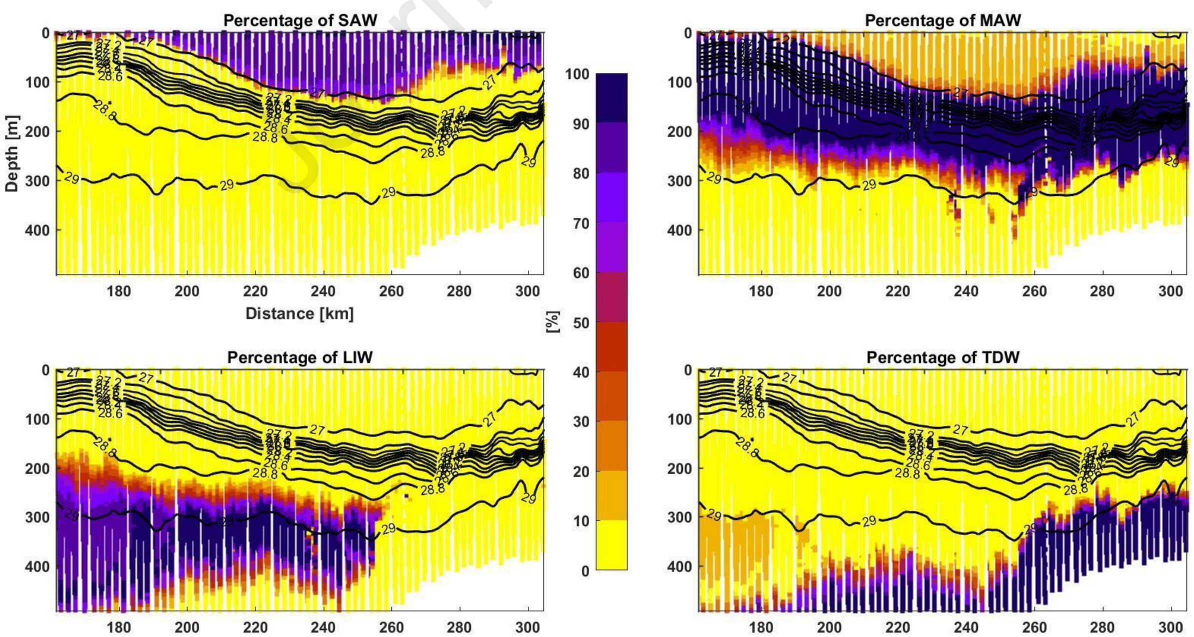


(b)

Figure 7: Percentage of the AWs and MWs along the first (a) and second (b) transects using (σ-π) diagram. The sum of the four contributions leads to 100% in the Atlantic and Mediterranean layers.



(a)



(b)

Figure 8: Percentage of the AWs and MWs along the first (a) and second (b) transects using ($\theta - S$) diagram. The sum of the four contributions leads to 100% in the Atlantic and Mediterranean layers.

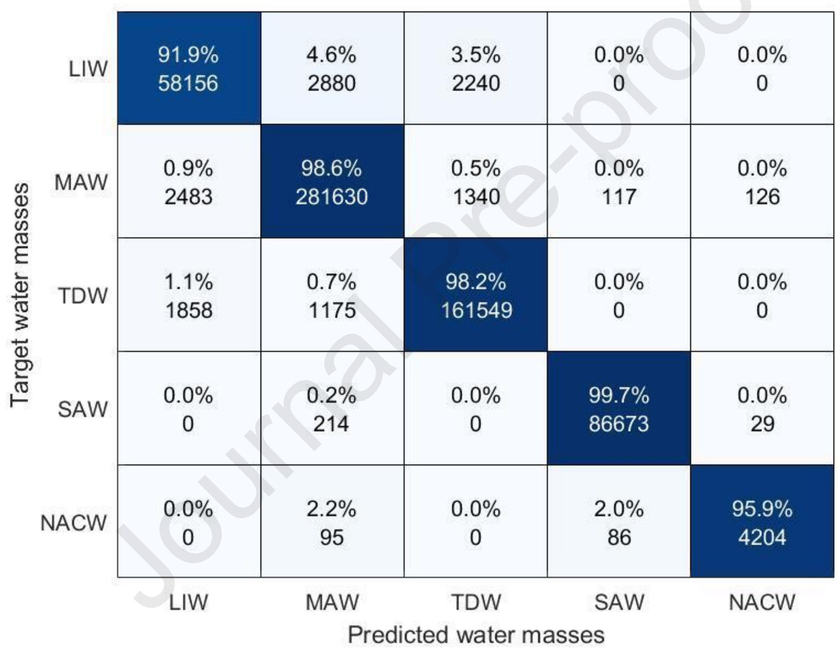


Figure 9: Confusion matrix for the 5 water masses deduced from the classifcation during the training stage.

