Clustering analysis of the *Sargassum* transport process: application to ~~stranding~~beaching prediction in the Lesser Antilles

a mis en forme : Police :Italique

Didier Bernard[1], Emmanuel Biabiany[2], Raphaël Cécé[1], Romual Chery[1], and Naoufal Sekkat[1]

[1]LARGE, University of the French West Indies, 97157 Pointe-à-Pitre, Guadeloupe, France.

5 [2]LAMIA, University of the French West Indies, 97157 Pointe-à-Pitre, Guadeloupe, France.

Correspondence to: Didier Bernard (didier.bernard@univ-antilles.fr)

Abstract. The massive *Sargassum* algae ~~strandings~~beachings observed over the past decade are the new natural hazard that currently impacts the island states of the Caribbean region (human health, environmental damages, and economic losses). This study aims to improve the prediction of the surface current dynamic leading to beachings in the Lesser Antilles, using clustering analysis methods.

10 The input surface currents ~~including windage effect~~ were derived from the Mercator model and the Hybrid Coordinate Ocean Model (HYCOM~~).~~) outputs in which we integrated the windage effect. Past daily observations of *Sargassum* ~~stranding~~beaching on Guadeloupe coasts and satellite-based *Sargassum* offshore abundance were also integrated. Four representative current regimes were identified for both Mercator and HYCOM data. The analysis of the ~~backward~~ current sequences leading to ~~strandings~~beachings showed that the recurrence of two current regimes is related to the beaching peaks observed respectively in March and in August.

15 ~~A decision tree classifier was built and its accuracy reaches 73.3% with 0.04° scale~~ The performance score of the predictive model showed that the HYCOM data ~~and 50.8% with 0.08° scale Mercator data. This significant accuracy difference highlights the need of very small-scale current data (i.e., lower than 5 km scale)~~seem more suitable to assess coastal *Sargassum* hazard in the Lesser Antilles. For one year of tests (i.e. 2021), the decision tree accuracy reached respectively 70.1% and 58.2% for HYCOM and Mercator with a temporal uncertainty range +/-3 days around the forecast date. The present clustering analysis predictive system

20 requiring lower computational resources compared to conventional forecast models would help improve this risk management in the islands of this region.

1 Introduction

During the periods 2011-2012, then 2014-2019, massive *Sargassum* ~~strandings~~beachings impacted most coasts of the Lesser Antilles (LA), mainly those facing east and southeast (Franks et al., 2012, Gower et al., 2013, Johnson. et al., 2014, Hu et al., 2016, Wang and Hu, 2016~~)~~. The LA received large amounts of algae on the windward Atlantic coastline, while leeward Caribbean coastal areas remained slightly affected (~~Franks et al., 2012, Gower et al., 2013, Johnson. et al., 2014, Hu et al., 2016, Wang and Hu, 2016~~, Maréchal et al., 2017). These beachings in terms of frequency and intensity can now be considered as a new natural hazard for the Caribbean islands and American coasts. ~~Strandings were also be observed in Africa (Széchy et al., 2012).~~

Indeed, while it has been demonstrated that *Sargassum* algae provide ecosystem services, habitat and shelter for various organisms

30 in a structurally sterile ocean ecosystem (Witherington et al., 2012; Bertola et al., 2020), the ~~strandings~~beachings over the past decade have induced health risks for the population and have had considerable socio-economic impacts (Franks et al., 2012). For

1

example, when looking at the French ~~west indies,~~West Indies, the Guadeloupe archipelago and Martinique, the findings are as follows:

(1) Apart from 2013, the recent inflow of *Sargassum* rafts on the coasts of Guadeloupe and Martinique, although irregular, has not ceased since 2011, reaching a paroxysm in 2015 (Florenne et al., 2016; Berline et al., 2020). State services estimated that the volumes stranded on the shores were in the order of 1.5 million m³, from October 2014 to October 2015 in Guadeloupe (Florenne et al., 2016). Only a third of these could be collected by the authorities and priority ~~being~~was given to areas at stake, ~~such as,~~ inhabited areas, shores with economic or tourist activities and ecosystems or other environmental niches. The particularity and the difficulty lay in the fact that 60% of this coastline and/or of the volume stranded remained inaccessible to the techniques currently proven and at costs currently bearable.

(2) There was an impact on human health and ecosystems because in shallow and small bays, the accumulated algae degrade by fermentation and emit chemical compounds such as hydrogen sulfide (H2S) and ammonia (NH3) (Anses, 2017, Van Tussenbroek et al., 2017; Resiere et al., 2018).

(3) The survey conducted by the organizations responsible for socio-economic development estimated that the decline in tourism resulted in an economic loss of $5.5 million for the first half of 2015 (https://eos.org/features/s*Sargassum*-watch-warns-of-incoming-seaweed).

The volumes ~~needed~~ to be collected ~~being~~were considerable compared to the size of these islands (< 1200 km² each) and the vulnerability of these territories. This new phenomenon has raised several scientific questions relating to *Sargassum* rafts, such as their transports, origins, the sources of nutrients promoting their growth but especially the physical factors that led to their occurrence and their development ~~of Sargassum rafts~~ in the tropical and equatorial Atlantic.

Using large-scale observations with ocean color satellite remote sensing, historical hydrographic observations, time series of *Sargassum* volume collected on ships, multi-year reanalysis of wind and current, and numerical models ~~estimated,~~ both the roles of both subsurface nutrient supply and surface current transport were estimated. Several authors have contributed to the understanding of the mechanisms and physicochemical processes governing the phenomenon (Gower et al., 2006; Gower and King, 2011; Gower et al., 2013; Maréchal et al., 2017; Johns et al., 2020). Operational systems have been developed such as the satellite based *Sargassum* Watch System SaWS (Hu, 2009; Hu et al., 2015) and the *Sargassum* Early Advisory System (SEAS) (Webster and Linton, 2013). They provide a temporal and spatial assessment of annual seasonal increases and decreases in *Sargassum* algae amount over wide areas of the tropical Atlantic and Caribbean (Wang and Hu, 2016; Wang and Hu, 2017; Wang et al. 2019). Time series from remote sensing were coupled with spatial distribution models to determine the mechanisms that aggregate *Sargassum* algae along a zonal band in the tropical Atlantic considering possible nutrient sources promoting the observed annual blooms (Wang et al., 2018; Wang et al., 2019; Johns et al., 2020~~)).;~~ Jouanno et al., 2021b).

Tropical Atlantic currents and winds seasonally aggregate and carry these algae towards the Caribbean (Franks et al., 2016; Brooks et al., 2018; Cuevas et al., 2018). Modelling studies mainly focused on the transport properties of *Sargassum* rafts by offshore currents (Wang and Hu, 2017; Brooks et al., 2018; Maréchal et al., 2017; Putman et al., 2018, 2020; Wang et al., 2019; Berline et al., 2020).

Johns et al. (2020) extended this analysis to highlight anomalous transport due to the 2009-2010 NAO anomaly and seasonal aggregation by the Inter Tropical Convergence Zone (ITCZ).

A combination of ~~MODIS~~ satellite-based Alternative Floating Algae Index (AFAI ~~Satellite images,~~ Wang and Hu, 2016) fields with Hybrid Coordinate Ocean Model (HYCOM) surface current forecast data were used by Maréchal et al. (2017) to short-term predict *Sargassum* ~~strandings~~beachings for Guadeloupe and the French Antilles islands. Maréchal et al. (2017) showed that this short-term prediction system (i.e~~.,~~. detection starting within 50-100 km of the coasts) worked efficiently during the year 2015 with a performance percentage of 62% and a ~~stranding~~beaching forecast date uncertainty below one day.

Trinanes et al. (2021) presented the *Sargassum* Inundation Reports (SIR), a product based on satellite observations to weekly predict *Sargassum* coastal inundation potential throughout the Caribbean Sea region, the Gulf of Mexico, and extending to the east coast of Florida and the Bahamas. As described by Trinanes et al. (2021), the SIR algorithm uses the Floating Algae density values within 50 km of each coastal pixel to predict three inundation potential levels (low, medium, and high). This algorithm does not include ocean currents, winds, and waves which may modify the movement of *Sargassum.*

In the above works, the implementation of methods based on several independent data sets has led to the production of scientific knowledge and even to the development of large-scale forecasting systems. None of them used predictive modelling~~,~~ (Geisser, 1993; Kuhn and Johnson, 2013) including classifiers~~,~~ (Friedl and Brodley, 1997) to determine the probability of ~~a set of data belonging to another set in order to discover~~ repeatable patterns~~, allowing~~ in a dataset so as to produce a decision for risk prevention managers. Predictive modelling refers to mathematical and computational methods of predicting future events based on the analysis of the repeatable patterns in the input dataset (Geisser, 1993; Friedl and Brodley, 1997; Kuhn and Johnson, 2013). Compared to other conventional forecast, predictive modelling methods requiring low computational costs are characterized by their flexibility, and their intuitive simplicity (Friedl and Brodley, 1997).

In this paper, we propose to use clustering and decision tree classifier methods, combining ocean surface current~~and~~, wind reanalysis, an satellite-based *Sargassum* offshore abundance with past observed ~~strandings~~beachings to obtain a first predictive model of *Sargassum* beaching on the Caribbean coasts. This model will be used with forecast data as input to produce an operational decision support system.

As ocean data are spatio-temporal fields, machine learning methods such as K-Means (KMS) may be used to obtain a finite number of possible k-cluster partitions of the surface currents. These methods have been widely used in weather forecasting (Michelangeli et al., 1995; Cassou et al., 2004; Boé and Terray, 2008) but are much less common in physical oceanography (Harms and Winant, 1998; Hisaki, 2013). ~~To optimize the final partitioning, an additional metric based on the Kullback Leiber divergence (Kulback and Leibler, 1951, Biabiany et al., 2020) will be included.~~

We focused on the offshore region covering either side of the Lesser Antilles, between 55-66°W and 8-17°N (Fig. 1a). Visual analysis of the monthly SaWS maps indicates that this region remains the primary pathway for *Sargassum* rafts from the Atlantic Ocean to the Caribbean Sea. The North Equatorial Current (NEC), the Guiana Current (GC), the eddies and the retroflection front of the North Brazil Current (NBC) are the main contributors of this transport. Putman et al. (2018) modelled the percentage of *Sargassum* which follows these routes. Figure 1b describes the focused area divided into a first sub-set "LA1" for the Caribbean

3

Sea, a second one, ""LA2"" between 18°N and 14.5°N (Guadeloupe, Dominica, Martinique, Saint Lucia) and a third one ""LA3"" south of 14.5°N (Saint Vincent, Barbados, Trinidad and Tobago). This ocean region corresponds to the CA and TA1 boxes in Johns (2020).

The questions are as follows. Can dynamic patterns of surface currents in the Lesser Antilles be summarized as a discrete set of cases? What is their temporal recurrence? What combinations of currents enhance *Sargassum* rafts arrival and strandingsbeachings on the Lesser Antilles coasts? What is the contribution of this type of predictive modelling to the prevention of this new natural hazard?

The overall methodology, database, clustering methods and decision tree used in this study are described in Sect. 2. The obtained current regimes, their relationship to *Sargassum* hazard and the decision support system performances are presented in Sect. 3. These results are discussed in Sect. 4.

2 Datasets and methods

The overall methodology is presented in Fig. 2. The main goal of the first step was to use clustering analysis to identify the main current patterns in the Lesser Antilles during the period 2019-2020. The 30-day current patterns sequences leading to beachings were deduced based on beaching observations in Guadeloupe. An additional clustering analysis was conducted on these sequences to study the main patterns (orange box in Fig. 2). A decision tree classifier was built with the following input data: current patterns, 30-day sequences before beaching, satellite-based Sargassum abundance offshore Guadeloupe, surface currents from HYCOM and Mercator. This decision support system was tested for the period of the year 2021. The performance scores were assessed for each decision day and three temporal uncertainty ranges around this day, respectively: +/-1 days, +/-2 days, +/-3 days.

2.1 HYCOM surface current dataset

Daily 12 UTC (i.e. Coordinated Universal Time) surface current components from the 41-layer Hybrid Coordinate Ocean Model (HYCOM) global 1/12-degree analysis (HYCOM GLBy0.08 version), were examined. The HYCOM surface forcing including 10-m wind velocities are extracted from Climate Forecast System Version 2 (CFSv2). The Navy Coupled Ocean Data Assimilation (NCODA) system is used to assimilate available observational data: satellite altimeter sea surface height, satellite and in-situ sea surface temperature, temperature vertical profiles and salinity vertical profiles (Cummings, 2005; Cummings and Smedstad, 2013; Helber et al., 2013). The Bathymetry used is the GEBCO8 (Becker et al., 2009) with 30 arc second of resolution. The HYCOM GLBy0.08 grid resolution is 0.08 degree in longitude and 0.04 degree in latitude. To perform the present study, the native HYCOM fields were preliminarily interpolated on the Mercator uniform lon/lat 0.08-degree grid with a bilinear method. Putman et al. (2018) and Johns et al. Fine scale surface current data from the 1/25 degree HYCOM + NCODA Gulf of Mexico analysis model (GOMu0.04/expt_90.1m000 version, Hogan et al, 2014; Helber et al., 2013; Cummings and Smedstad, 2013; Cummings, 2005) between 1st January 2019 (i.e., available data starting date) and 31 December 2020 were analyzed. Daily 12Z fields giving the u and v components of the current at 50 cm depth were used. These fine resolution current data were not used in previous studies dealing with Sargassum hazard (Putman et al., 2018; Johns et al., 2020). (2020) used a previous version of HYCOM model including uniform lon/lat 0.08° scale grid to successfully simulate *Sargassum* trajectories.

4

a mis en forme : Police :Italique

a mis en forme : Police :Italique

a mis en forme : Police :Non Gras, Anglais (États-Unis)

a mis en forme : Anglais (Royaume-Uni)

2.2 Mercator surface current dataset

The daily ~~50-cm depth~~12 UTC surface current components from the 50-layer PSY4V3R1 Mercator 1/12-degree 3D analysis system ~~including the version 3.1 of the NEMO ocean model~~ (Lellouche et al., 2018; Gasparin et al., 2019) were also analyzed ~~along the same period as HYCOM~~. The atmospheric surface forcing are extracted from the 3-hourly ECMWF (European Centre for Medium-Range Weather Forecasts) IFS (Integrated Forecast System). Assimilated observational data ~~one. These data have a~~types are quite similar to HYCOM model. Unlike the HYCOM GLBy0.08 native grid including higher resolution ~~which is half that of the HYCOM data (1/25-degree scale). Therefore,~~in latitude (i.e. 0.04 degree), the Mercator ~~model gives a less accurate representation of the island shapes. Comparison between HYCOM~~ native grid is uniform in longitude and ~~Mercator results~~latitude with 0.08-degree scale. This would ~~help to~~ suggest that HYCOM may better ~~understand the effects of spatial resolution on surface current~~ reproduce small scale patterns ~~in the focused region~~than Mercator. Moreover, as described by Lellouche et al. (2018), the Mercator bathymetry includes GEBCO8 data in regions shallower than 200 m and the coarse 1 arc-minute ETOPO1 data (Amante and Eakins, 2009) in regions deeper than 300 m. The complex bathymetry of the Lesser Antilles Arc studied here could be less realistic in Mercator than in HYCOM.

2.3 ERA-5 dataset: surface winds

Surface wind influences the transport of floating seaweed rafts ~~, with an optimal factor of C_w = 0.01, which corresponds to the~~ and a drag or windage coefficient ~~or windage, following~~must be added to the surface currents. The value of $C_w = 0.01$ was used by Putman et al. (2018), Johns et al. (2020 ~~. A first clustering (KMS-L2) on Mercator analysis without windage had been proposed by Bernard et al. (2019).~~) and Berline et al. ~~(2017),~~(2020). The use of other windage values should be investigated ~~Putman et al. (2018) and Johns et al. (2020) have shown that the windage improves the Lagrangian simulations of Sargassum rafts transport~~ in ~~the Caribbean region. The windage was included in the present surface current clustering. Surface~~a further study. The daily 12 UTC surface wind data (at 1000 hPa) from the 31-km scale ERA-5 model ~~for the time period 2019 to 2020~~ were integrated with Mercator and HYCOM currents following this formula:

$$u_s(x,t) = u_m(x,t) + C_w u_w(x,t) \tag{1}$$

where u_s represents the oceanic surface currents with windage, u_m the oceanic surface currents velocity, C_w the windage and u_w the surface winds velocity. This approach is consistent with Putman et al. (201~~6~~8) and Johns et al. (2020) studies. ~~The region analyzed in the present work corresponds to the CA - TA1 region defined in Johns et al. (2020).~~

2.4 ~~Stranding~~Beaching observational data (Guadeloupe)

A referencing database including observed ~~strandings~~beachings on Guadeloupe coasts was used in the present study. The selected time period is the same as the one for surface current data: from 1st January 2019 to 31 December 2020. ~~This~~During this period ~~includes~~of 730 ~~observational~~ days ~~with,~~ only 110 days of observed ~~strandings~~beaching were recorded (i.e. 30 days in 2019 and 80 days in 2020). During the year of 2021, 78 days of beaching were observed in Guadeloupe. These observational data based on remote sensing and in situ data are archived online by the Regional Directorate for Environment, Development and Housing in Guadeloupe (http://www.guadeloupe.developpement-durable.gouv.fr/sargasses-r999.html).