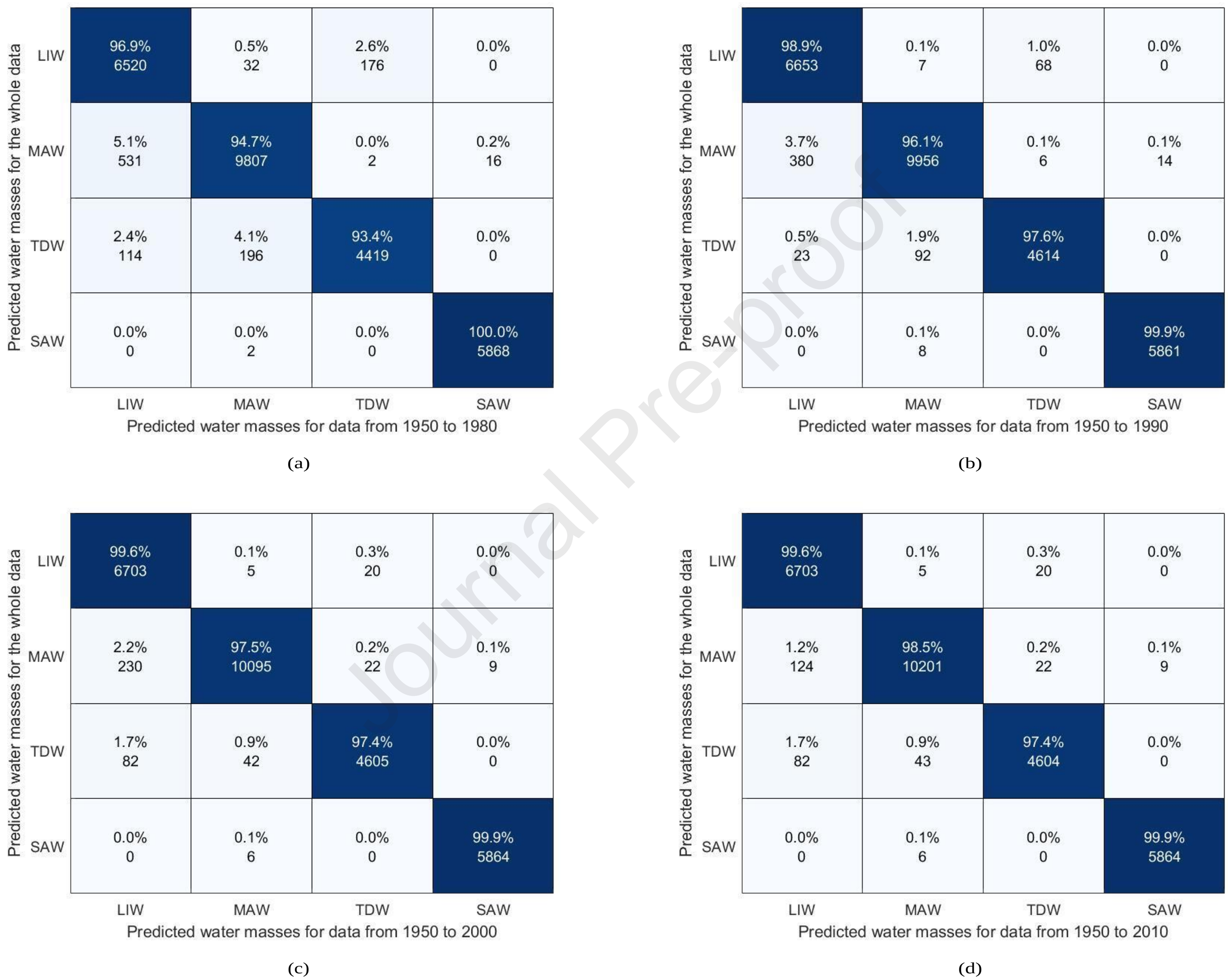
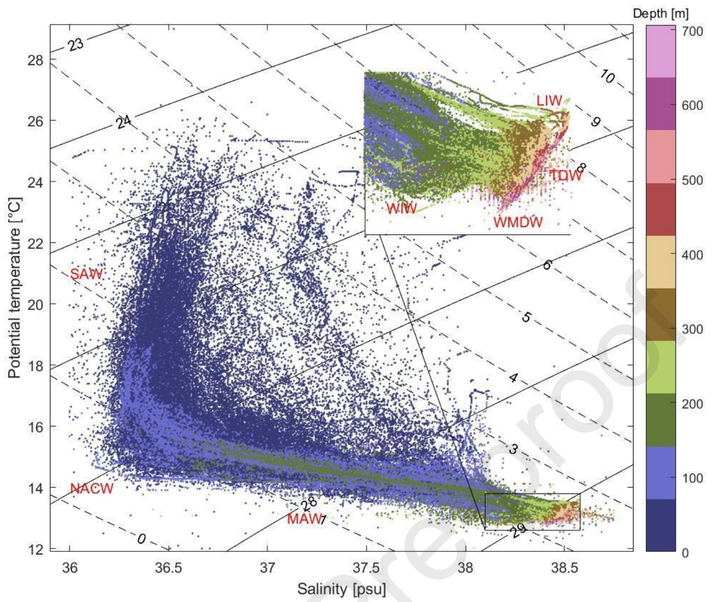
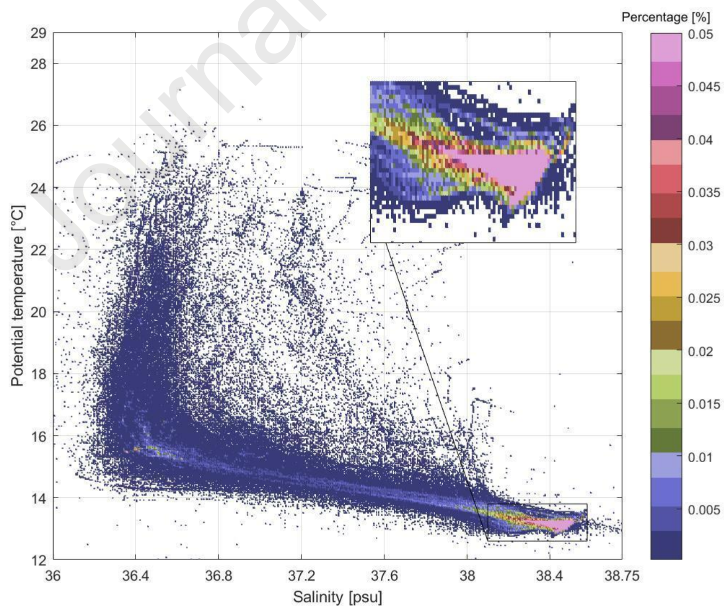


Figure 10: Confusion matrices of training dataset profiles gathered from 1950 to 1980 (a), from 1950 to 1990 (b), from 1950 to 2000 (c) and from 1950 to 2010 (d).

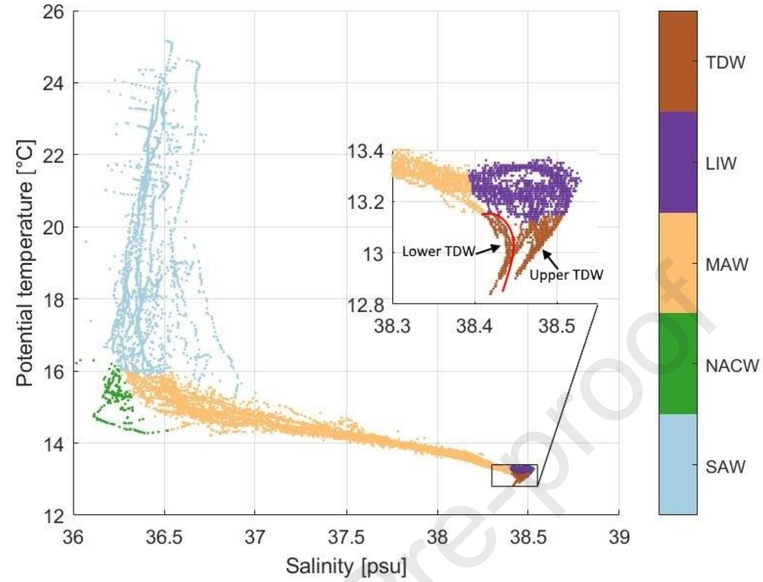


(a)

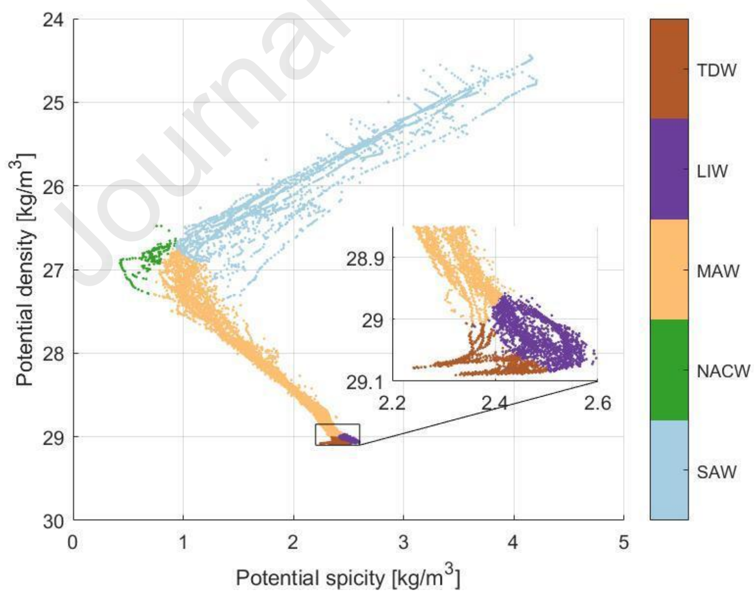


(b)

Figure 11: (a) $\theta - S$ diagram for all the database. Depths between the surface and 700 m are illustrated in different colors. isopycnals (solid lines) and spicity isopleths (dotted lines) are plotted 1 kg.m^{-3} apart. (b) Occurrence of water types as a function of temperature and salinity over temporal range of the WOD18 (1951-2020). Bin is scaled to represent percentage of total points. The color scales go from 0 to 0.05 for the sake of clarity.

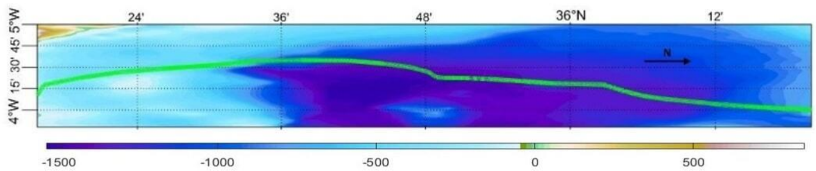


(a)

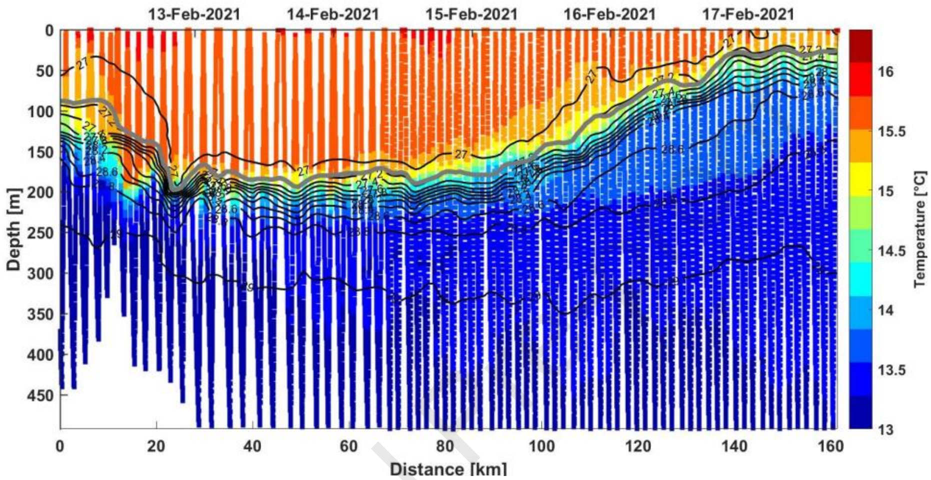


(b)

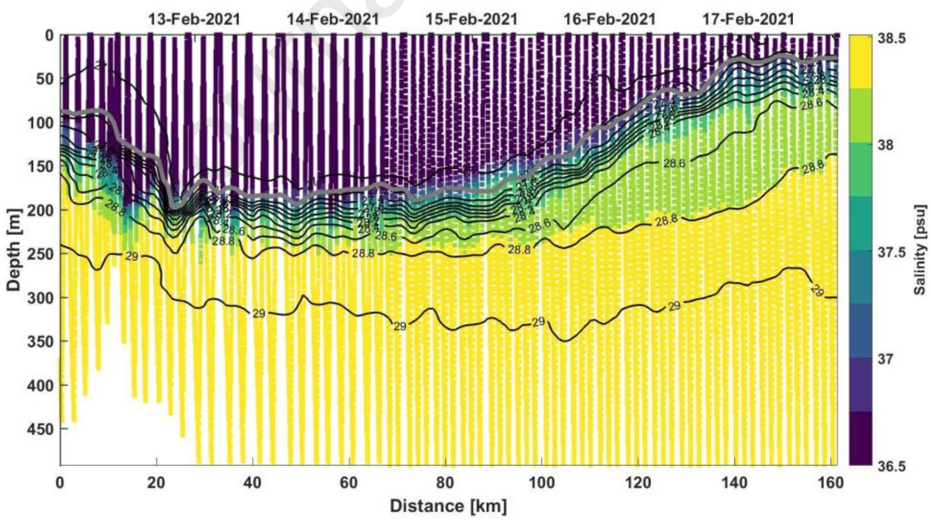
Figure 12: Example of a θ−S diagram (a) labelled by the Atlantic (SAW, NACW, MAW) and Mediterranean (LIW, TDW) water masses and its equivalent σ−π diagram (b). Upper-TDW and lower-TDW are separated by the red curve plotted in the inset (a).



(a)

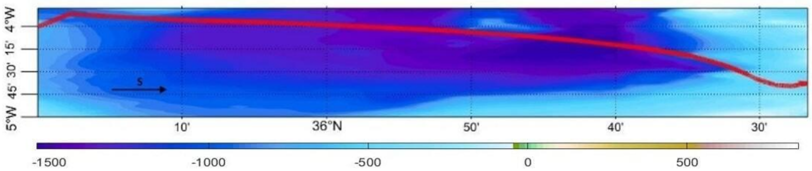


(b)

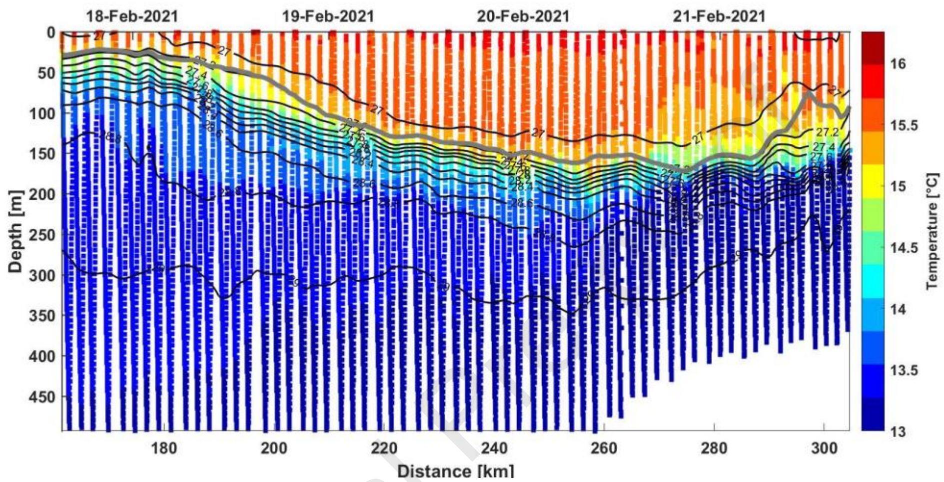


(c)