2.5 Satellite-based offshore abundance of *Sargassum*

Sargassum satellite observations were included in the present decision support system. To quantify the abundance of *Sargassum* in an area of 100 km radius offshore Guadeloupe, the 7-day Floating Algae (FA) density fields derived from the Alternative Floating Algae Index (Wang and Hu, 2016) were analyzed. As described by Trinanes et al. (2021), the 7-day Floating Algae (FA) density fields are accumulated on 7 days and have a 0.1° resolution. Due to optical complexity in nearshore waters, the FA density fields are masked with missing values within 30 km from shoreline (Trinanes et al. 2021). The cumulative FA density values were added up in the area 30-100 km offshore Guadeloupe (Fig. 1) then averaged over the two years 2019 and 2020 for each day.

2.6 Clustering analysis with expert ~~deviation~~distance

~~To process these~~Unsupervised learning methods such as Hierarchical Agglomerative Clustering (HAC) and K-means algorithms were used in the present study. Besides the measures and the classes of distance between objects such as the Euclidean distance for K-means and the Ward method which allows to identify homogeneous subsets of data~~, we shall use~~ (Ward, 1963), a new ~~approach called~~metric was also added. The Expert ~~Deviation~~Distance (ED) which integrates image analysis within unsupervised learning methods (Clustering) ~~to bring field knowledge into the analysis process (Biabiany et al., 2020).~~was used. This method allowed significant improvement in clustering analysis dealing with climate data characterized by high spatio-temporal variability, such as precipitation (Biabiany et al., 2020).

2.5.1 Clustering methods ~~process~~

~~Unsupervised learning methods such as Hierarchical Agglomerative Clustering (HAC) and K means algorithms are used in the present study. Besides the measures and the classes of distance between objects such as the~~using Euclidean distance ~~for K means and the Ward's method for HAC, a new metric was also added~~(L2) can lead to group different physical situations within the same cluster (Biabiany et al~~.~~., 2020). ~~This~~The ED metric integrates a set of knowledge about the dynamics of the data to be partitioned as well as their spatio-temporal properties. ~~The result is an automated analysis with its own expertise on the input data.~~

2.5.2 Use of Expert Deviation in clustering algorithms

~~The ED metric, which seems more suitable for this study, was used. L2 clustering methods can lead, within the same cluster, to gatherings of different physical situations (Biabiany et al., 2020). To remove these biases linked with L2 clustering, the first step of the method used here is to consider the spatial variability in the dynamics of the analyzed daily surface currents from L2. The LA study area was separated into three parts (Fig. 1b~~This ED is based on an empirical spatial subdivision and the use of Kullback-Leibler divergence, in order to quantify the similarity between two fields. Figure 3 shows the schematic of the Expert Distance process adopted here.

The LA study area was separated into three parts (Fig. 1b, Fig. 3) based on the *Sargassum* rafts transport centers of action reported in the literature (Franks et al., 2016; Berline et al., 2020). To the west of LA, the first zone, LA1, is centered on the Caribbean Sea. To the east, the Atlantic zone ~~has been~~was split into two areas towards 13.5°N, just above the island of Barbados ~~island~~. To the south-east is the LA3 zone under the influence of the North Equatorial Recirculation Region (NERR) and its retroflection rings, while to the north-east is the LA2 zone, more representative of the North Equatorial Current. The analyzed daily fields include a total of 14 279 ~~meshes~~grid points (4 282 ~~meshes~~grid points in LA1, 3 407 ~~meshes~~grid points in LA2 and 4 536 ~~meshes~~grid points

6

in LA3). The remainder corresponds to areas over land ares.(e.g., islands).

The second step was to group the information carried by the daily current velocity fields conditionally to the three given zones into histograms. The similarity of the most similar fields is estimated per pair and per zone based on the symmetrized Kullback-Leibler (KL) divergence computed from the histograms (Kullback and Leibler, 1951). This allows the entropy between two distributions to be expressed without having a priori reasoning concerning the probability distribution. The similarity between two histograms was quantified this way. The last step consisted in calculating the average of the divergence values for each zone. This allows to have a single value, named Expert Distance (ED) quantifying the similarity between the individuals of the database during clustering. The clustering results have been evaluated using the Silhouette Index (Rousseeuw, 1987).

The SaMkThe clustering results were evaluated using the Silhouette Index (Rousseeuw, 1987). The Silhouette (SaMk) index defined in Biabiany et al. (2020) was used. This allows to express the quality of a clustering, by the average of the quality of each cluster, which is itself the average of the silhouette indices $s(i)$ over the cluster elements. This index is defined as follows:

$$SaMk = \frac{1}{k} \times \sum_{j=1}^{k} \frac{1}{|Cj|} \times \sum_{i \in Cj} s(i) \tag{2}$$

where k is the number of clusters, C_j the set of days from the cluster j, i a day form C_j and $s(i)$ the silhouette index (Rousseeuw, 1987) value of day.

The current pattern clusters obtained are related by a set of days in common. Match percentages were calculated using the following formula.

$$p_{(m,h)} = \frac{|C_m \cap C_h|}{|C_m \cup C_h|} = \frac{N_{(m,h)}}{|C_m| + |C_h| - N_{(m,h)}} \tag{3}$$

where $p_{(m,h)}$ is the percentage of correspondence between cluster C_m and cluster C_h derived from Mercator and HYCOM datasets respectively. $N_{(m,h)}$ is the number of days shared by these two clusters.

2.67 Clustering analysis on stranding backward current sequences leading to beachings

To better understanding current regime dynamics which may lead to *Sargassum* strandings on the coasts ofbeaching in Guadeloupe, we analyzed the past stranding 30-day current backward sequences were analyzed.before beaching. The 30 days duration corresponds to the empirical transport time of a passive particle moving from the main entrance location of *Sargassum* rafts in the Lesser Antilles area (i.e. in LA3 zone, 8° N; -55° E) to Guadeloupe (i.e. LA2 zone). Based on the mean current magnitude of 0.2 m s^{-1} (average value over the LA zone, in HYCOM and in Mercator data) and the distance of 500 km between the main entrance location and the Guadeloupe coasts, 29 days are obtained for the transport. For simplicity, the duration of 30 days was selected instead of 29 days. While 110 observed strandingbeaching days were registered between January 2019 and December 2020, only 107 back sequences were studied here. This is explained by the fact that strandingbeaching days registered in January 202019 were removed to avoid back the sequences missing data of the December 2018 period. These 107 stranding back beaching sequences were examined with the highest resolution surface current model, i.e., HYCOM fields. Dissimilarities between these backward sequences were calculated with optimal matching methods before dividing the populationsequences dataset into several groups using a hierarchical classification (Larmarange et al., 2015). The Longest Common Subsequence (LCS) method was used to compute the distances between the backward sequences (Elzinga and Struder, 2015; Studer and Ritschard, 2016). A dendrogram

7

was calculated using Wa~~l~~rd's algorithm~~.~~ (Ward, 1963). The highest relative inertia loss criterion allowed to determine the optimal number of partitions (TraMiner package (Gabadinho et al., 2011)). The stages of this clustering process were summarized in Fig. 4.

2.~~7~~8 Decision support system

To determine the probability of *Sargassum* ~~stranding at a given location~~beaching, a decision tree was built using complementary elements called "modules" (Fig. ~~2). They each generate~~5). Each module generates information based on input data including surface currents (Mercator and HYCOM) with windage effects (~~Mercator, HYCOM and~~ ERA-5) and past observations of ~~strandings~~beachings in Guadeloupe. Thus, for a given day, the proposed system works as follow:

- Module A takes as input the ~~month~~week number of the selected day and returns the associated ~~monthly~~daily probability ~~(frequency)~~to reach the maximum offshore abundance of ~~stranding;~~*Sargassum* (based on observational FA density values during the two years 2019 and 2020);
- Module B ~~which~~ assigns a cluster number to the focused day after the ED clustering of the daily surface currents. Then, ~~it builds~~ from this day, it builds empirical ~~backward~~ sequences of numbers between 1 and 4 (type of cluster) over a period of 30 past days;
- Module C ~~which~~ takes as input the daily cluster number produced by module B and returns the probability (frequency) of ~~stranding~~beaching associated with the type of cluster. This probability is calculated, by cluster type, from the ~~strandings~~beachings observed on the coasts of the Guadeloupe archipelago. The system has 107 30-day ~~stranding backward~~current sequences~~.~~ before beaching. These ~~backward~~ sequences start on the day of ~~standing~~beaching on the coasts of Guadeloupe. This set of referenced ~~stranding backward~~30-day current sequences before beaching is called BASE (Fig. ~~2b~~5b);
- Module D~~, which~~ compares the ~~backward~~ sequence of the given day to the ~~stranding backward~~referenced current sequences before beaching with Jaccard distance. ~~The module~~Module D is interconnected to BASE and module B. It returns the percentage of correspondence between them.

In the literature, the average of the different modules is often used as the decision operator (Bo. et al., 2020; Swain and Hauska, 1977). In the present work, the percentage of ~~stranding~~beaching for a given day was determined using the percentages provided by modules A, C and D, according to the following formula:

$$P(i) = \left(A(i) + \frac{C(i)}{D(i)}\right) \tag{\sout{3}4}$$

where P(i) is the quantity used in the design of the decision rule. This rule is simply the linear combination of the percentages from modules A, C and D, calculated according to:

$$\sout{DECISION(i) = P(i) > \frac{1}{|R|} \times \sum_{j \in R} A(j) + \frac{C(j)}{D(j)}} \tag{4}$$

$$DECISION(i) = P(i) > Mean(P(j)) \tag{5}$$

where $j \in R$~~.~~. ~~t~~The set of past days (2019-2020) and DECISION(i) is a (logical) response of the decision tree for a given day i.e.~~.~~. expressed in binary form. ~~The proposed tree in Fig. 2 was experimented on the first 120 days of the year 2021, from 1ˢᵗ January 2021 to 30 April 2021, i.e., 120 tests.~~

8

The proposed tree in Fig. 5 was tested on the full year of 2021 except 31 December 2021 including missing data, i.e. in total 364 tests. The performance assessment of the decision support system was assessed for each decision day and three temporal uncertainty ranges around this day, respectively: +/-1 days, +/-2 days, +/-3 days.

3 Results

3.1 Surface current patterns in the focused area

In view of the lack of study dealing with surface current patterns in the Lesser Antilles area, this preliminary analysis is presented here. The deciles of surface current velocities including windage, according to equation (1), are presented in Table 1. For both models HYCOM and Mercator, the maximum surface velocity intensities do not exceedis 2.57 m s^{-1} and 90% of them remain below 0.65 m s^{-1}. (the respective 90th centile values are respectively 0.6515 m s^{-1} and 0.6458 m s^{-1} for HYCOM and Mercator). The Mercator data have a median of 0.28 m s^{-1}, the mean of 0.33 m s^{-1}, while for HYCOM these values are respectively equal to 0.32 m s^{-1} and 0.36 m s^{-1}. The ratio between the first and the last decile is close to 6. Figure 36 shows skewed distributions with skewness equal to 1.31 and 1.21. for HYCOM and Mercator, respectively. A normal kernel was used to obtain these distributions. The distribution mass is concentrated on the left. There are extreme values indicating surface current speeds with deviations 5 times greater than the standard deviation.

To assess the contribution of each of the three regions (i.e.,. LA1, LA2, LA3) to the deciles, the relative frequency against the decile thresholds given in Table 1 is shown in Fig. 47. Three different shapes can be seen. In the Caribbean Sea, the LA1 relative frequency distributions from HYCOM and Mercator are almost horizontal, indicating a quite constant contribution (~3%) over all velocity classes. In the Atlantic Ocean (i.e., are. in the area including LA2 and LA3), HYCOM and Mercator current speed distributions are quite similar. The frequency distributions show two opposite behaviors respectively for LA2 and LA3. In the Atlantic north LA part, LA2 area, the frequency decreases with current speed. The current speeds above 0.65 m s^{-1} are very uncommon. On the contrary, in the Atlantic south LA part, LA3 area, the frequency increase is observed with maximum frequency linked with current speeds above 0.65 m s^{-1}. These three significant specific current speed distributions associated with LA1, LA2 and LA3 confirm the need to separate these three areas in the ED metric clustering process.

The differences between HYCOM and Mercator current vectors were also examined for each grid cellpoint (Fig. 58). Globally, at open sea, the current speed differences are relatively small and remain below 0.15 m s^{-1}. These differences between HYCOM and Mercator increase close to the islands with an average value of 0.3 m s^{-1}. The largest differences, above 0.5 m s^{-1} are observed inIn the South part of the LA arc, around Trinidad and Tobago, Mercator current magnitudes are globally higher than HYCOM current magnitudes. Thus, Mercator surface currents might induce higher *Sargassum* influx from the Western Central Atlantic to the Caribbean Sea in this area.

At each grid point, the angular deviations found between the medians of the surface current velocity vector directions can be divided into three magnitude groups of 45°.° intervals. The current direction differences between 0 and 45° are the most frequent group in the region, while those between 45 and 90° remain localized downstream of the islands. Finally, those above 90° occur exclusively around Trinidad.

3.2 Clustering analysis

9

To identify surface current patterns in the region, and then those that lead the transport of *Sargassum* rafts to the LA islands coasts, the clustering of the gridded data according to equation (1) was performed.

3.2.1 Clustering assessment

One of the known uncertainties in the k-means method is induced by the selected number of clusters. To find an optimal number of clusters and identify the best partition (Biabiany et al. 2020), the silhouette index (SaMk) evolution against the number of clusters, k. is shown in Fig. ~~6~~9. The silhouette indices obtained by the KMS-ED method, are in general above 0.2 for any k<15, and remain higher than those ~~from~~obtained by the KMS-L2, HAC-L2 and HAC-ED methods. These values indicate that the quality of the clusters is much better with the KMS-ED method. The inflection point of the KMS-ED curve occurs for the same number of clusters, k=4, for both Mercator and HYCOM data. This highlights four representative current regimes in the studied region, respectively named MC1, MC2, MC3, MC4 for Mercator and HC1, HC2, HC3 and HC4 for HYCOM.

3.2.2 Visual analysis of current regimes

The four types of surface current circulation, obtained in intensity and direction, are shown in Figs. ~~7~~10 and ~~8~~11, respectively for the Mercator and HYCOM analysis. The parangon which is the closest day to the centroid, was chosen to represent each type of cluster. The four clusters may be distinguished by the NBC expansion and by the induced retroflection ring locations. The surface current velocities and their associated streamlines are driven by the following structures:

- those which enter through the Caribbean Sea from the south, remaining almost parallel to the continental shelf. They occur in the LA3 and LA1 regions;
- Those due to the propagation of the eddies dynamic characteristics related to the retroflection rings of the NBC. They are coming from the south of the LA3 region, along the Atlantic side of the Lesser Antilles arc, before passing through the Caribbean Sea towards 12-14° N;
- Those generally coming from the northeast of the LA1 and LA2 regions, representing the southern limit of the subtropical gyre which cuts the Lesser Antilles at about 15° N. They keep their initial direction and are sheared by the South-East currents.

The number of days corresponding to each cluster is given in Table 2. MC1, HC2 and HC3 are the most common along the studied period. Each of them represents almost 30% of the daily output. However, none of the four clusters really stands out. For both analyses, the differences between cluster occurrences stay lower than 10%.

3.2.3 Matching days between clusters

The clusters found are also related by a set of days in common. Match percentages ~~have been~~ calculated using the following formula.

$$p_{(m,h)} = \frac{|C_m \cap C_h|}{|C_m \cup C_h|} = \frac{N_{(m,h)}}{|C_m| + |C_h| - N_{(m,h)}} \quad (5)$$

~~where $p_{(m,h)}$ is the percentage of correspondence between cluster C_m and cluster C_h derived from Mercator and HYCOM datasets respectively. $N_{(m,h)}$ is the number of days shared by these two clusters.~~were calculated using the following Eq. 3. Table 3 shows the results.

of the match percentages (Eq. 3). MC4 - HC2 is the cluster pair with the ~~most important~~highest match score (69.8%). It is followed

10

by the pair MC2 - HC1 (60.4 %), then MC3 - HC3 (56.7 %) and MC1 - HC4 (50,6%).

3.2.4 Distribution and comparison of intensities

335 Deciles were used to study and analyze the velocity distributions characterizing each cluster. Evolutions of the relative frequency of Us(x,y,t) as a function of the deciles (Table 1) are shown in Figs. 912 and 1013. For the entire analysis, the values of the deciles remain fixed and constant, and the curves are plotted for the three regions described in Fig. 1.

For both models, globally, three main patterns are identified. The first pattern includes the following clusters: MC1, MC3, HC1 and HC3. This pattern is characterized by the increase of the relative frequency curve in the LA1 and LA3 regions and its decrease in 340 the LA2 region. The elements of these clusters include strong current velocities above the median of 0.28 m s^{-1}. The second pattern includes MC2 and HC2 clusters which are characterized by the decrease orof the relative frequency for the three regions (i.e., LA1, LA2, LA3). The last pattern includes MC4 and HC4 clusters and corresponds to three concave curves with maximums located at different velocity thresholds depending on the region under study.