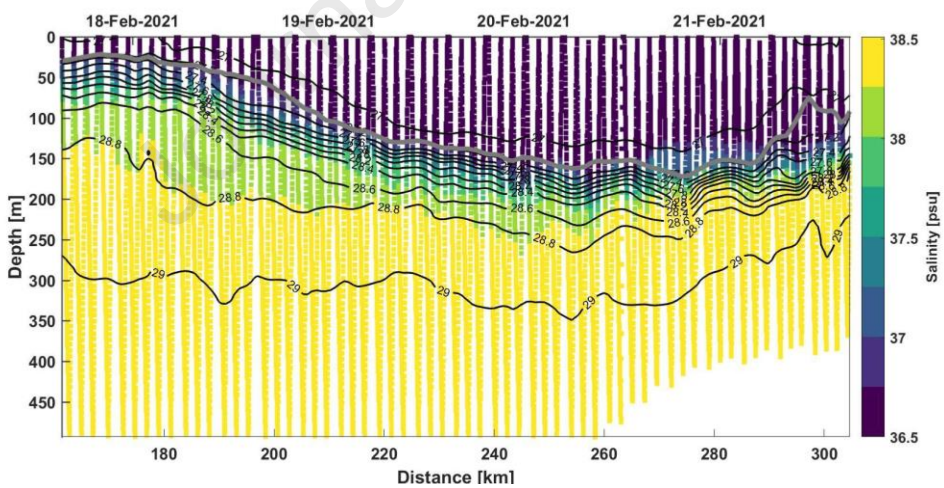
Figure 13: Temperature (b) and salinity (c) along the first glider transect (a). The black lines are the isopycnal levels and the gray line is the Mixed Layer Depth, defined using the threshold method with a finite difference criterion (density criterion of 0.03 kg.m^{-3}). The black arrow in (a) points in the North direction.



(a)

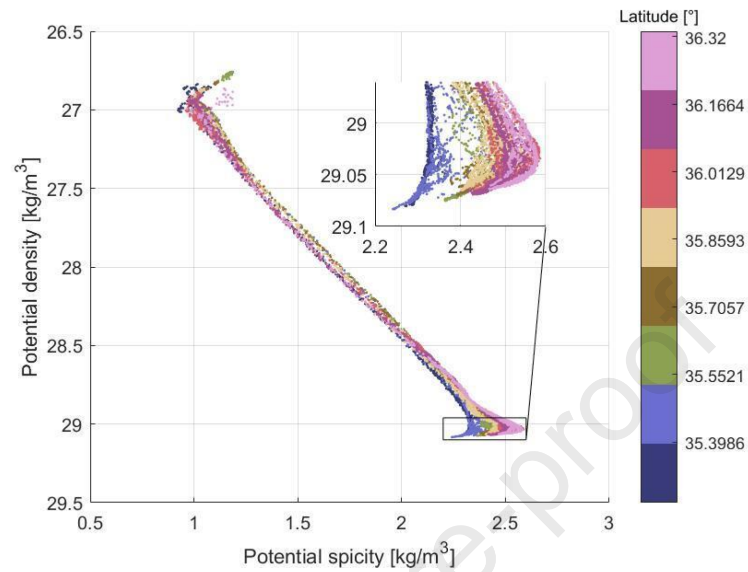


(b)

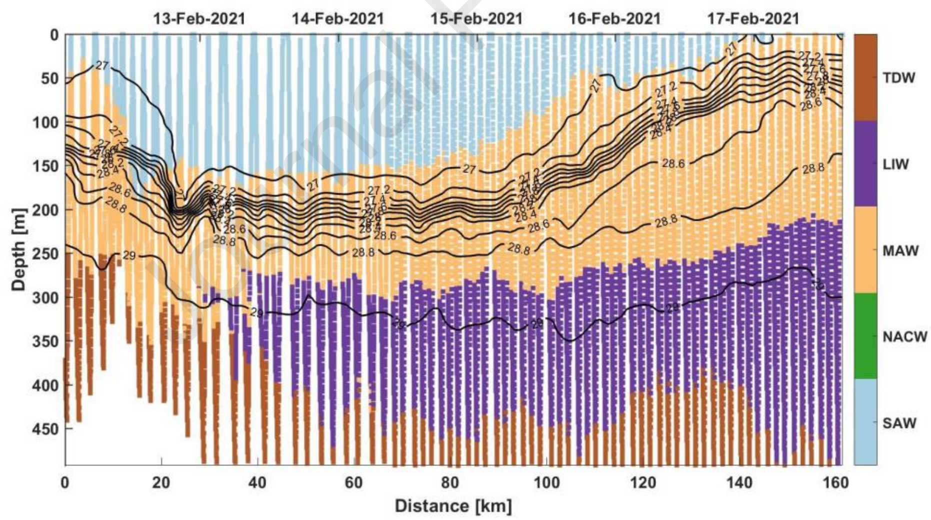


(c)

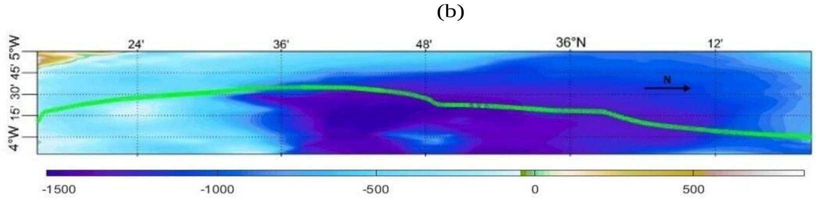
Figure 14: Temperature (b) and salinity (c) along the second glider transect (a). The black lines are the isopycnal levels and the gray line is the Mixed Layer Depth, defined using the threshold method with a finite difference criterion (density criterion of 0.03 kg.m^{-3}). The black arrow in (a) points in the South direction.



(a)

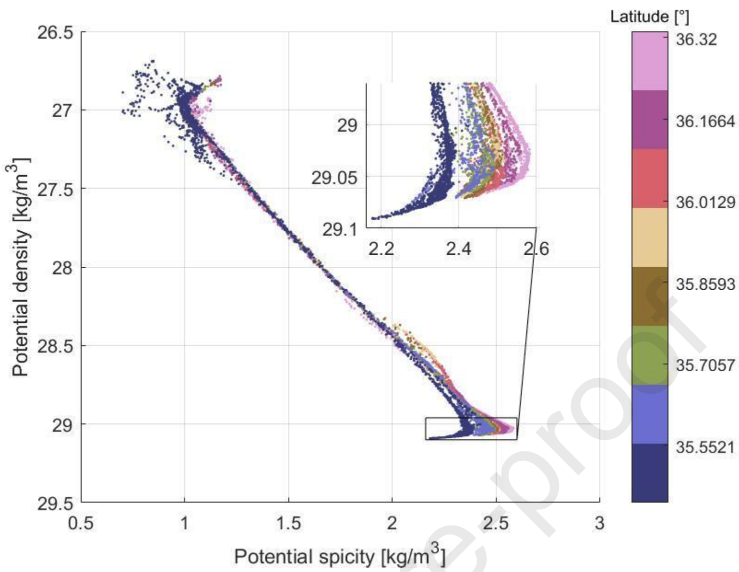


(b)

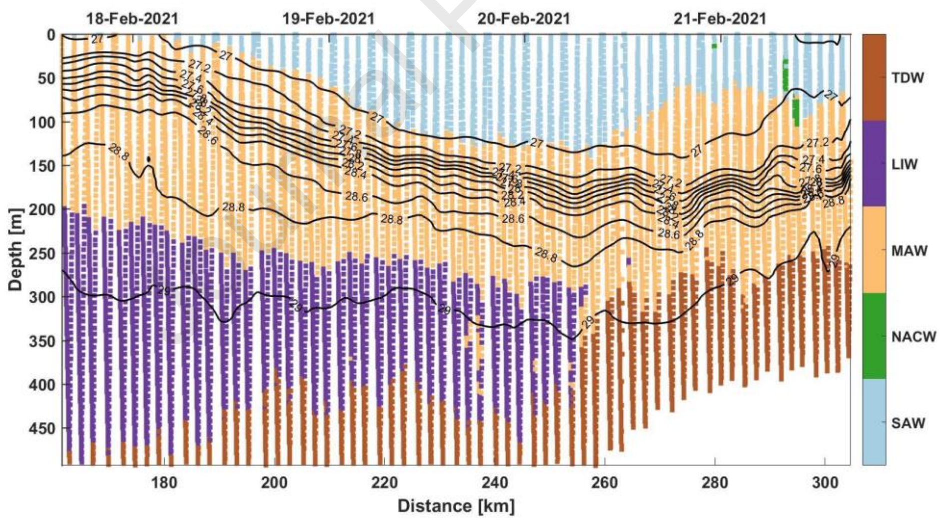


(c)

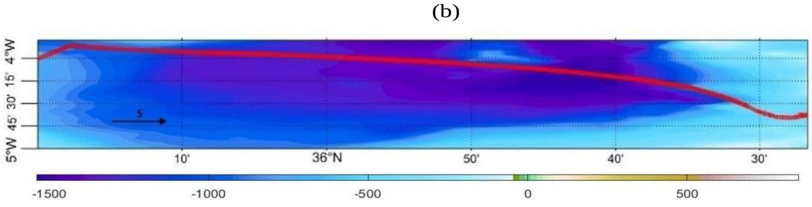
Figure 15: The $\sigma - \pi$ diagram (a) and the classification (b) of the water masses in the first glider transect (c). The black arrow in (c) points in the North direction.



(a)

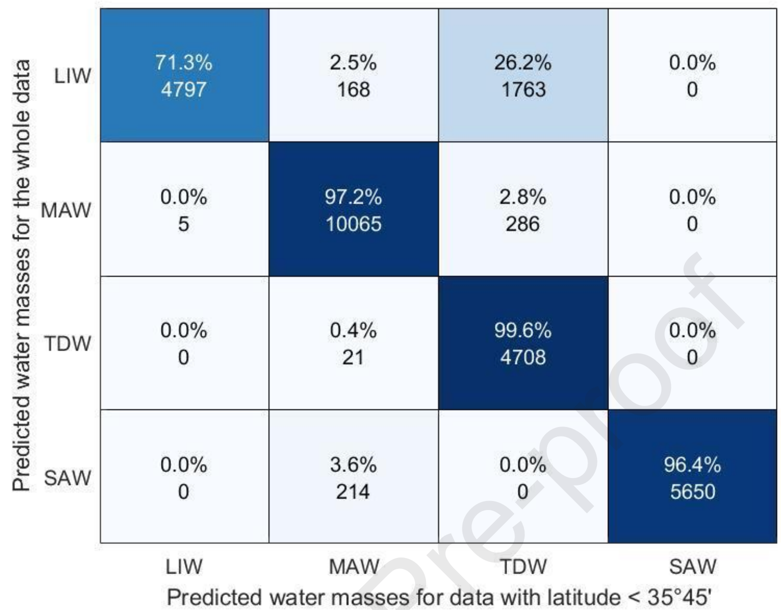


(b)

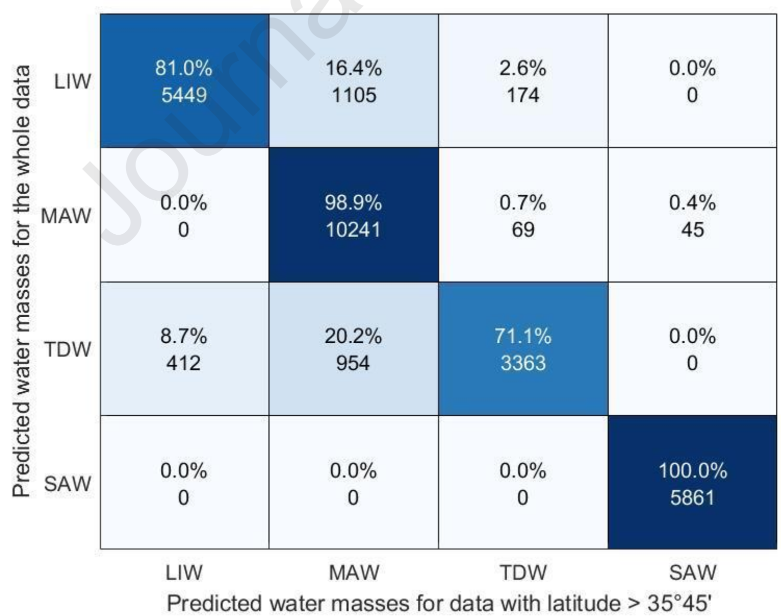


(c)

Figure 16: The σ − π diagram (a) and the classification (b) of the water masses in the second glider transect (c). The black arrow in (c) points in the South direction.

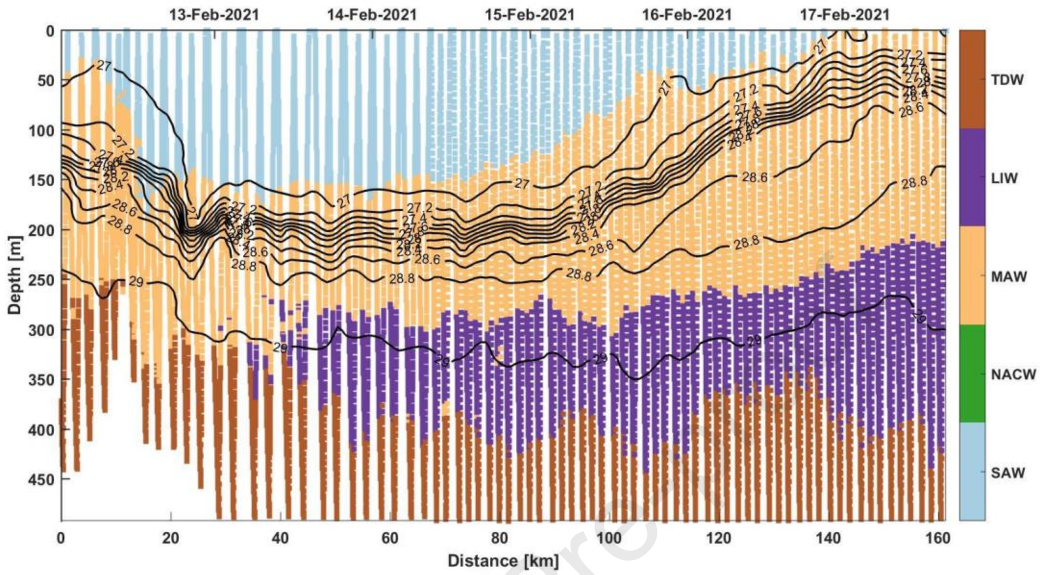


(a)