To examine possible relationships, for a given region, between the two variables, decile speed thresholds and identified clusters, 345 contingency tables were constructed (not shown) and the chi-squared test was performed. For the three areas, the p-value was much lower than 0.01. The chi-squared test results indicated that for the LA1, LA2 and LA3 regions, the speed distribution depends on the identified cluster.

3.2.5 Seasonality

The monthly distribution of each cluster is plotted (Figs. 1114 and 1215). Differences are relatively clear for both model analyses 350 with a marked seasonal variation. The MC3 and HC3 regimes are observed during the first half of the year with a maximum in March, followed by MC2 and HC1 from April to July. The last two regimes are observed from August to December. The pair MC4 HC2, reaches a maximum in September while MC1 and HC4 persist until February of the following year.

3.3 Links with *Sargassum* strandingsbeachings

As with many floating objects, before coming ashore on the coasts of the LA, *Sargassum* algae accumulate on the ocean surface in 355 large amounts and form slicks, or filamentary structures, interspersed with void areas, under the influence of currents. These dynamic structures regularly observed from satellites, aircraft, and ships, have a certain inertia (Maximenko et al., 2012).

Beyond biological production, it is therefore the specific dynamic conditions of the surface currents and the surface winds which may lead to massive *Sargassum* strandingsbeachings on Caribbean coastal areas.

The monthly evolution of observed stranding days on the Guadeloupe coasts, the monthly evolution of *Sargassum* abundance overin 360 the Central Atlantic region (SaWS, https://optics.marine.usf.edu/projects/SaWS.html)area 30-100 km offshore Guadeloupe were also analyzed on the focused period 2019-2020 (Figs. 1114 and 1215). During these two years, the amount of *Sargassum* overlikely to enhance the Central Atlantic regionbeaching risk in Guadeloupe increased significantly from February to JulyMay, then decreased from JulyMay to November.

Two strandingbeaching peak values are found: one in March and the second in August. The strandingsbeachings dates and the 365 cluster occurrence dates were also compared in Table 4. The MC3 - HC3 pair gather the greatest number of similarities, followed by the MC1 and HC2 clusters.

11

a mis en forme : Police :Italique

a mis en forme : Police :Italique

a mis en forme : Police :Italique

a mis en forme : Police :Italique

a mis en forme : Anglais (États-Unis)

a mis en forme : Police :Italique, Anglais (États-Unis)

a mis en forme : Anglais (États-Unis)

a mis en forme : Anglais (États-Unis)

a mis en forme : Anglais (États-Unis)

a mis en forme : Anglais (États-Unis)

a mis en forme : Anglais (États-Unis)

~~The pairs (MC1, HC2) and (MC3, HC3) include the greatest number of observed stranding days in Guadeloupe (Table 4).~~ These pairs of clusters would be favourable to the transport of these algae toward the coasts of the Lesser Antilles islands. MC2 and HC1 are the two clusters with the smallest number of ~~stranding~~beaching days.

3.4 Current ~~regime backward~~ sequences leading to ~~strandings~~beachings

The HAC clustering analysis on the current regime ~~backward~~ sequences leading to observed ~~stranding~~beaching days allowed to distribute the 107 ~~backward~~ sequences into four classes, respectively called Seq1, Seq2, Seq3 and Seq4. This analysis integrated only the HYCOM surface current data which have a greater resolution than Mercator. During the focused period (i.e~~.,~~. 2019-2020), Seq4 (39.3%) and Seq2 (37.4%) have the greatest occurrence (Table 5). Seq1 and Seq3 have a respective occurrence of 16.8% and 6.5%. Figure ~~13~~16 shows that Seq2, Seq3, and Seq4 are characterized by the respective modal current regimes HC3, HC1, and HC2. For the Seq1 ~~backward~~ sequences, there is no clear prevalent current regime. The monthly distribution of the main ~~backward~~ sequence classes, Seq2 and Seq4, highlights a significant seasonal splitting (Fig. ~~14~~17). The Seq2 ~~backward~~ sequences occurred from December to June while the Seq4 ones occurred from July to November. These two distributions also seem ~~also~~ significantly correlated with the monthly occurrences of observed ~~strandings~~beachings. While the first ~~stranding~~beaching peak occurring in March is linked with the Seq2 maximum occurrences, the second ~~stranding~~beaching peak occurring in August is linked with the Seq4 maximum occurrences.

3.5 Decision support system results

~~Table 6 presents the results obtained for Mercator and HYCOM. For HYCOM, results show that the model (clustering + decision tree) correctly classified 41.7% (37.5% for Mercator) of the observations in stranding class, and 81.3% (54.2% for Mercator) of the observations among non-strandings (Table 6). Overall, the performance of the decision tree reached 50.8% for the Mercator database and 73.3% for HYCOM.~~ The behavior of each module is presented in Fig. ~~15. In general, modules A and C remain with~~18. Globally, Module A probabilities ~~below 0.2 and 0.5 respectively. These two values are expected. They are induced by the strong spatio-temporal variability of the phenomenon.~~

~~The percentages of stranding per cluster associated with module C show empirical probabilities close to 0.3 indicating that one third of the days in the concerned clusters are stranding days. Module~~ based on *Sargassum* offshore abundance probability seem the most correlated with the DSS decision. However, during the months with low *Sargassum* offshore abundance (i.e. from September to December) Modules C and D ~~produces empirical probabilities~~ related to ~~the links~~current patterns are the main contributors to the decision. Module D based on the comparison between ~~the~~ past observed sequences and the sequence~~s~~ corresponding to the forecast day~~. In our case, they~~ remains with high probabilities above 0.5. These probabilities can reach 0.95 ~~(Fig. 15a)~~ indicating strong similarities between the sequences. Module C associated with the percentages of beaching per cluster shows empirical probabilities close to 0.3 indicating that one third of the days in the concerned clusters are beaching days.

Table 6 presents the performance results of the predictive model (i.e. clustering + decision tree) for Mercator and HYCOM. Overall, the use of HYCOM data allows to improve the prediction of beaching and non-beaching days (Table 6). With forecast date uncertainty below one day, the HYCOM DSS has an accuracy of 54.1% (i.e. beaching and non-beaching days) and predicts 59.0% of the observed beachings (i.e. true positive) in the year 2021. At the same date precision, the Mercator DSS has an accuracy of

12

50.6% and predicts 55.1% of the observed beaching days. The performance differences between the two datasets tend to increase with the temporal uncertainty ranges around the forecast date. With a temporal uncertainty range +/-3 days, the HYCOM DSS reaches an accuracy of 70.1% and predicts 73.1% of the observed beachings in the year 2021. At the same date precision, the Mercator DSS presents an accuracy of 58.2% and predicts 65.4% of the observed beachings.

4. Discussion

4.1. Performance indices and clustering quality

The performance of the clustering and the quality of the clusters were assessed using the silhouette coefficient. The evolution of this coefficient (Fig. 6) shows9) clearly shows that on the one hand, the methods based on the HAC algorithm produce lower values than those obtained by the KMS algorithms. On the other hand, for ED, silhouette indices are largely above those found by the L2 distance as written by Biabiany et al. (2020). This silhouette coefficient evolution allows us to keep four representative types of current regimes in this part of the Caribbean region. However, due to the lack of works for this region, comparisons between the present results and other studies were very limited. In other studies, authors have proposed a similar number of dominant regimes on a large scale, in the tropical Pacific (Fereday et al., 2008), for the determination of robust modes of Northern Hemisphere Sea ice variability (Fučkar et al., 2016), or for ocean mapping from environmental data (Zhao et al., 2020).

In our case, the velocity distributions show four singular profiles confirming the good performance of the clustering. Each cluster also had distinct monthly distributions. This analysis allowed to better understand the variability of the surface current circulations in this region.

4.2. Surface current analysis applied to Sargassum hazard

In terms of spatial distribution, clusters show notable differences for both types of model analysis and three variability factors can be identified.

The first one is the seasonal evolution of the NBC retroflexction front (Baklouti et al. 2007). The NBC feeds the Guiana Current (GC) but also separates sharply, near 6°–8°N, from the South American coastline and retroflects to feed, this time, the eastward NECC. Isolated large rings move north-westward toward the Caribbean Sea, on a course parallel to the South American coastline, then interact with the Lesser Antilles (Fratantoni et al., 2002, 2006). These two dynamic structures, GC and NBC rings, contribute significantly to the transfer of South Atlantic surface water to the Caribbean. These dynamic structures were found on the four identified clusters and seem to work year-round with intensity variations.

Another part of this variability is caused by the rings of the NBC that move northwestward from the equatorial Atlantic and interact with the steep topography of the Lesser Antilles arc. MC2 and HC1 are two typical cases. Interactions with the island chain cause significant disturbances of the inflow through the southern passages with a blocking. This provides a meridional transport of surface water northward, along the LA arc (Fratantoni and Richardson, 2006; Huang et al., 2021). The Lesser Antilles arc clearly diverted the initially northwestwardnorth-westward drift of the NBC rings to a more northward course parallel to the island arc. Johns et al. (2002) have shown that the crossing of the Atlantic inflow to the Caribbean Sea through the passages of the Windward Islands (i.e., Lesser Antilles south islands from Trinidad to Martinique) has a highly asymmetric seasonal cycle, with a maximum in June and a minimum in September–October. The annual distribution of MC2 and HC3 clusters is close to that found by Johns et al. (2002).

13

The last identified factor, is related to ~~surface currents present in~~ the North Atlantic ~~region due to the North Current~~Gyre and the associated North Equatorial Current. As the seasons change from winter to summer, the gyre ~~circulation~~shifts South by a few degrees in latitude. In this part of the study area, several clusters show lower current speeds and areas with large angular deviations in direction have also been identified. In the LA2 area ~~(i.e.,~~ under influence of the north-east Trade Winds (i.e. Atlantic area between 14.5°N and 18°N), the relative frequencies of above-average speeds are the lowest ~~with the northeast Trade Winds.~~. The wind-current shear zones are also the most extensive. The wind-driven flow occurs from the subtropical gyre location to 15°N, near the Martinique island (Johns et al., 2002). Passages through the Leeward islands have a maximum inflow in September and a minimum one in June.

The comparison between the large-scale meteorological situations corresponding to the parangons showed that the main differences between the current regime clusters are related to the location and the extension of the high-pressure centers, the positioning of the ITCZ, and the intensity of the low Caribbean Level Jet.

4.2. *Sargassum* beachings

All clusters contain ~~stranding~~beaching days in relative abundance, 12 to 36 % of ~~stranding~~beaching days for the two years 2019, 2020. ~~The monthly distribution of clusters and the distribution of observed strandings in Guadeloupe are out of sync.~~ The first peak of ~~strandings~~beachings, in March ~~and~~, seems linked with the maximum frequency of MC3 and HC3 clusters. The second peak of observed ~~strandings~~beachings occurs in August and seems associated with the MC1, HC2 and HC4 clusters. Johns et al (2020) found that windage forcing induced by the wind convergence accumulates *Sargassum* rafts within the ITCZ between April and September. This accumulation would contribute to the observed ~~stranding~~beaching peak in August. The clustering analysis on the ~~stranding~~beaching current ~~backward~~ sequences confirmed that the recurrence of HC3 (between December and June) and HC2 (between July and November) would induce large ~~strandings~~beachings on the Guadeloupe coasts during these respective periods. The HC2 current regime is characterized by the prevalence of the North Atlantic gyre with weak velocities in the Western Central Atlantic and zonal streamlines. As for the HC3 current regime, it is characterized by strong Guiana Current with high velocities in the LA3 region and meridional streamlines almost parallel to the Lesser Antilles Arc.

4.3 Predictive model performance

A machine learning based method for predicting *Sargassum* beaching was proposed and was built from a decision tree. This method has already been used for other parameters and it allows to improve both the prediction accuracy and the fully black-box effect of the neural network. Compared to usual parametric statistical methods, it can effectively overcome the multicollinearity of ~~independent~~explanatory variables (e.g., ocean current and surface wind). ~~The~~Depending on the temporal uncertainty ranges, the accuracy of the present decision tree ~~reaches 73.3~~is between 54.1% and 70.1% for HYCOM against 50.~~86~~% and 58.2% for Mercator (Table 6). The best performance scores are reached with the largest temporal uncertainty range +/-3 days. Similar performance scores were found for decision trees predicting summer rainfall in Chongqing (China) (Bo et al., 2020) or landslide hazard in the Yen Bai Province (Vietnam) (Pham et al., 2020). ~~However, asymmetric performances have been highlighted with better results for true negatives than for true positives (Table 6). These can be attributed to the algorithm and to the weak ability of the model to handle different data sets. These prediction errors are greater for Mercator.~~

14

During the year of testing (i.e. 2021) only 78 beachings days were observed. Despite this large difference between beaching and non-beaching occurrences, the predictive model produces quite symmetric performances for both true positives and true negatives (Table 6). This fact highlights the good ability of the model to handle the different chosen datasets. This stability of the decision support system tends to increase with the temporal uncertainty range. For HYCOM, this asymmetric performance difference drops below 4% at +/-3 days precision.

Several ways to improve the predictive model were identified. The lack of observational data in time (i.e., only two years) may weaken the final decision and induce overfitting. The tree could also be improved by weighting and prioritizing the different modules so as to increase their relevance. The improvement of the results can be found by optimizing the proposed decision calculation rule (3) to better integrate the characteristics of the observed phenomenon. The daily probability of Sargassum offshore abundance produced by Module A would also be improved with better quality *Sargassum* remote sensing observations, particularly the FA density data gap within 30 km of the coasts. The present study does not take into account the effects of other factors (e.g., presence of nutrient, sinking of algae and waves) which would allow a more realistic understanding of the *Sargassum* beachings.

5. Conclusion

For a decade, the Caribbean countries, and particularly the LA, have suffered from the impacts induced by the massive and regular arrival of *Sargassum* on their coastal areas. This study presents the application of a clustering approach to determine the types of surface current circulations integrating the additional wind drift and their possible links with the *Sargassum* beachings observed on the LA coasts. The Guadeloupe archipelago was chosen as beaching observational site for the period 2019-2020. This analysis was performed using the most recent versions of ocean current analysis, Mercator and HYCOM. The surface wind speed data from the ERA-5 model were also used. The Clustering of the spatiotemporal surface current fields including windage was produced using the K-Means algorithm combined with the expert distance metric. Silhouette index was used to determine the optimal number of clusters.

For this region (8-17°N, 66-55°W) divided into three sub-regions, we identify four coherent patterns from data sets. They contain the current structures related to the Guiana currents, the branches of the subtropical Atlantic gyre, the front and the retroflection rings related to the NBC.

The finer resolution of HYCOM analysis provided more detailed information on surface current velocities near the islands than Mercator fields (i.e., mean local velocity difference of 0.3 m s^{-1}). Offshore, these differences remain very small.

Links between clusters and observed beachings in Guadeloupe were studied considering windage, parangon velocity distributions and monthly abundance maps. The surface current circulations characterizing the (MC3; HC3) and (MC4; HC2) cluster pairs seemed the most favorable for the transport and the beaching of *Sargassum* on the Lesser Antilles coasts.

The clustering analysis on the beaching current sequences based on HYCOM fields confirmed that the recurrence of HC3 (Seq2, between December and June) and HC2 (Seq4, between July and November) would induce large beachings on the Guadeloupe coasts during these respective periods. While the HC2 current regime is characterized by the prevalence of the North Atlantic gyre with weak zonal velocities, the HC3 current regime is marked by the influence of the NBC, the induced retroflection rings and strong Guiana Current leading to higher meridional velocities in the LA3 region.

Machine learning algorithms (KMS, ED, decision tree classifier) were applied to estimate the probability of *Sargassum* ~~strandings~~beachings in Guadeloupe, based on: surface current forecasts, current regime ~~backward~~ sequences and several combinations of probabilities. The performance score of this predictive model showed that ~~the finer resolution of HYCOM (i.e., lower than 5 km scale)~~HYCOM seems more suitable to reproduce small-scale current patterns inducing or not ~~strandings~~beachings in the Lesser Antilles. ~~The~~ For one year of tests (i.e. 2021), the decision tree accuracy reached respectively ~~50.8~~70.1% and ~~73.3~~58.2% for HYCOM and Mercator ~~and HYCOM~~.with a temporal uncertainty range +/-3 days around the forecast date. This accuracy could be improved by weighting and prioritizing the different modules. ~~New modules~~The daily probability of Sargassum offshore abundance used in the decision tree would also be ~~added like~~improved with better quality Sargassum remote sensing observations, particularly the FA density data gap within 30 km of the coasts.

Due to the very recent availability of the selected HYCOM new generation version, the present study was conducted only on two years (i.e.~~.,~~ 2019-2020). The studied period could be extended to more years to integrate the inter-annual variability of the surface currents.

Nevertheless, the obtained results are very encouraging and open new possibilities for the forecasting of this natural hazard type. Machine learning methods developed in this analysis proved to be useful in the prevention of a natural risk depending on physical multifactorial combinations.

The present clustering analysis predictive system could be applied to other islands of the Lesser Antilles changing the observational ~~stranding~~beaching site. The association of clustering methods and decision trees requiring low computational costs may enhance existing operational systems to help decision-makers in the *Sargassum* risk management. Maréchal et al. (2017) restrained the starting point of their operational short-term forecast system within 50-100 km of the LA coasts in order to reduce prediction errors. This geographical limit would correspond to a forecast period of 1-2 days before beaching. The present regional information on current dynamics leading to the arrival of *Sargassum* near the islands would be useful to extend this limit. In this way, it could be easier to anticipate the implementation of the resources needed to collect the *Sargassum* algae on the shorelines. Like the *Sargassum* Inundation Reports (Trinanes et al., 2021), the present small-scale *Sargassum* beaching predictive model may contribute to the region-wide efforts to help coastal communities managing this hazard.

References

16

~~Anses: Expositions aux émanations d'algues sargasses en décomposition aux Antilles et en Guyane, Technical report, ANSES, Maisons-Alfort, France, 162 pp, available at: https://www.anses.fr/en/system/files/AIR2015SA0225Ra.pdf, 2017.~~

Allshouse, M. R., Ivey, G. N., Lowe, R. J., Jones, N. L., Beegle-Krause, C. J., Xu, J., and Peacock, T.: Impact of windage on ocean

540 surface Lagrangian coherent structures, Environ. Fluid Mech. 17, 473–483, https://doi.org/10.1007/s10652-016-9499-3, 2017.

Amante, C. and Eakins, B. W.: ETOPO1 1 Arc-minute global relief model: procedures, data sources and analysis, NOAA Technical Memorandum NESDIS NGDC-24, Marine Geology and Geophysics Division, Boulder, Colorado, 25 pp., https://doi.org/10.1594/PANGAEA.769615, 2009.

Anses: Expositions aux émanations d'algues sargasses en décomposition aux Antilles et en Guyane, Technical report, ANSES,

545 Maisons-Alfort, France, 162 pp, available at: https://www.anses.fr/en/system/files/AIR2015SA0225Ra.pdf, 2017.

Arnault, S., Thiria, S., Crépon, M., and Kaly, F.: A tropical Atlantic dynamics analysis by combining machine learning and satellite data, Advances in Space Research, 68, 2, 467-486, https://doi.org/10.1016/j.asr.2020.09.044. 2021.

Baklouti, M., Devenon, J.-L., Bourret, A., Froidefond, J.-M., Ternon, J.-F., and Fuda, J.-L.: New insights in the French Guiana continental shelf circulation and its relation to the North Brazil Current retroflection, J. Geophys. Res., 112, C02023,

550 doi:10.1029/2006JC003520., 2007.

Becker, J. J., Sandwell, D. T., Smith, W. H. F., Braud, J., Binder, B., Depner, J., Fabre, D., Factor, J., Ingalls, S., Kim, S. H., Ladner, R., Marks, K., Nelson, S., Pharaoh, A., Trimmer, R., Von Rosenberg, J., Wallace, G., and Weatherall, P.: Global Bathymetry and Elevation Data at 30 Arc Seconds Resolution: SRTM30_PLUS, Mar. Geod., 32, 355–371, 2009.

Berline, L., Ody, A., Jouanno, J., Chevalier, C., André, J.-M., Thibaut, T., and Ménard, F.: Hindcasting the 2017 dispersal of

555 *Sargassum* algae in the Tropical North Atlantic, Marine Pollution Bulletin, 158, https://doi.org/10.1016/j.marpolbul.2020.111431, 2020.

Bernard, D., Biabiany, E., Sekkat, N., Chery, R., and Cécé, R.: Massive stranding of pelagic *Sargassum* seaweeds on the french Antilles coasts: Analysis of observed situations with Operational Mercator global ocean analysis and forecast system. 24e Congrès Français de Mécanique, Brest, https://cfm2019.sciencesconf.org/258628/document, 2019.

560 Bertola, L. D., Boehm, J. T., Putman, N. F., Xue, A. T., Robinson, J. D., Harris, S., Baldwin, C. C., Overcast, I., and Hickerson, M. J.: Asymmetrical gene flow in five co-distributed syngnathids explained by ocean currents and rafting propensity, Proc. R. Soc. B., 287, https://doi.org/10.1098/rspb.2020.0657, 2020.

Biabiany, E., Bernard, D., Page, V., and Paugam-Moisy, H.: Design of an expert distance metric for climate clustering: The case of rainfall in the Lesser Antilles, Computers & Geosciences, 145, https://doi.org/10.1016/j.cageo.2020.104612, 2020.

565 Bo, X., Chunfen, Z., Xinning, D., and Jiayue, W.: The Application of a Decision Tree and Stochastic Forest Model in Summer Precipitation Prediction in Chongqing. Atmosphere, 11(5), 508, https://doi.org/10.3390/atmos11050508, 2020.

Boé, J. and Terray, L.: Weather regimes and downscaling, La Houille Blanche, 94, 45-51, https://doi.org/10.1051/lhb:2008016, 2008.

Brooks, M. T., Coles, V. J., Hood, R. R., and Gower, J. F.: Factors controlling the seasonal distribution of pelagic *Sargassum*, Mar.

570 Ecol. Prog. Ser., 599, 1-18, https://doi.org/10.3354/meps12646, 2018.

17

Cassou, C., Terray L., Hurrell, J. W., and Deser, C.: North Atlantic winter climate regimes: spatial asymmetry, stationarity with time and oceanic forcing. J. Climate. 17, 1055-1068, https://doi.org/10.1175/1520-0442(2004)017<1055:NAWCRS>2.0.CO;2, 2004.

Copernicus PSY4V3R1 Mercator 1/12 degree 3D: https://datastore.cls.fr/catalogues/mercator-model-psy4v3-velocity-112, last access: 15 June 2021.

Cuevas, E., Uribe-Martínez, A., and Liceaga-Correa, M.: A satellite remote-sensing multi-index approach to discriminate pelagic *Sargassum* in the waters of the Yucatan Peninsula, Mexico, Int. J. Remote Sens., 39, 3608-3627, https://doi.org/10.1080/01431161.2018.1447162, 2018.

Cummings, J. A.: Operational multivariate ocean data assimilation. Quart. J. Royal Met. Soc., Part C, 131(613), 3583-3604, https://doi.org/10.1256/qj.05.105, 2005.

Cummings, J. A., Smedstad O. M.: Variational Data Assimilation for the Global Ocean. In: Park S., Xu L. (eds) Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II), Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-35088-7_13, 2013.

Elzinga, C. H. and Studer, M.: Spell Sequences, State Proximities, and Distance Metrics, Sociological Methods & Research, 44, 3-47, https://doi.org/10.1177/0049124114540707, 2015.

Fereday, D. R., Knight, J. R., Scaife, A. A., Folland, C. K., and Philipp, A.: Cluster Analysis of North Atlantic–European Circulation Types and Links with Tropical Pacific Sea Surface Temperatures, Journal of Climate, 21(15), 3687-3703, https://doi.org/10.1175/2007JCLI1875.1, 2008.

Fossette, S., Putman, N. F., Lohmann, K. J., Marsh, R., and Hays, G. C.: A biologist's guide to assessing ocean currents: a review, Mar. Ecol. Prog. Ser., 457, 285-301, http://www.jstor.org/stable/24876354, 2012.

Florenne, T., Guerber, F., and Colas-Belcour, F.: Le phénomène d'échouages des sargasses dans les Antilles et en Guyane, Ministry of Overseas, Ministry of the Environment, Energy and the Sea, Ministry of Agriculture, Agri-Food and Forestry, Paris, France, 406 pp, https://agriculture.gouv.fr/sites/minagri/files/cgaaer_15113_2016_rapport.pdf, (last access: 10 April 2021), 2016.

Franks, J. S., Johnson, D. R., Ko, D. S., Sanchez-Rubio, G., Hendon, J. R., and Lay, M.: Unprecedented Influx of Pelagic *Sargassum* along Caribbean Island Coastlines during Summer 2011. In: Proceedings of the Gulf and Caribbean Fisheries Institute, 64, 6-8, http://aquaticcommons.org/21307/, 2012.

Franks, J. S., Johnson D. R., and Ko D. S.: Pelagic *Sargassum* in the Tropical North Atlantic. Gulf and Caribbean Research, 27, https://doi.org/10.18785/gcr.2701.08, 2016.

Fratantoni, D. M., and Glickson, D. A.: North Brazil Current Ring Generation and Evolution Observed with SeaWiFS, Journal of Physical Oceanography, 32(3), 1058-1074. https://doi.org/10.1175/1520-0485(2002)032<1058:NBCRGA>2.0.CO;2, 2002.

Fratantoni, D. M., and Richardson, P. L.: The Evolution and Demise of North Brazil Current Rings, Journal of Physical Oceanography, 36, 1241-1264, https://doi.org/10.1175/JPO2907.1, 2006.

Friedl, M. A., and Brodley, C. E.: Decision tree classification of land cover from remotely sensed data, Remote Sensing of Environment, vol. 61, Issue 3, pp. 399-409, ISSN:0034-4257, https://doi.org/10.1016/S0034-4257(97)00049-7, 1997.

18

605 Fučkar, N. S., Guemas, V., Johnson, N. C., Massonet, F., and Doblas-Reyes, F. J.: Clusters of interannual sea ice variability in the northern hemisphere. Clim Dyn, 47, 1527-1543, https://doi.org/10.1007/s00382-015-2917-2, 2016.

Gabadinho, A., Ritschard, G., Müller, N., Studer, M.: Analyzing and Visualizing State Sequences in R with TraMineR, Journal of Statistical Software, 40(4), 1–37, https://doi.org/10.18637/jss.v040.i04, 2011.

Gasparin, F., Guinehut, S., Mao, C., Mirouze, I., Rémy, E., King, R. R., Hamon, M., Reid, R., Storto, A., Le Traon, P-Y., Martin,
610 M. J., and Masina, S.: Requirements for an Integrated in situ Atlantic Ocean Observing System From Coordinated Observing System Simulation Experiments, Frontiers in Marine Science, 83(6), https://doi.org/10.3389/fmars.2019.00083, 2019.

Geisser, S.: Predictive Inference: An Introduction, Chapman & Hall, ISBN: 978-0-412-03471-8, 1993.

Gower, J., Hu, C., Borstad, G., and King, S.: Ocean Color Satellites Show Extensive Lines of Floating *Sargassum* in the Gulf of Mexico, in IEEE Transactions on Geoscience and Remote Sensing, 44, 3619-3625, https://doi.org/10.1109/TGRS.2006.882258,
615 2006.

Gower, J., and King, S.: Distribution of floating *Sargassum* in the Gulf of Mexico and the Atlantic Ocean mapped using MERIS, International Journal of Remote Sensing, 32, 1917-1929, https://doi.org/10.1080/01431161003639660, 2011.

Gower, J., Young, E. and King, S.: Satellite images suggest a new *Sargassum* source region in 2011, Remote Sensing Letters, 4:8,
764-773, https://doi.org/10.1080/2150704X.2013.796433, 2013.

620 Haza, A. C., Paldor, N., Ozgokmen, T. M., Curcic, M., Chen, S. S., and Jacobs, G. A.: Wind-based estimations of ocean surface currents from massive clusters of drifters in the Gulf of Mexico. Journal of Geophysical Research: Oceans, 124, 5844– 5869. https://doi.org/10.1029/2018JC014813., 2019.

Helber, R. W., Townsend, T. L., Barron, C. N., Dastugue, J. M., Carnes, M. R.: Validation Test Report for the Improved Synthetic Ocean Profile (ISOP) System, Part I: Synthetic Profile Methods and Algorithm, Report, Naval Research Laboratory, Mississippi,
625 127 pp, https://www7320.nrlssc.navy.mil/pubs/2013/helber1-2013.pdf, 2013.

Hu, C.: A novel ocean color index to detect floating algae in the global oceans, Remote Sens. Environ., 113, 2118–2129, https://doi.org/10.1016/j.rse.2009.05.012, 2009.

Hu, C., Feng, L., Hardy, R. F, and Hochberg, E. J.: Spectral and spatial requirements of remote measurements of pelagic *Sargassum* macroalgae, Remote Sens. Environ., 167, 229-246, https://doi.org/10.1016/j.rse.2015.05.022, 2015.

630 Hu, C., Murch, B., Barnes, B., Wang, M., Maréchal, J., Franks, J., Johnson, D., Lapointe, B., Goodwin, D., Schell J., and Siuda, A.: *Sargassum* Watch Warns of Incoming Seaweed, Eos, 97, https://doi.org/10.1029/2016EO058355, 2016.

Huang, M., Liang, X., Zhu, Y., Liu, Y., and Weisberg, R. H.: Eddies connect the tropical Atlantic Ocean and the Gulf of Mexico. Geophysical Research Letters, 48, e2020GL091277. https://doi.org/10.1029/2020GL091277, 2021.

HYCOM GLBy0.08 version, https://www.hycom.org/dataserver/gofs-3pt1/analysis, last access: 17 January 2022.

635 Johns, W. E., Townsend, T. L., Fratantoni, D. M., Douglas Wilson, D. W.: On the Atlantic inflow to the Caribbean Sea, Deep Sea Res. Part I Oceanogr. Res. Pap., 49, 211-243, https://doi.org/10.1016/S0967-0637(01)00041-3, 2002.

Johns, E. M., Lumpkin, R., Putman, N. F., Smith, R. H., Muller-Karger, F. E., Rueda, D., Hu, C., Wang, M., Brooks, M. T., Gramer, L. J., Werner F. E.: The Establishment of a Pelagic *Sargassum* Population in the Tropical Atlantic: Biological Consequences of a

19

Basin-Scale Long Distance Dispersal Event, Progress in Oceanography (2020), 182, 25 pp, https://doi.org/10.1016/j.pocean.2020.102269, 2020.

Johnson, D., Ko, D. S., Franks, J. S., Moreno, P., Sanchez-Rubio, G.: The *Sargassum* invasion of the eastern Caribbean and dynamics of the equatorial North Atlantic, Gulf and Caribbean Fisheries Institute Proceedings, 65, 102-103, 2014.

Jouanno, J., Benshila, R., Berline, L., Soulié, A., Radenac, M.-H., Morvan, G., Diaz, F., Sheinbaum, J., Chevalier, C., Thibaut, T., Changeux, T., Menard, F., Berthet, S., Aumont, O., Ethé, C., Nabat, P., and Mallet, M.: A NEMO-based model of *Sargassum* distribution in the tropical Atlantic: description of the model and sensitivity analysis (NEMO-Sarg1.0), Geosci. Model Dev., 14, 4069–4086, https://doi.org/10.5194/gmd-14-4069-2021, 2021a.

Jouanno, J., Moquet, J.-S., Berline, L., Radenac, M.-H., Santini, W., Changeux, T., Thibaut, T., Podlejski W., Ménard, F., Martinez, J.-M., Aumont, O., Sheinbaum, J., Filizola N. and Moukandi N'Kaya G. D.: Evolution of the riverine nutrient export to the Tropical Atlantic over the last 15 years: is there a link with Sargassum proliferation?, Environ. Res. Lett., 16, 8 pp, https://doi.org/10.1088/1748-9326/abe11a, 2021b.

Jouini, M., Béranger, K., Arsouze, T., Beuvier, J., Thiria, S., Crépon, M., and Taupier-Letage, I.: The Sicily Channel surface circulation revisited using a neural clustering analysis of a high-resolution simulation, J. Geophys. Res. Oceans, 121, 4545– 4567, doi:10.1002/2015JC011472, 2016.

Kuhn, M., and Johnson, K.: Applied predictive modeling. New York, Springer. isbn:978-1-4614-6848-6, doi:10.1007/978-1-4614-6849-3, 2013.

Kullback, S., and Leibler, R.: On information and sufficiency, Annals of Mathematical Statistics, 22, 79-86, https://doi.org/10.1214/aoms/1177729694, 1951.

Larmarange, J., Mossong, J., Bärnighausen T., Newell M. L.: Participation Dynamics in Population-Based Longitudinal HIV Surveillance in Rural South Africa, PLOS ONE, 10(4), 16 pp, https://doi.org/10.1371/journal.pone.0123345, 2015.

Lellouche, J.-M., Greiner, E., Le Galloudec, O., Garric, G., Regnier, C., Drevillon, M., Benkiran, M., Testut, C.-E., Bourdalle-Badie, R., Gasparin, F., Hernandez, O., Levier, B., Drillet, Y., Remy, E., and Le Traon, P.-Y.: Recent updates to the Copernicus Marine Service global ocean monitoring and forecasting real-time 1/12° high-resolution system, Ocean Science, 14, 1093-1126, https://doi.org/10.5194/os-14-1093-2018, 2018.

Lumpkin, R. and Garzoli S. L.: Near-surface circulation in the tropical Atlantic Ocean, Deep Sea Res. Part I Oceanogr. Res. Pap., 52, 495-518, https://doi.org/10.1016/j.dsr.2004.09.001, 2005.

Maréchal, J.-P., Hellio, C., Hu, C.: A simple, fast, and reliable method to predict *Sargassum* washing ashore in the Lesser Antilles, Remote Sens. Appl.: Soc. Environ., 5, 54-63, https://doi.org/10.1016/j.rsase.2017.01.001, 2017.

Michelangeli, P., Vautard, R., Legras, B.: Weather regime occurrence and quasi stationarity., J. Atmos. Sci., 52, 1237-1256, https://doi.org/10.1175/1520-0469(1995)052, 1995.