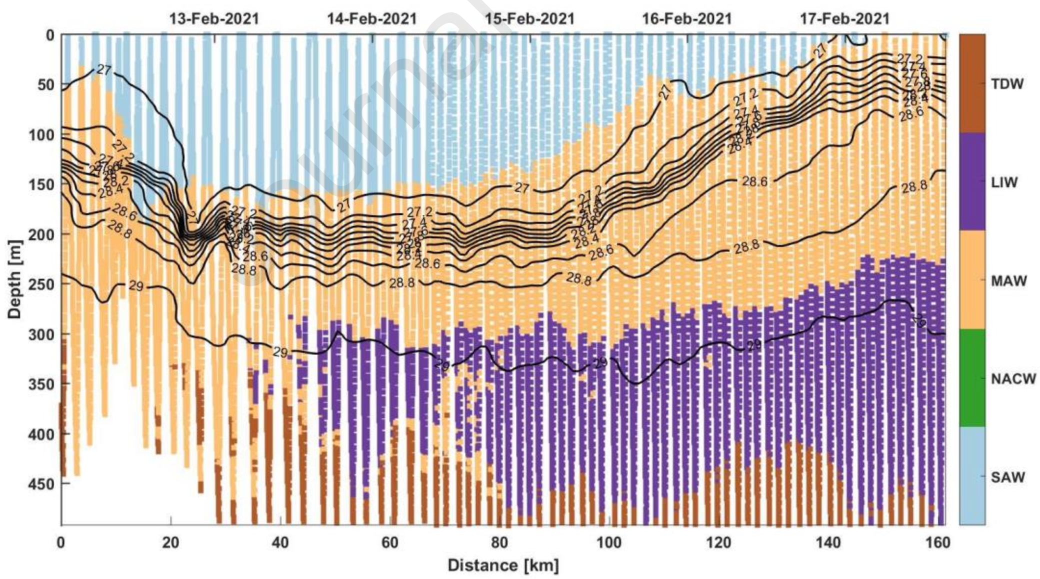


(b)

Figure 17: Confusion matrices of training dataset profiles gathered below (a) and beyond (b) 35°45’N.

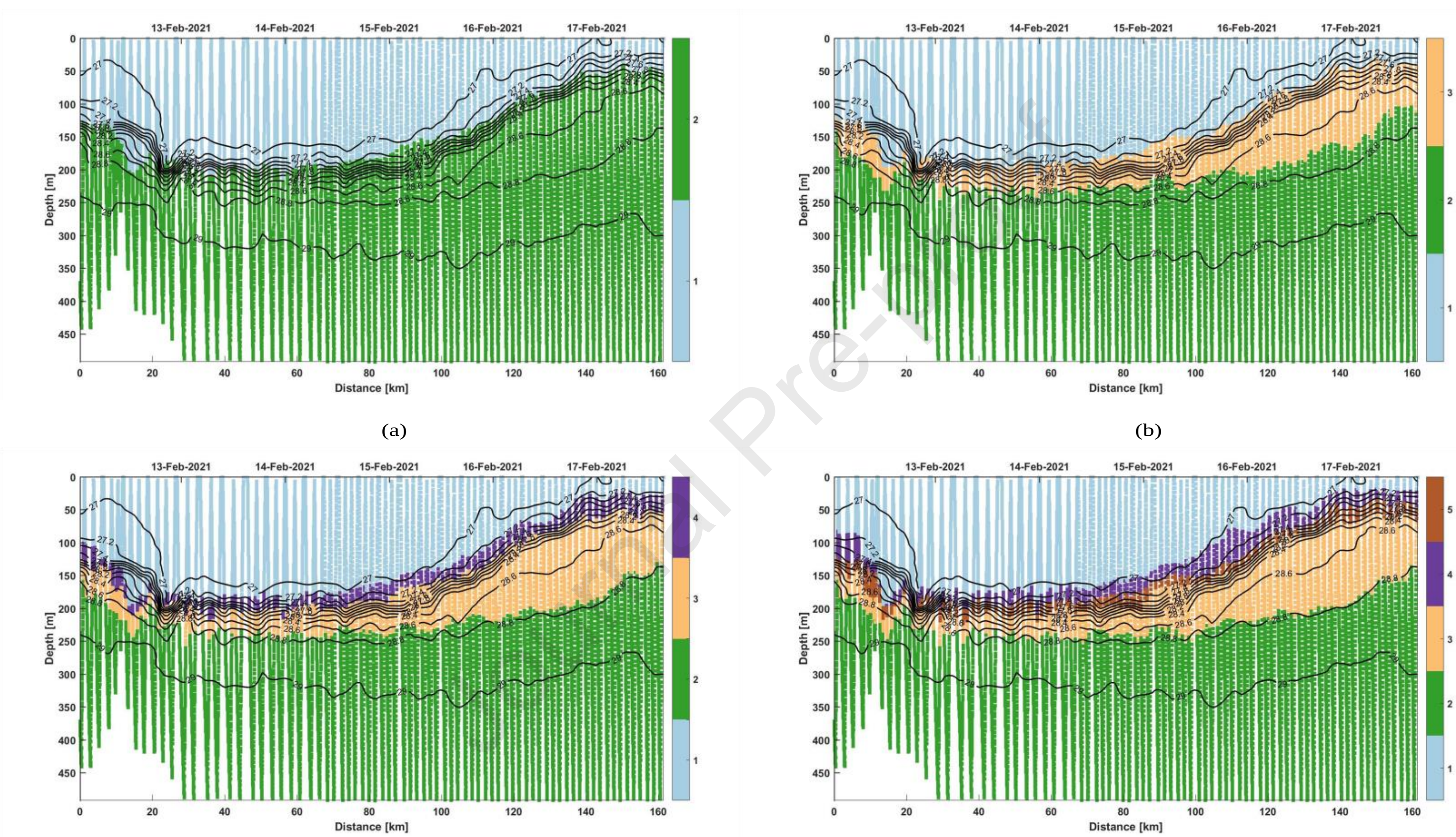


(a)



(b)

Figure 18: Classification of the water masses in the first transect using training dataset profiles gathered below (a) and beyond (b) 35°45'N.



(a) (b)

(c) (d)

Figure 19: Classification of water masses in the first transect (Figure 15c) provided by the k-means clustering. The number of clusters used in: (a) k = 2, (b) k = 3, (c) k = 4 and (d) k = 5.