Miron, P., Olascoaga, M. J., Beron-Vera, F. J., Putman, N. F., Triñanes, J., Lumpkin, R., and Goni, G. J.: Clustering of marine-debris- and *Sargassum*-like drifters explained by inertial particle dynamics, Geophys. Res. Lett., 47, https://doi.org/10.1029/2020GL089874, 2020.

20

Pham, B. T., Phong, T.V., Nguyen-Thoi, T., Parial, K., Singh S. K., Ly, H.-B., Nguyen, K. T., Ho, L. S., Van Le, H., and Prakash, I., Ensemble modeling of landslide susceptibility using random subspace learner and different decision tree classifiers, Geocarto International, https://doi.org/10.1080/10106049.2020.1737972, 2020.

675 Putman, N. F., Goni, G. J., Gramer, L. J., Hu, C., Johns, E. M., Trinanes, J., and Wang, M.: Simulating transport pathways of pelagic *Sargassum* from the equatorial Atlantic into the Caribbean Sea, Prog. Oceanogr., 165, 205-214, https://doi.org/10.1016/j.pocean.2018.06.009, 2018.

Putman, N. F., Lumpkin, R., Olascoaga, M. J., Trinanes, J., and Goni, G. J.: Improving transport predictions of pelagic *Sargassum*, 680 Journal of Experimental Marine Biology and Ecology, 529, https://doi.org/10.1016/j.jembe.2020.151398, 2020.

Swain, P. H., and Hauska, H.: The decision tree classifier: Design and potential, IEEE Transactions on Geoscience Electronics, 15, https://doi.org/10.1109/TGE.1977.6498972, 1977.

Resiere, D., Valentino, R., Nevière R., Banydeen, R., Gueye, P., Florentin, J., Cabié, A., Lebrun, T., Mégarbane B., Guerrier G., and Mehdaoui, H.: *Sargassum* seaweed on Caribbean islands: an international public health concern, The Lancet, 392, 2691, 685 https://doi.org/10.1016/S0140-6736(18)32777-6, 2018.

Rousseeuw, P. J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, 20, 53-65, https://doi.org/10.1016/0377-0427(87)90125-7, 1987.

Sargassum Watch System (SAWS), University of South Florida, https ://optics.marine.usf.edu/projects/SaWS.html, last access: 15 June 2021.

690 Schell, J. M., Goodwin, D. S., and Siuda, A. N. S.: Recent *Sargassum* Inundation Events in the Caribbean: Shipboard Observations Reveal Dominance of a Previously Rare Form, Oceanography, 28, 8–11, https://www.jstor.org/stable/24861895, 2015.

Széchy, M. T. M., Guedes, P. M., Baeta-Neves, M. H., and Oliveira, E. N.: Verification of Sargassum natans (Linnaeus) Gaillon (Heterokontophyta: Phaeophyceae) from the Sargasso Sea off the coast of Brazil, western Atlantic Ocean, Check List, 8, 638-641, http://dx.doi.org/10.15560/8.4.638, 2012.

695 Studer, M. and Ritschard, G.: What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures, Journal of the Royal Statistical Society, Series A, 179(2), 481–511, https://doi.org/10.1111/rssa.12125, 2016.

Trinanes J., Putman, N. F., Goni, G., Hu, C., Wang, M.: Monitoring pelagic Sargassum inundation potential for coastal communities, Journal of Operational Oceanography, 14, https://doi.org/10.1080/1755876X.2021.1902682, 2021.

Van Tussenbroek, B. I., Hernández Arana, H. A., Rodríguez-Martínez, R. E., Espinoza-Avalos, J., Canizales-Flores, H. M., 700 González-Godoy, C. E., Barba-Santos, M. G., Vega-Zepeda, A., and Collado-Vides, L.: Severe impacts of brown tides caused by *Sargassum* spp. on near-shore Caribbean seagrass communities, Mar. Pollut. Bull., 122, 272-281. https://doi: 10.1016/j.marpolbul.2017.06.057, 2017.

Wang, M., and Hu, C.: Mapping and quantifying *Sargassum* distribution and coverage in the central West Atlantic using MODIS observations, Remote Sens. Environ., 183, 350-367, https://doi.org/10.1016/j.rse.2016.04.019, 2016.

705 Wang, M., and Hu, C.: Predicting *Sargassum* blooms in the Caribbean Sea from MODIS observations, Geophys. Res. Lett., 44, 3265– 3273, https://doi.org/10.1002/2017GL072932, 2017.

21

Wang, M., Hu, C., Cannizzaro, J., English, D., Han, X., Naar, D., B., Lapointe, R., Brewton, F., Hernandez.: Remote sensing of *Sargassum* biomass, nutrients, and pigments. Geophys. Res. Lett., 45, 12,359– 12,367. https://doi.org/10.1029/2018GL078858, 2018.

Wang, M., Hu, C., Barnes, B. B., Mitchum, G., Lapointe, B., Montoya, J. P.: The great Atlantic *Sargassum* belt, Science, 365, 83-87, https://doi.org/10.1126/science.aaw7912, 2019.

Ward, J. H. Jr.: Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association, 58:301, 236-244, DOI: 10.1080/01621459.1963.10500845, 1963.

Webster, R. K., and Linton T.: Development and implementation of *Sargassum* Early Advisory System (SEAS), Shore & Beach, 81, 1–6, http://www.sargassoseacommission.org/storage/Webster_et_linon_2013_1.pdf, 2013.

Witherington, B., Hirama S., and Hardy, R.: Youngsea turtles of the pelagic *Sargassum*-dominated drift community: habitat use, population density, and threats, Mar. Ecol. Prog. Ser., 463, 1-22, https://doi.org/10.3354/meps09970, 2012.

Zhao, Q., Basher, Z., and Costello, M. J.: Mapping near surface global marine ecosystems through cluster analysis of environmental data. Ecological Research.; 35: 327– 342. https://doi.org/10.1111/1440-1703.12060, 2020.
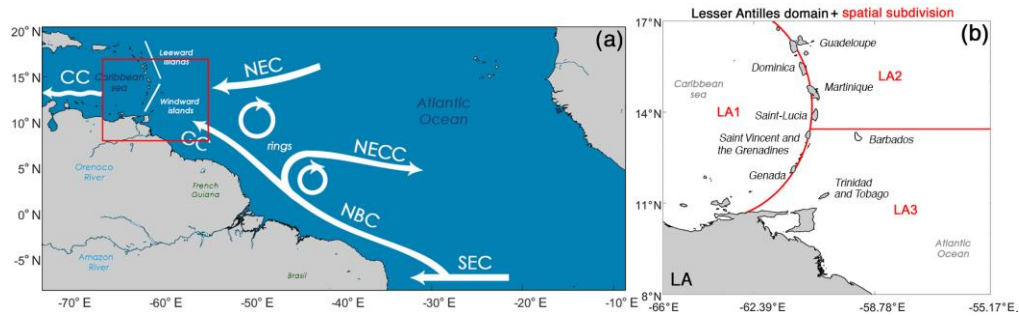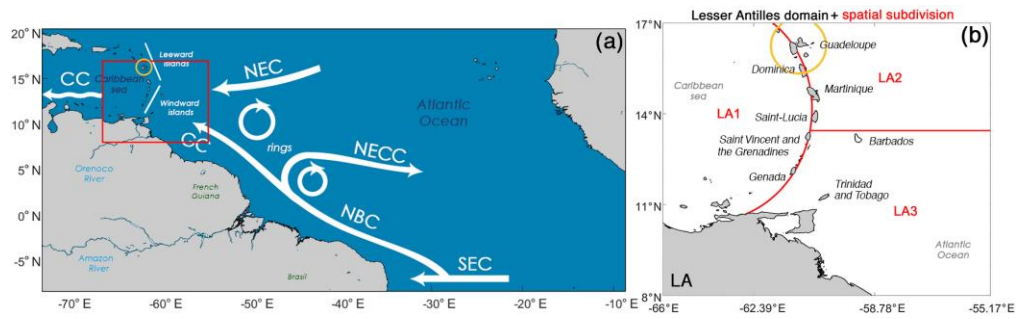
Figure 1: (a) Main oceanic currents occurring and interacting in the central Atlantic and the Lesser Antilles regions; Caribbean Current (CC), North Equatorial current (NEC), North Brazil current (NBC), North equatorial Counter Current (NECC), South Equatorial current (SEC). Lesser Antilles domain (LA): the red rectangle corresponds to the study area (55-66° W, 8-17° N); (b) Spatial subdivision of the study area into three sub-areas: LA1 (i.e., Caribbean Sea), LA2 (i.e., North Tropical Atlantic above Barbados (13.2° N)) and LA3 (i.e., North Tropical Atlantic below 13.2° N). The yellow circle corresponds to the 100 km offshore Guadeloupe area in which the satellite-based Sargassum abundance is analysed.

A: annual frequency (%), B: pattern compare with KMS-ED bary-centers, C: sargassum stranding frequency per class (%), D: matching with base history on dt days (%).
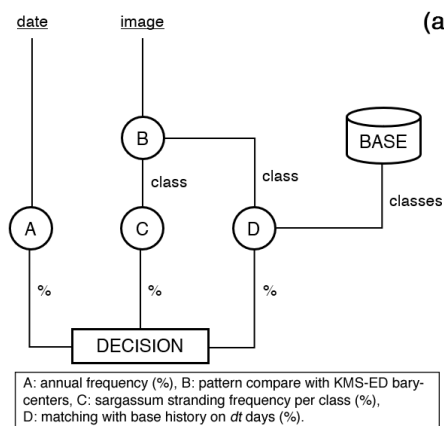
$Cl(i, d)$ operator gives the **cluster label** of oceanic current (determined by KMS-ED clustering) for the index i of the base of sargassum stranding days and the date d corresponding to it.
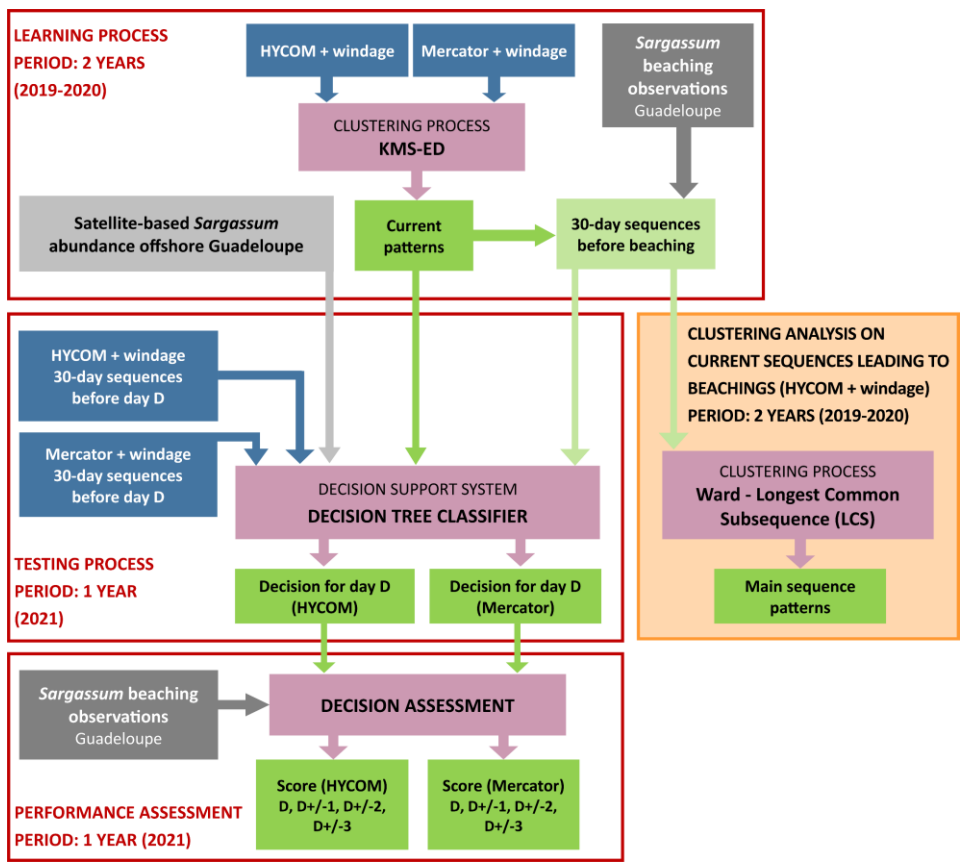
Figure 2: The schematic of the overall methodology.
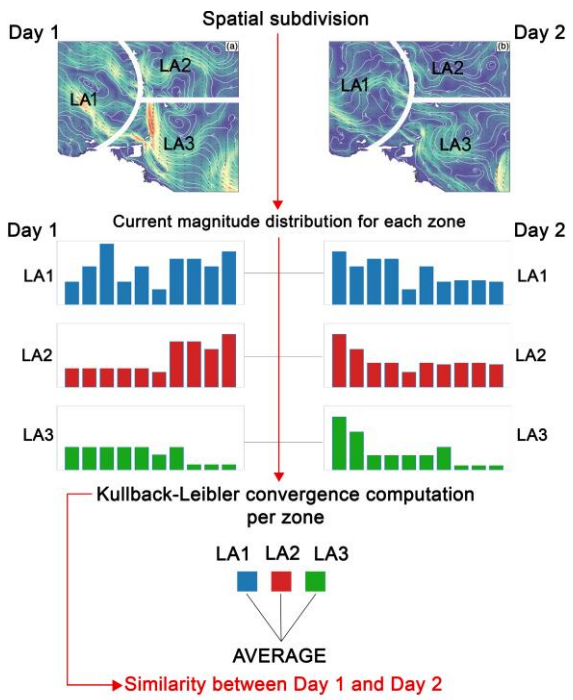
Figure 3: The schematic of the Expert Distance process.

Daily HYCOM clusters
Surface current patterns

Base
Set of the referenced 30-day current clusters before Sargassum beaching in Guadeloupe

Clustering process

Clustering method:
Agglomerative Hierarchical Clustering (HAC)

Linkage method:
Ward algorithm

Metric:
Longest Common Subsequence (LCS)

Clusters of sequences
Main patterns of sequences leading to Sargassum beaching
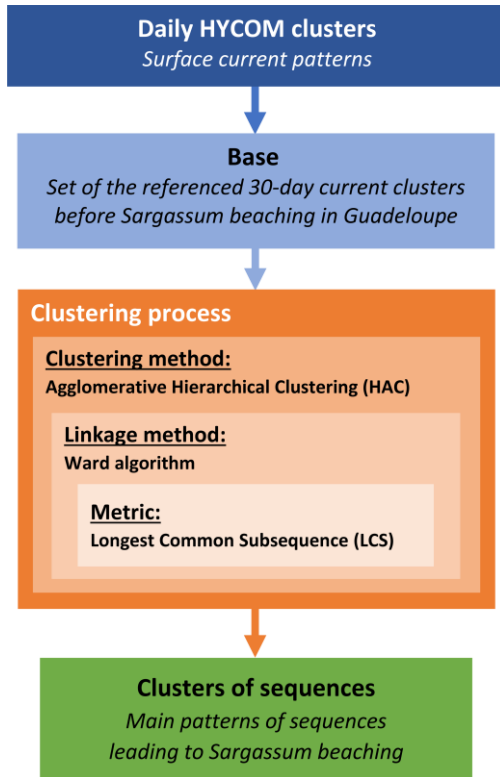
755

Figure 4: The schematic of the clustering process on the current sequences leading to beachings.

(a)

date image

B

Base

class   class   sequences

A   C   D

%   %   %

DECISION
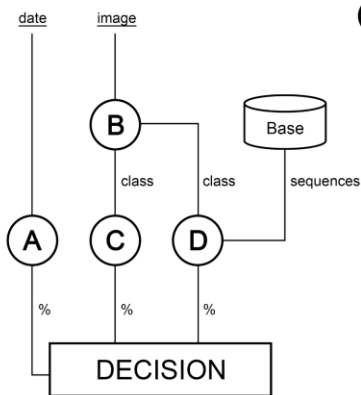
A: daily probability to reach the maximum *Sargassum* abundance in the area 30-100 km offshore Guadeloupe, B: current pattern compare with KMS-ED barycenters, C: *Sargassum* beaching frequency par class (%), D: matching of current sequences with base history on Δt days (%), Base: set of the referenced 30-day current sequences before *Sargassum* beaching.

(b)

Base index	Beaching day	$d-1$	$d-2$	...	$d-\Delta t$
1	$Cl(1,d)$	.	.	..	$Cl(1, d-\Delta t)$
2	.	.	.	..	.
. . .	. . .	. . .	. . .	... ...	... ...
N	$Cl(N,d)$	.	.	...	$Cl(N, d-\Delta t)$

$Cl(i,d)$ operator gives the cluster label of oceanic surface current (computed by KMS-ED) for the index i of the base of sargassum beaching days and the corresponding date d.
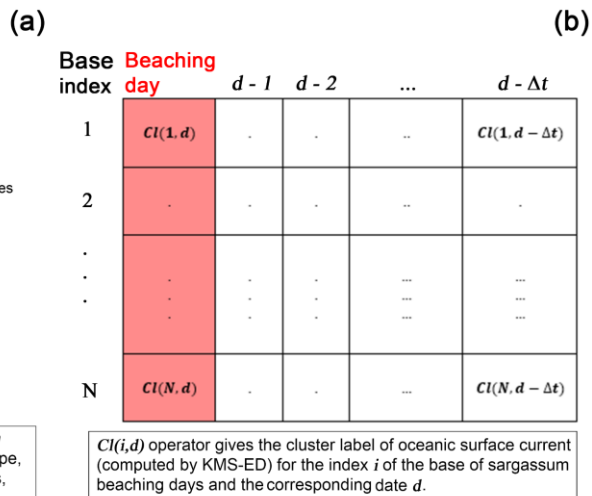
Figure 5: (a) Scheme of the decision tree classifier to predict *Sargassum* strandingbeaching probability. (b) Combination base of oceanic currents clusters labels obtained by KMS-ED from each strandingbeaching day to Δt days before.

a mis en forme : Police :Italique
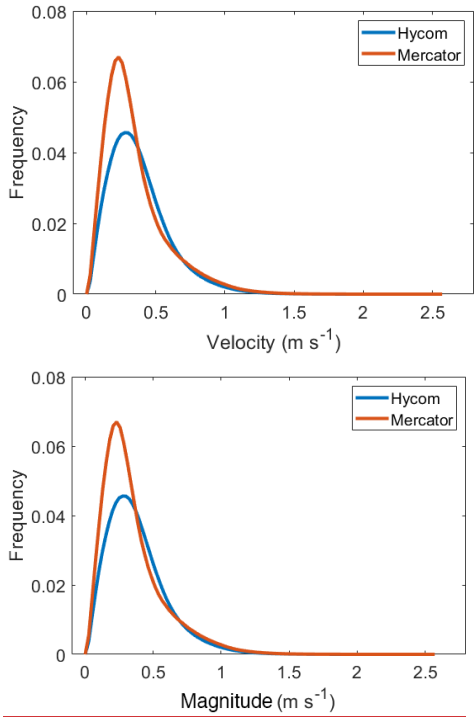
Figure 36: Distributions of oceanic surface ~~currents~~current magnitudes including windage for both models, HYCOM (blue) and Mercator (red) datasets.

775

29
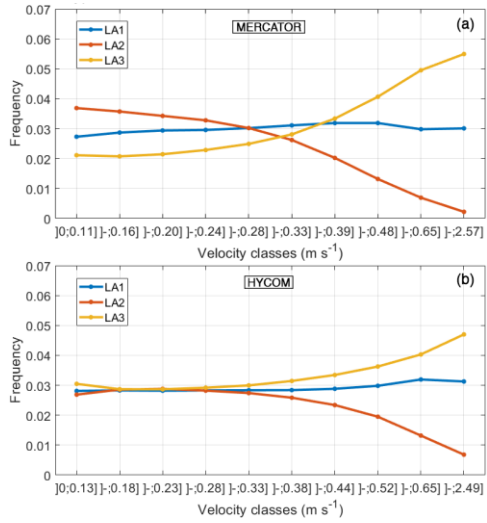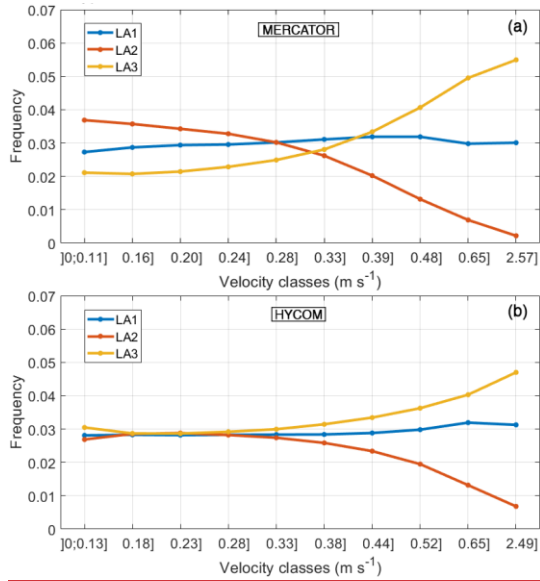
Figure 47: Relative frequency distribution of current speeds for the three offshore sub-regions around the Lesser Antilles (2019-2020), LA1 (blue), LA2 (red), LA3 (yellow). (a) Mercator with ERA-5 windage and (b) HYCOM with ERA-5 windage.

780
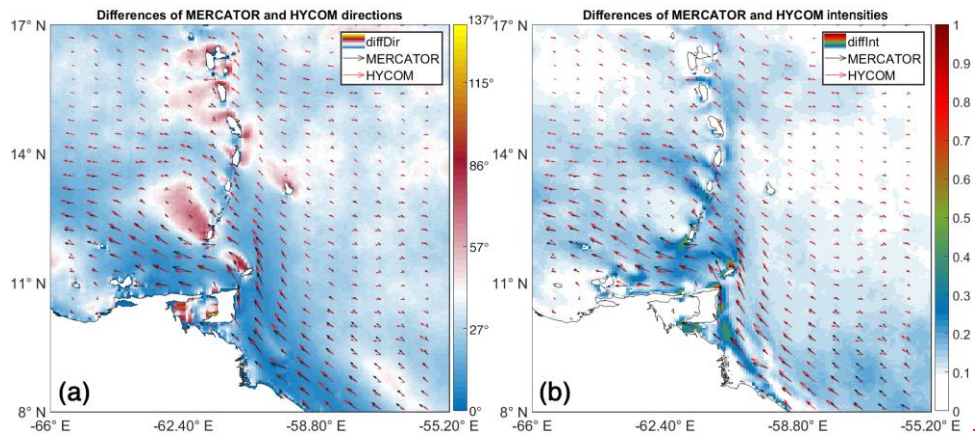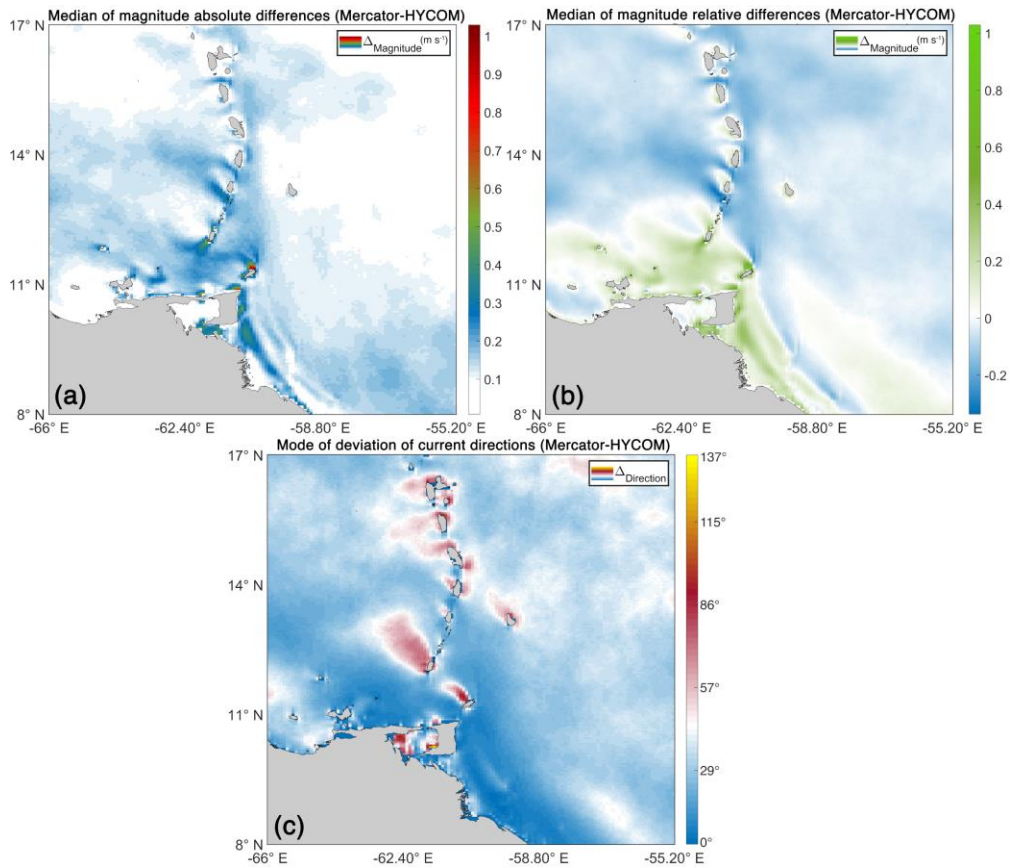
Figure 58: Comparison between Mercator and HYCOM surface currents from 2019 to 2020~~: current direction~~ on the same 0.08° grid: (a) median of magnitude absolute differences ~~in degree (a), current intensity~~(Mercator-HYCOM) in m s⁻¹ and (b) median of magnitude relative differences (Mercator-HYCOM) in m s⁻¹ ~~(b).~~and (c) mode of current direction differences (Mercator-HYCOM) in degree.

785
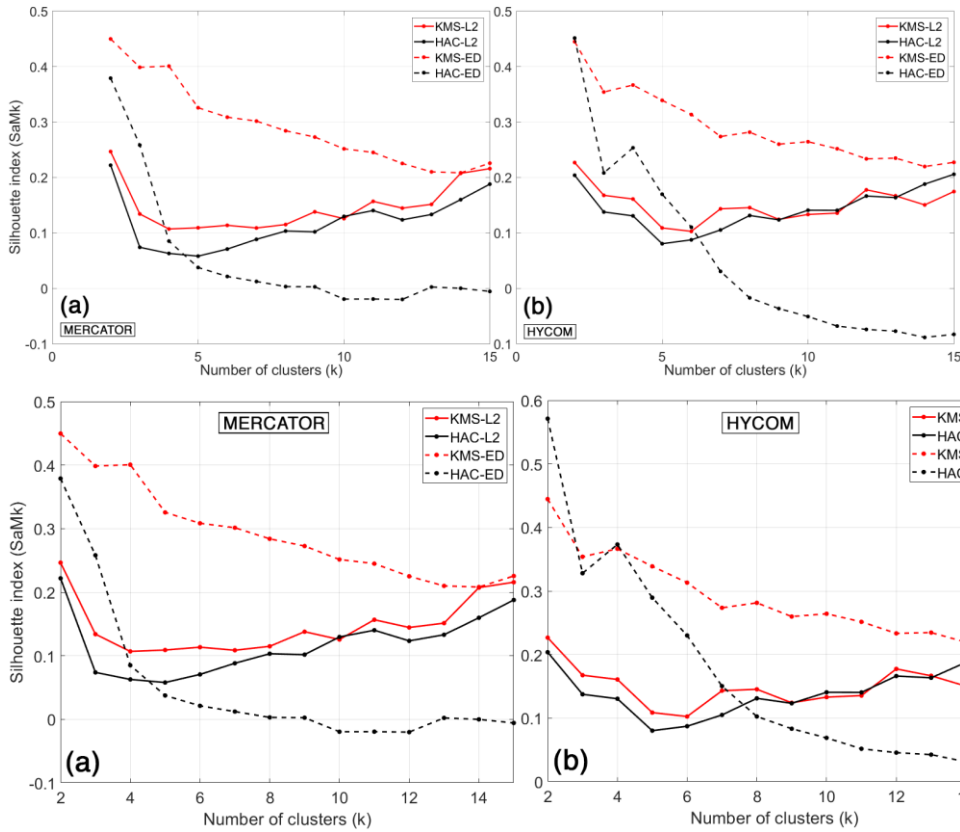
Figure 69: Evolution of the SaMk silhouette index (by method) as a function of the number of clusters k, Mercator (a) and HYCOM (b): HAC method (black), KMS method (red), with L2 metric (solid line) and ED metric (dashed line).
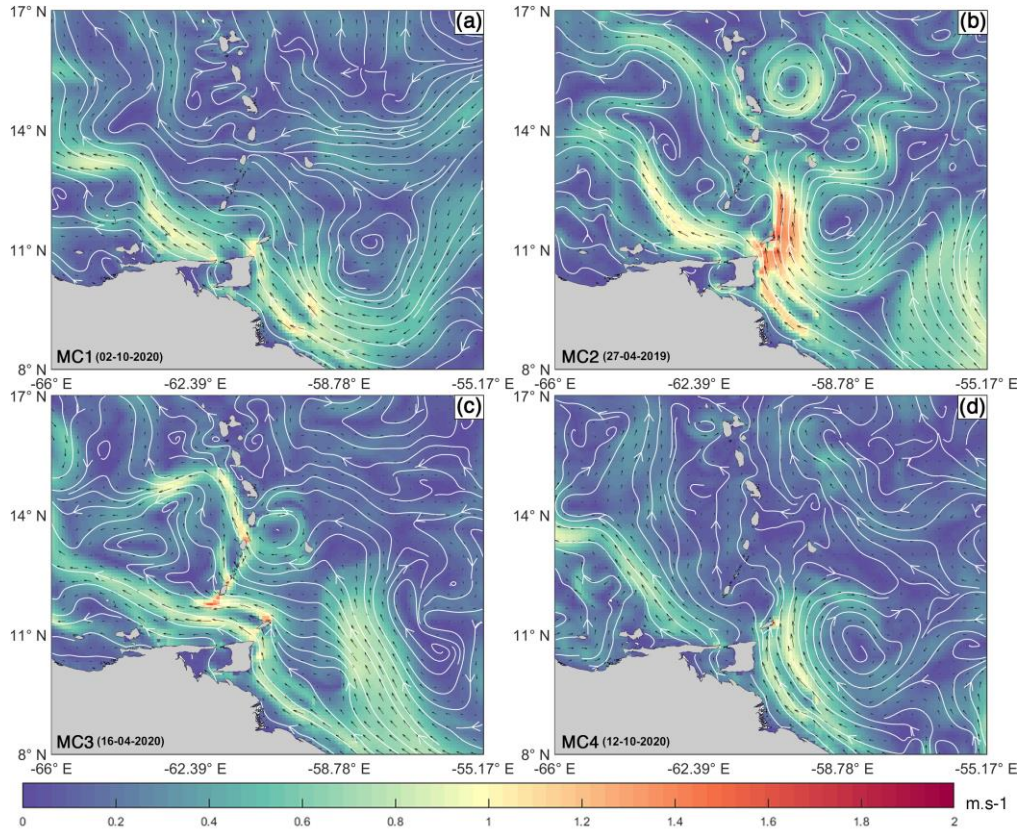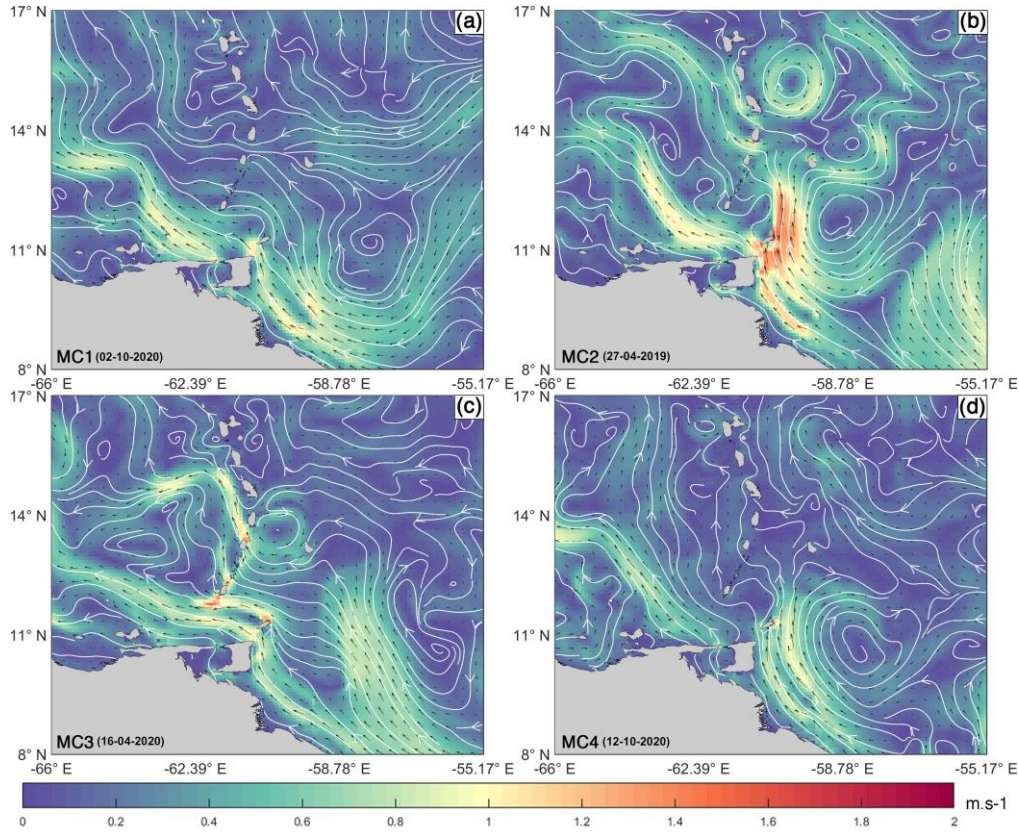
Figure 710: Representative elements of the clusters from Mercator current data combined with ERA-5 windage (KMS-ED method with k = 4): MC1 (day 02-10-2020) (a), MC2 (day 27-04-2019) (b), MC3 (day 16-04-2020) (c), MC4 (day 12-10-2020) (d).

35
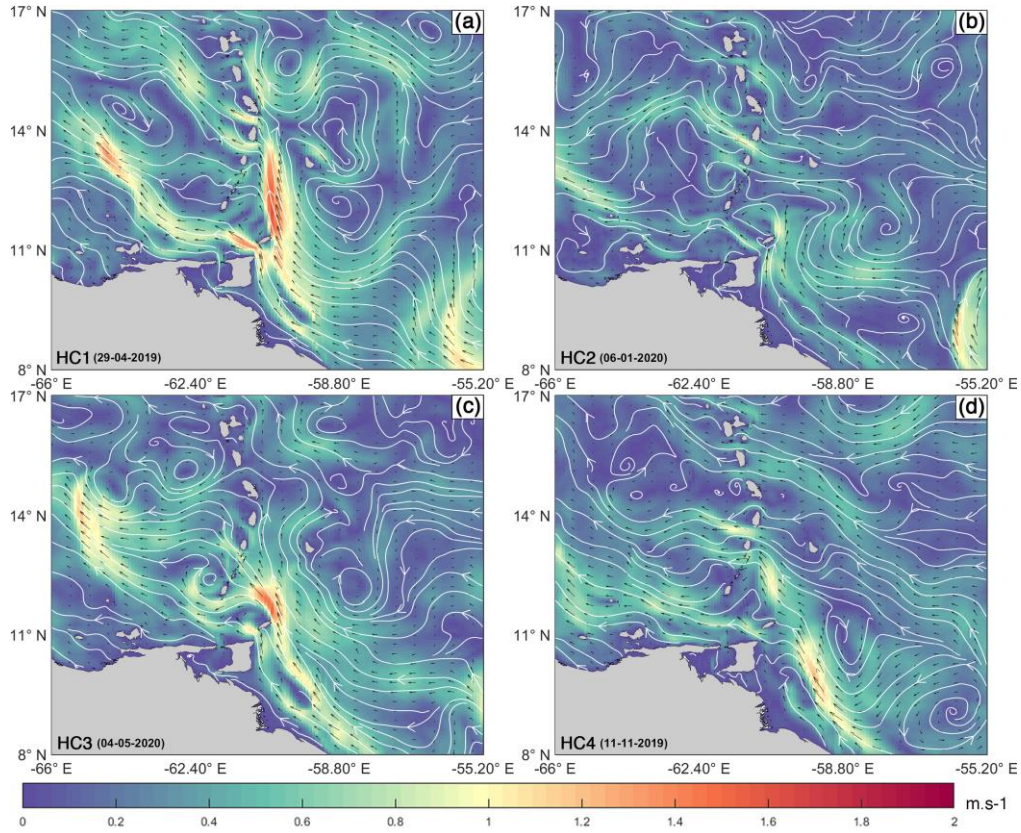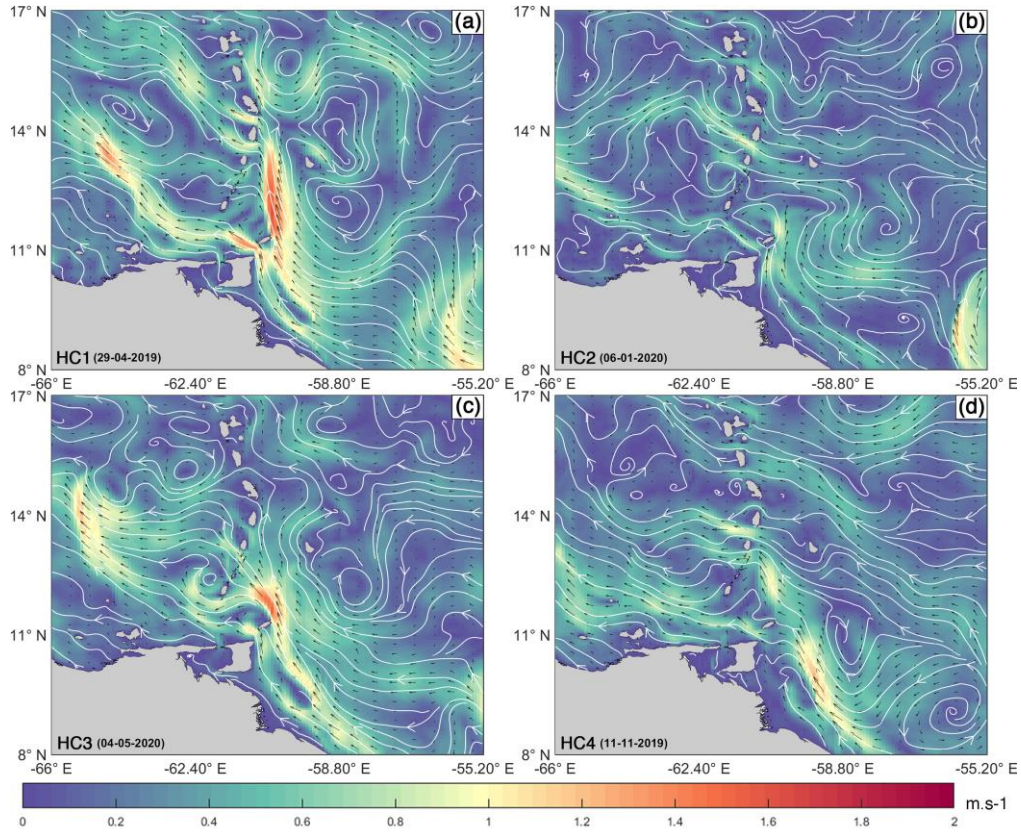
Figure ~~8~~11: Representative elements of the clusters from HYCOM current data combined with ERA-5 windage (KMS-ED method with k = 4): HC1 (day 29-04-2019) (a), HC2 (day 06-01-2020) (b), HC3 (day 04-05-2020) (c), HC4 (day 11-11-2019) (d).

37

810

38
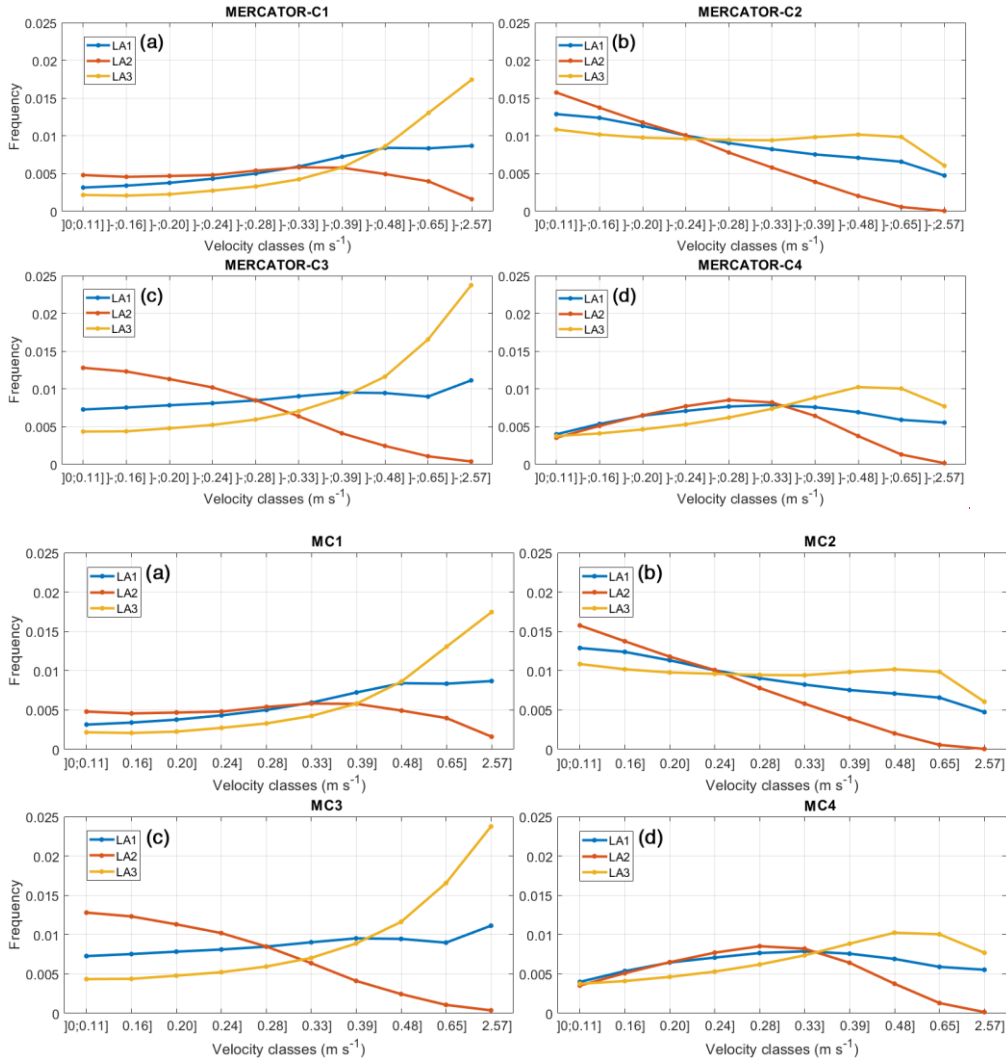
Figure 912: Relative frequency distribution of current speeds for the three offshore sub-regions: MC1 (a), MC2 (b), MC3 (c) and MC4 (d). The representative elements were obtained after KMS-ED clustering for Mercator.

815
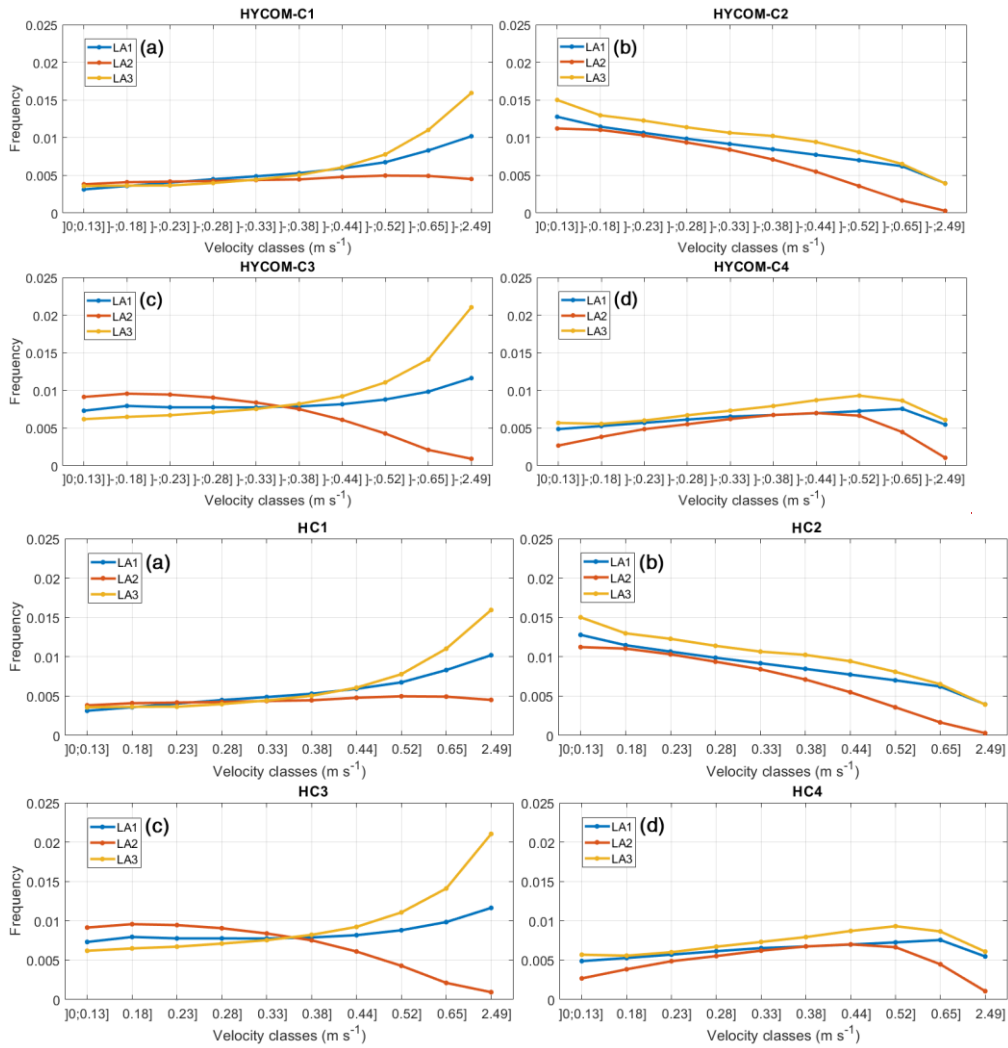
Figure 1013: Relative frequency distribution of current speeds for the three offshore sub-regions: ~~HYCOM-C1~~HC1 (a), ~~HYCOM-C2~~HC2 (b), ~~HYCOM-C3~~HC3 (c) and ~~HYCOM-C4~~HC4 (d).~~–~~ The representative elements were obtained after KMS-ED clustering for HYCOM.

820

41
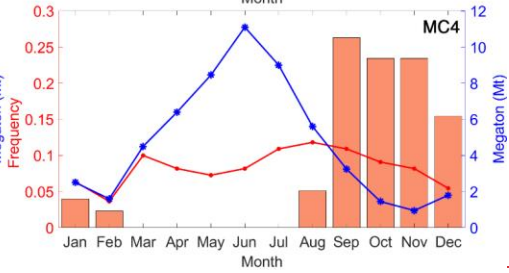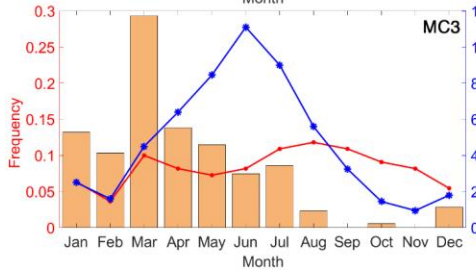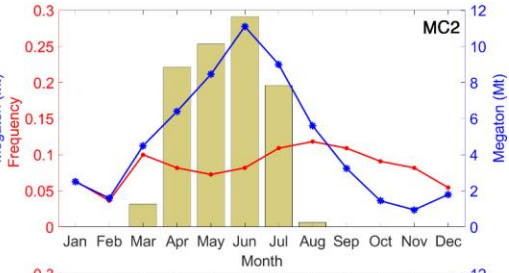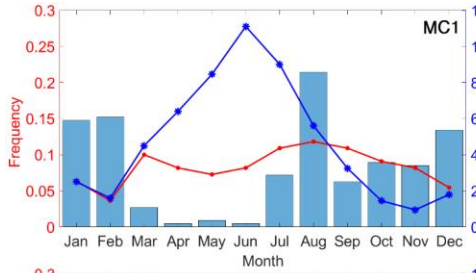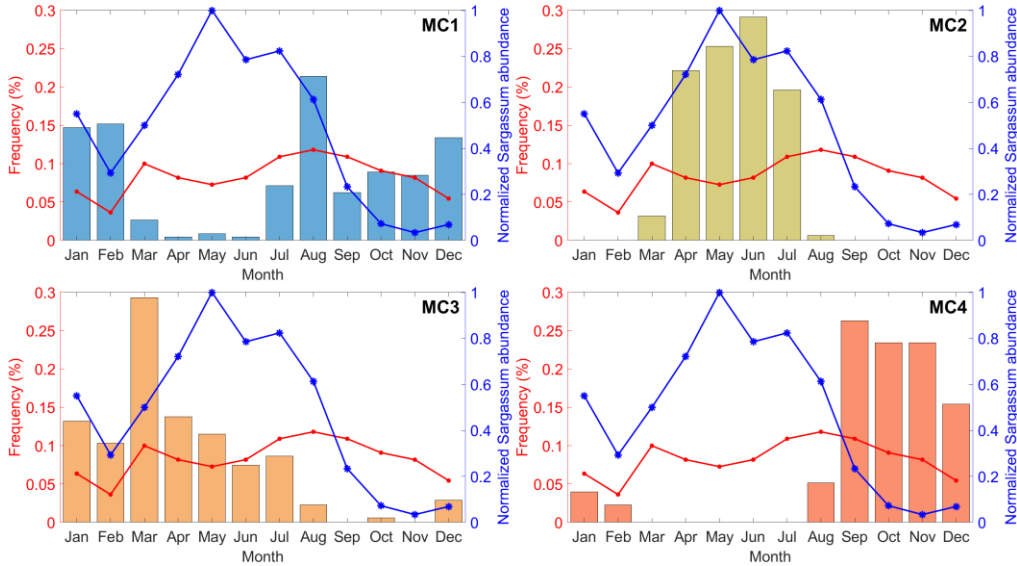
Figure ~~11~~14: Monthly distribution of cluster occurrence from Mercator outputs, from 2019 to 2020, in the Lesser Antilles (55-66°W, 8-17°N): MC1 (a), MC2 (b), MC3 (c) and MC4 (d). The red line shows the monthly distribution of *Sargassum* ~~strandings~~beachings on the coasts of Guadeloupe during the same period. The blue line indicates the monthly ~~distribution~~evolution of *Sargassum* abundanc~~ye~~ in the ~~Atlantic Ocean and~~area 30-100 km offshore Guadeloupe normalized on the ~~Caribbean Sea (in Megaton).~~maximum value.

825

a mis en forme : Police :Italique

a mis en forme : Police :Italique

43
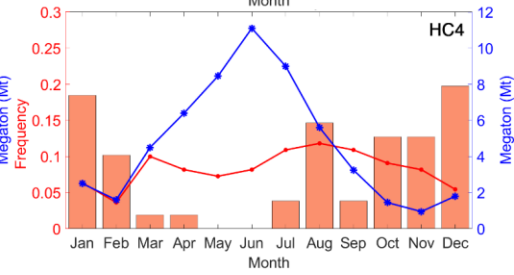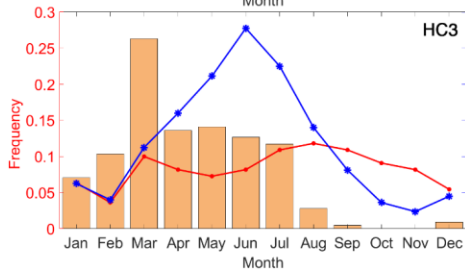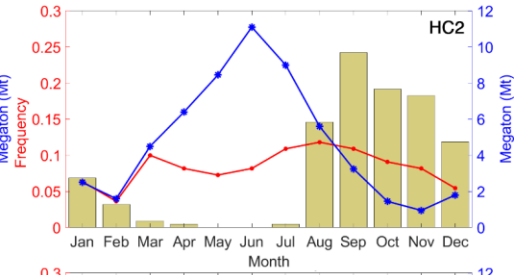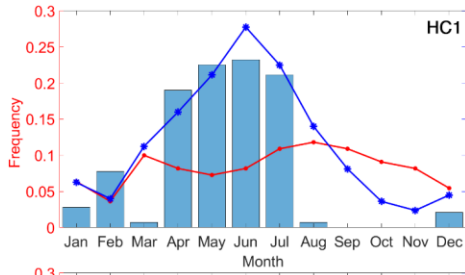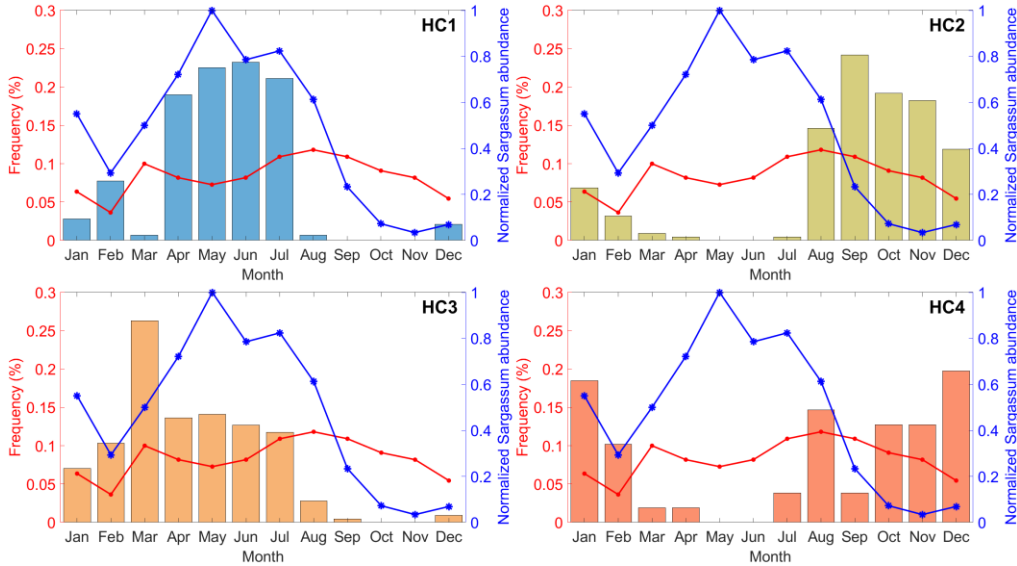
830   **Figure ~~12~~15: Monthly distribution of cluster occurrence from HYCOM outputs, from 2019 to 2020, in the Lesser Antilles (55-66°W, 8-17°N): HC1 (a), HC2 (b), HC3 (c) and HC4 (d). The red line shows the monthly distribution of *Sargassum* ~~strandings~~beachings on the coasts of Guadeloupe during the same period. The blue line indicates the monthly ~~distribution~~evolution of *Sargassum* abundanc~~y~~e in the ~~Atlantic Ocean and~~area 30-100 km offshore Guadeloupe normalized on ~~the Caribbean Sea (in Megaton).~~maximum value.**

835
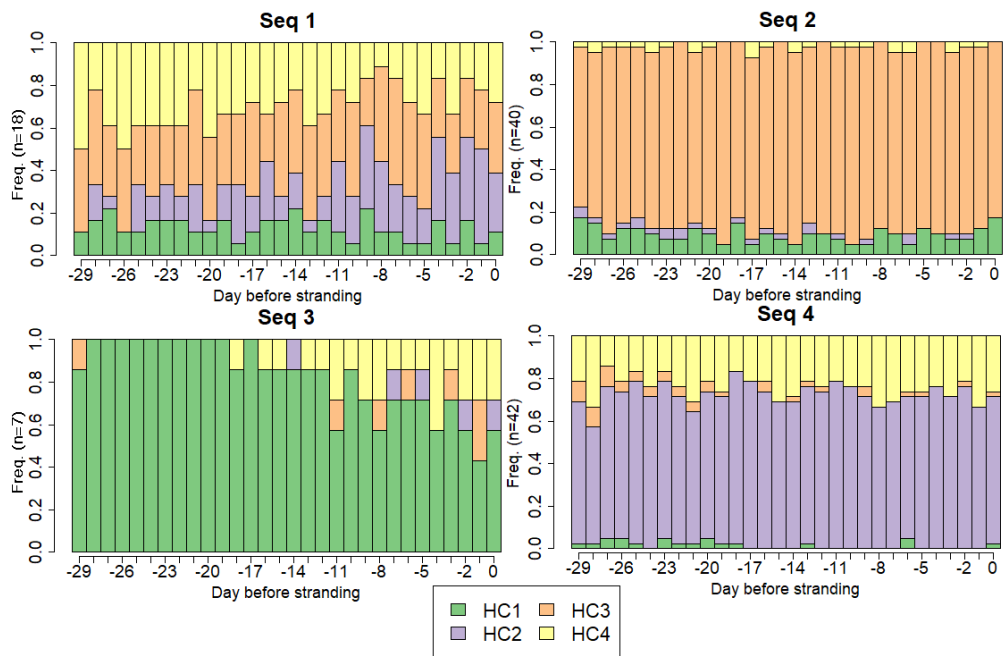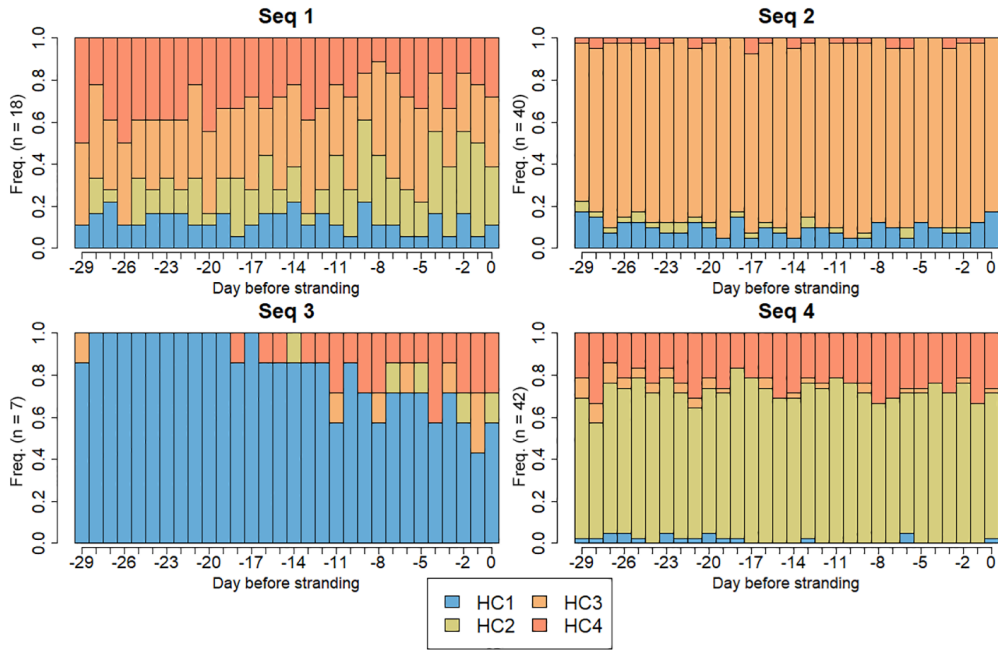
Figure 1316: Distribution of HYCOM current ~~regimes over the 30-day stranding backward sequences:~~regime clusters (i.e. **HC1 (in blue),**
HC2 (in green~~, HC2 in purple,~~**), HC3 (in orange**~~, and~~**), HC4 (in** ~~yellow,~~ **red)) in the 30-day sequence types (i.e. Seq1, Seq2, Seq3, Seq4).**
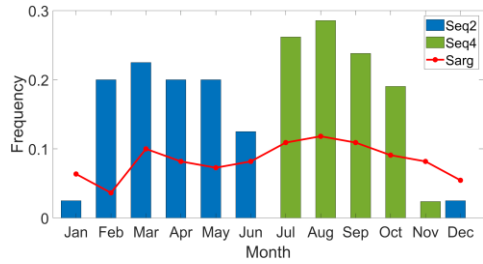
845
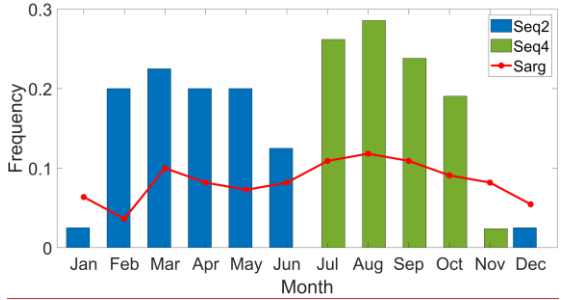
Figure ~~14~~17: Monthly distribution of the main observed ~~stranding backward~~current sequences leading to beaching: Seq2 (blue) and Seq4 (green). The red line represents the distribution of the observed ~~stranding~~beaching days.

850

47
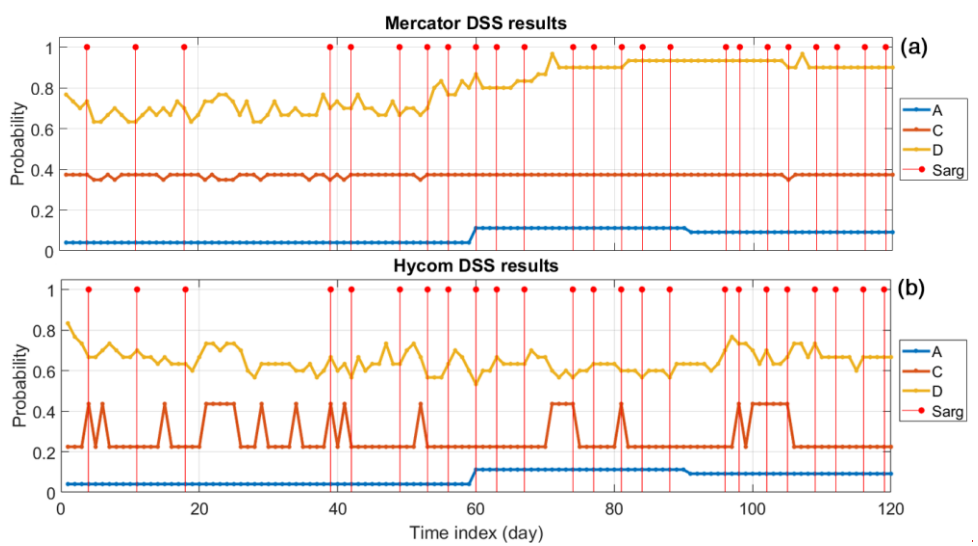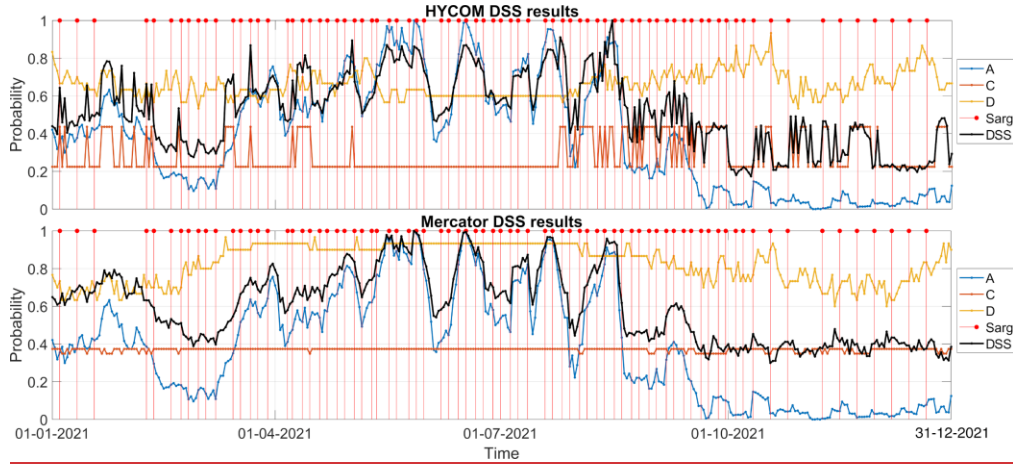
Figure ~~15~~18: Decision Support System (DSS) results; probability of ~~strandings~~beaching obtained per module. ~~Monthly stranding frequency obtained~~ Daily probability to reach the maximum *Sargassum* abundance in the area 30-100 km offshore Guadeloupe for module A (blue line), ~~stranding~~beaching frequency per cluster for module C (~~red~~orange line), match percentage for module D (yellow line~~). Day~~), DSS Decision (black line). Days of observed ~~stranding~~beaching on Guadeloupe coasts (red dots): HYCOM (a) and Mercator (~~a) and HYCOM (~~b).

Deciles (D_i)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Max	~~m~~Mea	Sigma

48

											n	
Mercator (m.s⁻¹)	0.11	0.16	0.20	0.24	0.28	0.32	0.39	0.48	0.65	2.57	0.33	0.22
HYCOM (m.s⁻¹)	0.13	0.18	0.23	0.28	0.32	0.38	0.44	0.52	0.65	2.49	0.36	0.21

Table 1: Boundaries of the histogram classes used to quantify surface currents velocity data with Sigma as Standard deviation.

Datasets	C1	C2	C3	C4
MERCATOR	**224**	158	174	175
	(30.-7%)	(21.6%)	(23.8%)	(23.9%)
HYCOM	142	**219**	213	157
	(19. 4%)	(29.9%)	(29.1%)	(21.5%)

Table 2: Number of days corresponding to each cluster for MERCATOR and HYCOM datasets.

		HYCOM			
		C1	C2	C3	C4
MERCATOR	C1	8,3 %	9,6 %	7,1 %	**50,6 %**
	C2	**60,4 %**	(-)	12,4 %	1,3 %
	C3	0,3 %	4,8 %	**56,7 %**	4,7 %
	C4	(-)	**69,8 %**	0,8 %	3,1 %

Table 3: Correspondence table between the 4 clusters generated with MERCATOR and HYCOM datasets. The percentage expresses the proportion of common days between two clusters ((-) for 0%).

Datasets	C1	C2	C3	C4
MERCATOR	33	15	34	28

HYCOM	14	35	40	21

Table 4: Distribution of observed *Sargassum* ~~stranding~~beaching days (Guadeloupe coasts) in MERCATOR and HYCOM clusters.

~~Backward~~30-day current sequence ~~classes~~before beaching (HYCOM)	Seq1	Seq2	Seq3	Seq4
n	18	40	7	42
%	16.8	37.4	6.5	39.3

Table 5: Distribution of ~~backward~~ sequence ~~classes.~~clusters, with (n) corresponding to the number of sequences in each cluster and (%) corresponding to the ratio of the number of sequences on the total.

Time range around D (day)	Datasets	TPTrue positive (recall %)	FNTrue negative (recall %)	TN	FP	Accuracy (ratio %)
0	**HYCOM**	**46 (59.0%)**	151 (52.8%)			**197 (54.1%)**
	Mercator (SUM)	943 (55.1%)	15141 (49.3%)	5218 4	44 61	(50.6%)
+/- 1	**HYCOM**	52 (66.7%)	175 (61.2%)			227 (62.4%)
	Mercator (%)	37547 (60.3%)	62,5151 (52.8%)	198	45,8 50,8	(54,2 4%)
+/- 2	**HYCOM** (SUM)	1055 (70.5%)	14189 (66.1%)	7824	18 88	(67.0%)
	Mercator	51 (65.4%)	155 (54.2%)			206 (56.6%)
+/- 3	**HYCOM** (%)	41,7 57 (73,3. 1%)	58,3 198 (69.2%)	81,3	18,8	255 (70.1%)
	Mercator	**51 (65.4%)**	161 (56.3%)			**212 (58.2%)**

Table 6: Decision tree ~~accuracy: True Positive (TP), False negative (FN), True Negative (TN) and False Positive (FP).~~ performance scores.

51