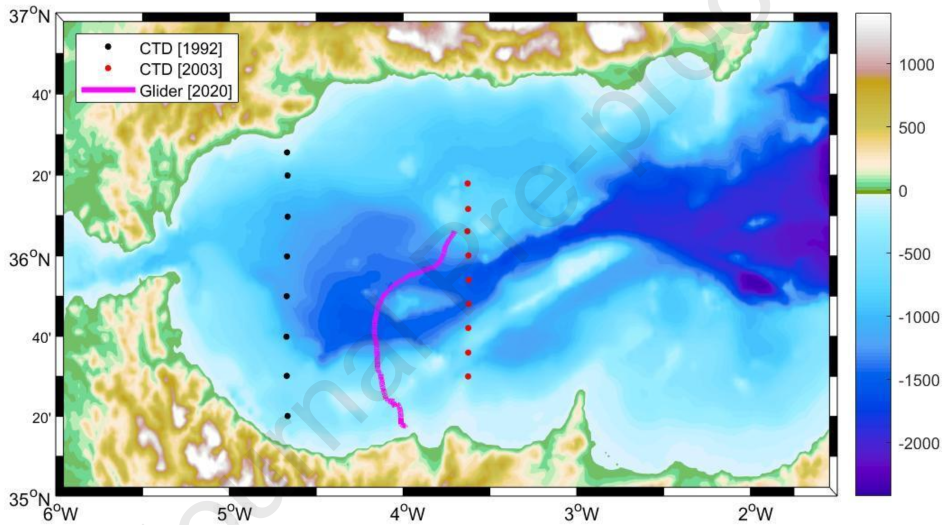
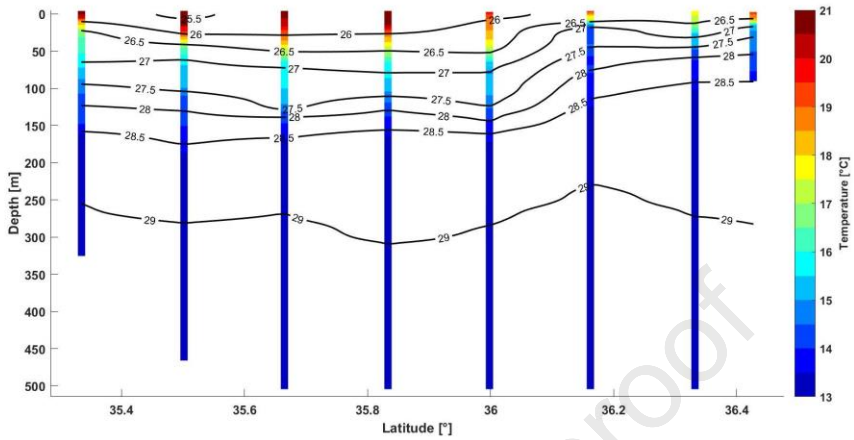
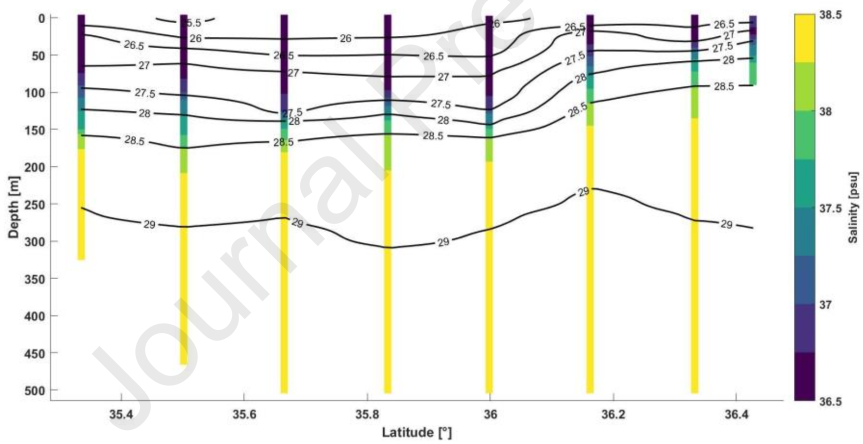


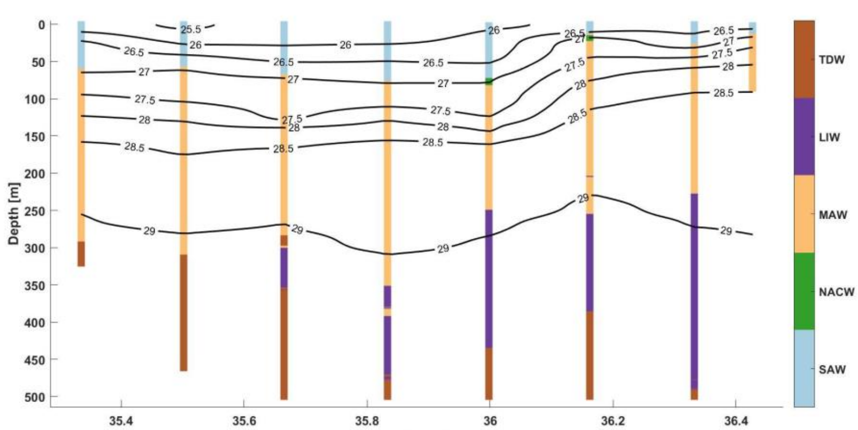
Figure A.20: Map of the Western Alboran Sea sketching the bathymetric depths and topographic elevations in meters (m) relative to the mean sea level. Black dots indicate the position of the CTD data gathered in 1992. Red dots represent the localization of the CTD data collected in 2003. The glider transect is sketched in magenta.



(a)

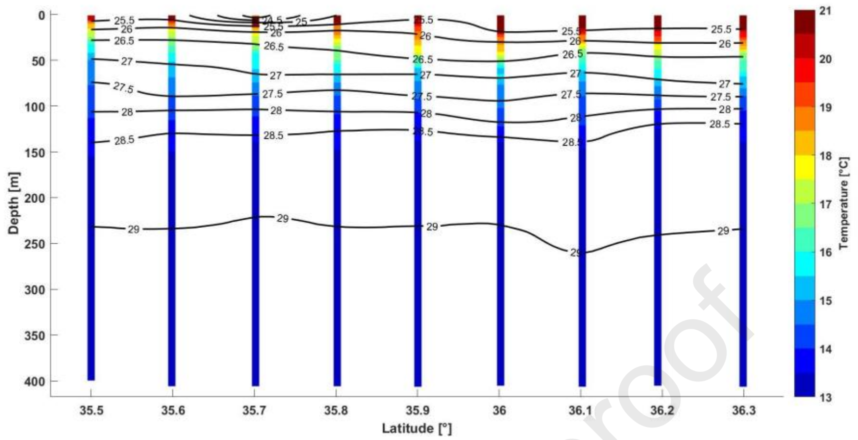


(b)

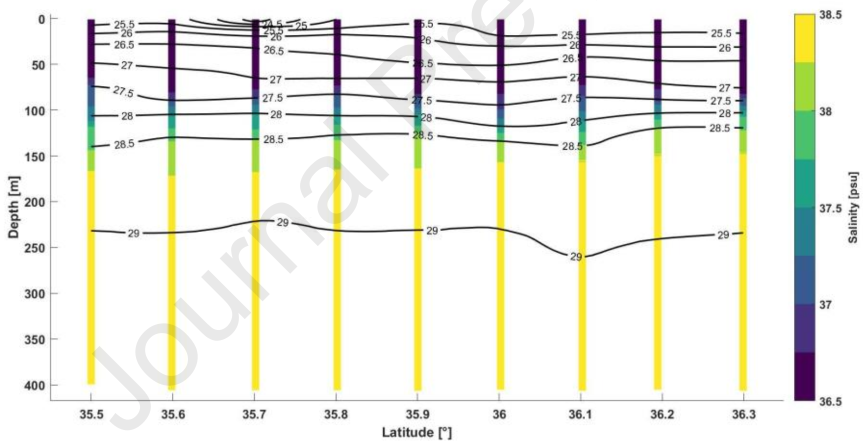


(c)

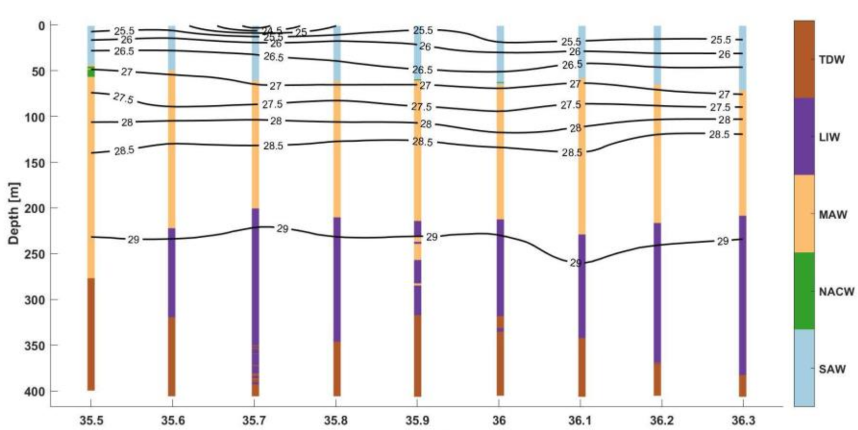
Figure A.21: Temperature (a) and salinity (b) along the westernmost transect (figure A.20). (c) represent the classification results. The black lines are the isopycnal levels



(a)

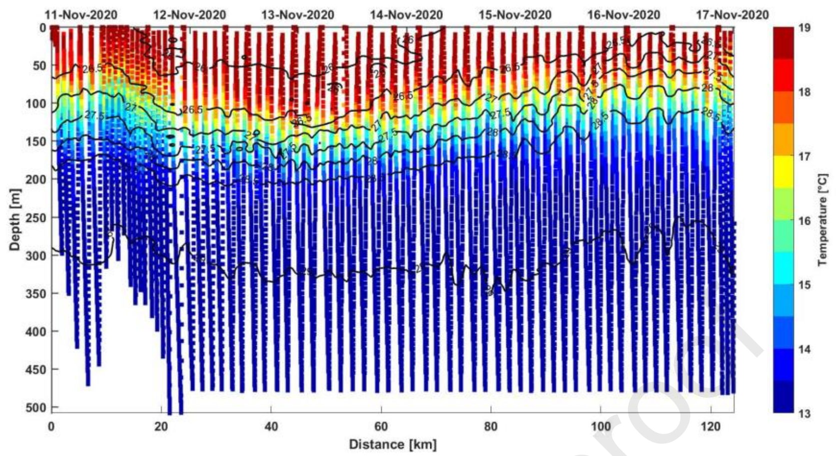


(b)

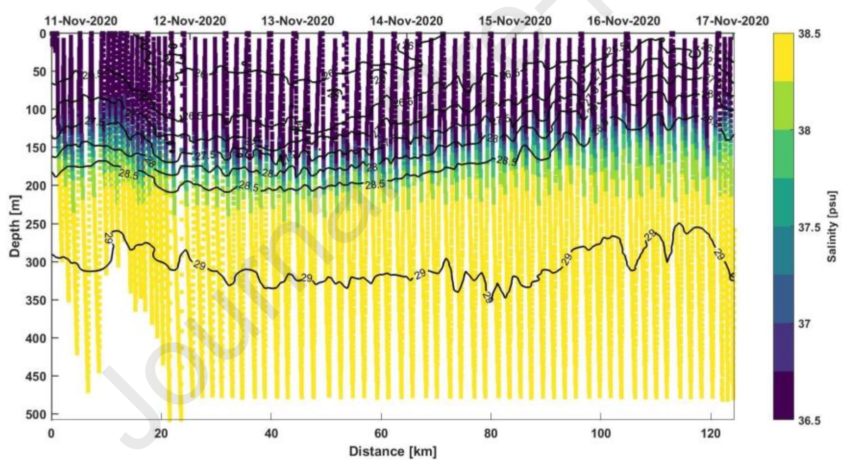


(c)

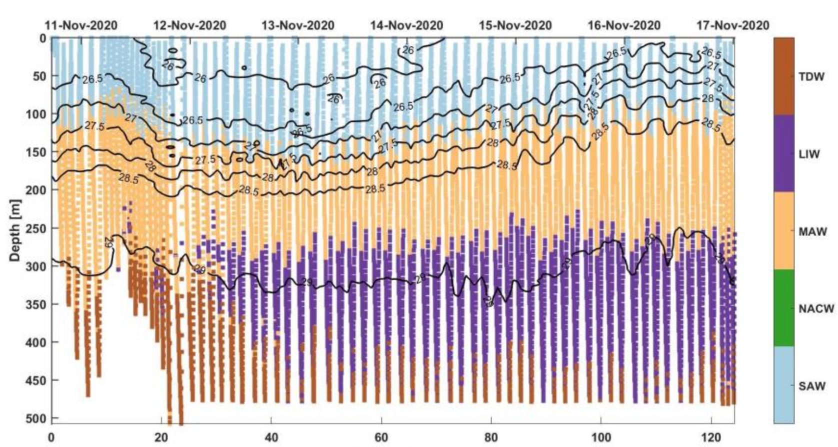
Figure A.22: Temperature (a) and salinity (b) along the easternmost transect (figure A.20). (c) represent the classification results. The black lines are the isopycnal levels.



(a)



(b)



(c)

Figure A.23: Temperature (a) and salinity (b) along the glider transect (figure A.20). (c) represent the classification results. The black lines are the isopycnal levels.

- High spatial resolution glider profiles of θ-S in the western Alboran sea ;

- Water masses derived on a (σ-π) diagram using Knn algorithm ;

- Classification results confirm earlier derived circulation schemes ;

- The proposed method outperforms classical clustering analysis in delineating water mass boundaries.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: