https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

Open Access Earth System Science Data Discussions

MAREL Carnot data and metadata from Coriolis Data Center

Raed Halawi Ghosn[1,2], Émilie Poisson-Caillault[2], Guillaume Charria[3], Armel Bonnat[4], Michel Repecaud[5], Jean-Valery Facq[6], Loïc Quéméner[5], Vincent Duquesne[1], Camille Blondel[1], Alain Lefebvre[1]

[1]Ifremer, Unité Littoral, Laboratoire Environnement et Ressources, 150 quai Gambetta, 62321 Boulogne-sur-mer, France
[2]LISIC EA 4491 ULCO/University Littoral, 62228 Calais, France
[3]Ifremer, Univ. Brest, Laboratoire d'Océanographie Physique et Spatiale (LOPS), IUEM, F-29280, Brest, France
[4]Ifremer, Univ. Brest, Service des Systèmes d'Informations Scientifiques pour la MER, F-29280, Brest, France
[5]Ifremer ,REM/RDT/DCM, Centre de Brest, Plouzané, France
[6]Ifremer, Marine Structure Laboratory, 150 QuaiGambetta, 62200 Boulogne-sur-mer, France

Correspondence to: Alain Lefebvre (alain.lefebvre@ifremer.fr)

Abstract. The French coast of the Eastern English Channel (ECC) is classified as potential eutrophication zone by the Paris and Oslo Convention (OSPAR), and as moderate to poor according to phytoplankton quality element of the Water Framework Directive (WFD). Indeed, the French part of the EEC is regularly affected by *Phaeocystis globosa* bloom events, which have detrimental effects on the marine ecosystem, economy as well as public health. Since phytoplankton is an important indicator of water quality, the MAREL Carnot oceanographic multi-sensor station was installed in the Eastern English Channel at the Carnot wall in Boulogne sur Mer in 2004. The aim of this station was to monitor water quality and phytoplankton in order to complement results from existing more conventional low resolution monitoring programs, with high frequency data (sampling every 20 minutes). The purpose of this paper is to introduce the MAREL Carnot dataset and show how it can be used for several research objectives. MAREL Carnot collects high frequency, multi-parameter observations from surface water, as well as meteorological measurements, and send data almost immediately to an inshore data center. In this paper, we present several physiochemical and biological parameters measured by this station. In addition, we demonstrated, based on previous research activities, that the MAREL Carnot dataset is useful for evaluating environmental or ecological statuses, marine phytoplankton ecology, physical oceanography, turbulence, as well as public policy. Most importantly, we showed its contribution to Marine Strategy Framework Directive (MSFD) and other regional or universal conventions.

1 Introduction

For millennia, the marine environment has been subjected to various sources of pollution. Major inputs of nitrate, phosphate, sulphate, metals, and others have been causing detrimental effects on the marine environment, including harmful algal blooms (HAB), and eutrophication (Le Moal et al., 2019). Since phytoplanktons are at the base of food webs, their blooms can affect the entire trophic levels, and can cause serious changes in the marine biodiversity and water quality (e.g. oxygen deficiency). HABs can produce toxins that degrade water quality and may cause health problems in humans and marine animals, in addition to their ability to form high biomass, which leads to foam accumulation with direct and indirect impacts. (Ross Brown et al.,

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

2022). Furthermore, they can detrimentally cause economic losses in sectors such as fish farms, shellfish aquaculture, tourism and recreational activities, as well as public health (Derot et al., 2020).

Understanding the processes underlying these problems necessitates continuous monitoring of marine environments in order to prevent the associated deterioration effects and help managers and stakeholders achieve optimised environmental assessment and management. Traditionally, monitoring aquatic and marine ecosystems was done using low frequency in-situ measurements (weekly to monthly sampling frequency). This was done by collecting water samples through Niskin bottles, and then performing several laboratory analysis to determine various physiochemical and biological parameters, including salinity, temperature, conductivity, organic and inorganic matter, as well as phytoplankton analysis. Despite the fact that these tests helped scientists to have an overview of the processes taking place in the marine environment, they failed to enhance their knowledge and understandings of marine ecosystems, particularly phytoplankton dynamics and eutrophication because of their too low sampling resolution.

In order to be able to set proper management to prevent further deterioration of marine ecosystems, continuous measurements are needed to derive the most relevant information, not just on a monthly or weekly scale, but rather on a daily or hourly scale. In other words, high frequency measurements are needed in order to enhance our understanding of harmful algal blooms, their dynamics, as well as processes such as eutrophication. Although satellite and earth modelling data provide high frequency data, they alone, remain incapable of providing all the needed information required to set better management practices. Indeed, in-situ data is essential to calibrate and validate algorithms used by these two complementary data sources. This urged scientists and stakeholders to study the marine environment using high frequency in-situ monitoring systems, such as ferry boxes, buoys etc.

Over the past decades, the advancement of sensor technology and data science shed light on the importance of time series in marine research. This urged the construction of autonomous systems capable of supporting long-term time series for key physical, chemical, and biological parameters. In other words, the implementation of such automated systems enabled the measurement of essential ocean variables (EOV) and important biodiversity variables (EBV) at high frequency, which aided in reorienting marine research from low frequency measurements to high frequency measurements (Blain et al., 2004).

In the Eastern English Channel (EEC), HABs are mainly caused by the Prymnesiophyceae *Phaeocystis globosa*, which is often associated with *Pseudo-nitzchia* causing severe paralytic shellfish poisoning (Karasiewicz & Lefebvre, 2022). When the temperature of the water rises in the spring and summer, *P. globosa* forms large biomass. In fact, *P.globosa* was identified as a potentially harmful species for several reasons. First, it releases dimethyl sulfide gas (DMS), which can irritate people's eyes, skin, and respiratory systems (Riegman & Van Boekel, 1996). Second, mucco-polysaccharides are abundant in its colonies. These polysaccharides are broken up by external factors like turbulence as well as internal factors like lysis and aging, which cause the accumulation of a thick, odorous foam on the coast. Besides, *Pseudo-nitzschia* complex needles can stick into *P. globosa* colonies and form structures that irritate filter feeders during *P. globosa* blooms. These structures' lesions may promote viral and bacterial infections in fish, thereby affecting higher trophic levels, and reducing biodiversity (Alain & David, 2022). Moreover, the neurotoxin domoic acid (DA) produced by *Pseudo-nitzschia* is responsible for the neurological disorder known

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

as amnesic shellfish poisoning (ASP) in humans. Additionally, marine mammals and seabirds can get poisoned if they consume DA-contaminated planktivorous prey (Delegrange et al., 2018).

The french monitoring of phytoplankton populations and associated environmental factors in the English Channel started in 1979 with RNO (Réseau National d'Observation) or RNC (Réseau Nationale de Contrôle). Then, in 1984, a national network called REPHY(le Réseau de Surveillance du Phytoplankton et des phycotoxines) was established by Ifremer, to estimate the abundance and taxonomic composition of phytoplankton, describe their spatio-temporal dynamics, detect toxin-producing species, and monitor and alert for harmful blooms (https://doi.org/10.17882/47248). After that, in 1992, the Artois-Picardy Water Agency and Ifremer decided to establish SRN (Suivi Régional des Nutriments) in response to the need of precise monitoring of nutrient concentration over a longer period of time, and to harmful algal blooms. Despite the fact these studies helped a lot in avoiding the detrimental effects of HABs, they alone remained insufficient to fully understand the dynamics of phytoplanktons and algal blooms (Dickey, 2003).

It was until 2004, when MAREL Carnot monitoring station has been installed in the French part of the English Channel. The MAREL (Mesures Automatisées en Réseau pour l'Environnement Littoral) Carnot station, developed and implemented by Ifremer (French Research Institute for Sea Exploitation), is a moored buoy protected by a tube and equipped with physicochemical and biological measuring devices and sensors that operate continuously and autonomously. This multi-sensor station is located in the Boulogne-sur-Mer harbor (Eastern English Channel) which is influenced by both marine and fresh waters. It is equipped with high-performance systems for seawater analysis and data transmission in near real time. It measures the following parameters with a high frequency resolution (20 minutes): estimated sea level, gust wind speed, wind direction relative to true north, horizontal wind speed, relative humidity, light irradiance surface PAR, sea temperature, practical salinity, pH, dissolved oxygen, oxygen saturation, fluorescence, and turbidity. For nutrients, including nitrate, phosphate, and silicate, the sampling frequency is set to 12 hours.

2. Objectives

The purpose of this article is to introduce the MAREL Carnot dataset and provide an overview of its variability. We will provide a detailed description of the MAREL Carnot station, including its deployment and measurements, for any future users of the associated dataset. Based on previous research publications, we aim to demonstrate that the MAREL Carnot dataset is useful for evaluating environmental or ecological statuses, marine phytoplankton ecology, physical oceanography, turbulence, as well as public policy.

3. Materials and Methods

3.1 Location and Study Area

In 2004, the MAREL station was installed on the Carnot sea wall, hence the name MAREL Carnot. It is situated on the French

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

Open Access Earth System Science Data Discussions

side of the Eastern English Channel at 50.7405N, and 1.5677W. In other words, this automated channel is situated at the exit of the Boulogne-sur-Mer harbour, which is France's first fishing port. Figure **1** below depicts the location of MAREL Carnot station on map.

In general, there is no seasonal pycnocline in the Eastern English Channel (ECC), and stratification is limited and sporadic depending on freshwater discharge levels. Water can be extremely turbid due to the continental shelf nature of its seabed, which can reach a maximal depth of 180 m depending on tidal regimes. Most importantly, ECC has macro-tidal regime in the Dover Strait which varies from 3 m to 9 m during neap tide and spring tide, respectively. This regime produces significant residual tidal currents from the English Channel to the North Sea, as well as high tidal currents that are nearly parallel to the shore. Fluvial supplies distributed throughout the French coast from the Bay of Seine to Cape Gris-Nez form a coastal water mass that floats near the shore and is protected from the open ocean by a frontal area (Brylinski et al., 1996) . This frontal area plays a significant role in structuring biological and non-biological exchange between coastal and offshore water masses. However, particle and nutrient transport, as well as exchanges between inshore and offshore water masses, are tide-dependent, with neap tides being stronger than spring tides.

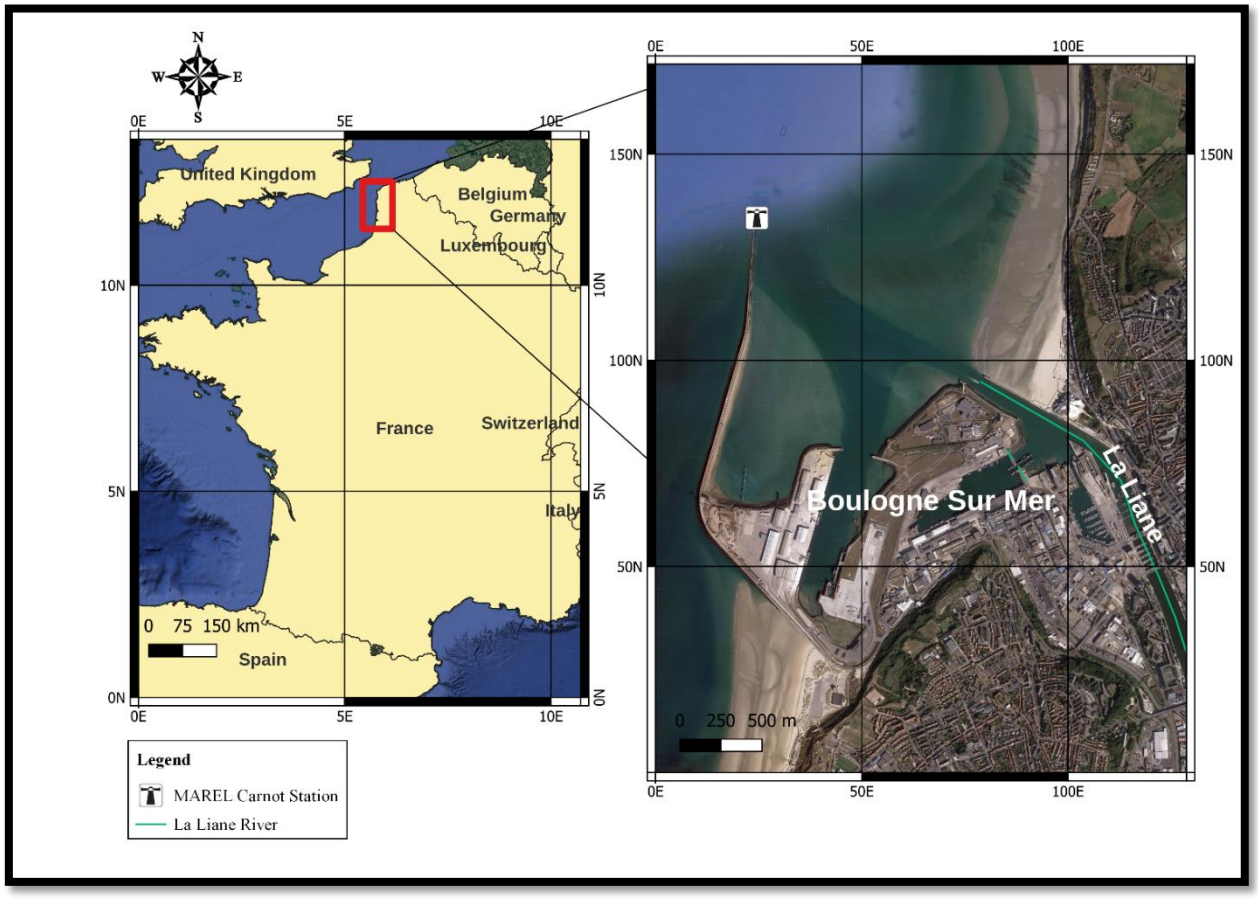


Figure 1 **The location of MAREL Carnot station in the Eastern English Channel (EEC) (Map data © 2022 Google Satellite).**

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

3.2 Description of the MAREL Carnot Station

MAREL is a French acronym for Mesures Automatisées en Réseau pour l'Environnement Littoral (automated sampling network for coastal waters). It belongs to a network of fixed platforms extending across the entire French coast called COAST-HF (https://coast-hf.fr). In fact, COAST HF is a component of the IR ILICO research infrastructure at the French national level (https://www.ir-ilico.fr/). MAREL Carnot station consists of a tube weighing 12 tons, and measuring 15 meters in length. Because MAREL Carnot is located in a megatidal zone, it is encased in a tube to protect it from strong currents, frequent storms, and boat collisions near the port. In other words, buoys are not designed for such environments, thus an infrastructure to retain the buoy in a specific location was required. However, such an infrastructure would be very huge and expensive, so the tube was the best solution. Figure 2 shows MAREL Carnot measuring station, which consists of the tube, along with the light house.

Its sensors are placed on a float inside the tube in order to follow tidal movements. Until 2014, it was made of a measurement cell containing several sensors. The seawater was pumped upwards to be analysed. During periods when there are no measurement cycles, the system was chlorinated via electrolysis to prevent biofouling. In other words, water was taken from the sub-surface at a depth of 1.5 meters, and then sent to the passage chamber to be redistributed to different sensors. In order to make measurements directly in-situ using a multi-parameter probe, the system was updated in 2014, and water circulation in the chamber was removed to avoid air intake, which would compromise measurements and data quality. The pulley system is placed in a chamber inside the harbor structure, allowing for the management of the cables during the tide's downward and upward movements, as well as the raising of the station for maintenance.

The MAREL Carnot automated station is built with 1990s electronic, computer, and mechanical components. The general aging process, which primarily affects marine-exposed systems, necessitates the replacement of a number of elements that are no longer functional and whose maintenance is impossible due to a lack of spare parts. This explains why the measurement system was replaced in 2014 with a new automated measuring probe. Hence, several data for the year 2014 are missing.

The core of the system is now composed of the following elements:

-a PLC type MAREL ESTRAN

-a small in situ circulation pump (pumping of the water on the probe)

- a chlorinator for the production of chlorine by electrolysis

-a multiparameter probe type MP6 nke

-a Systea nutrient analyzer (nitrate, phosphate, silicate)

-Seabird PAR Satlantic to measure the Photosynthetic Active Radiation,

-A PONCPC-EH-10 probe for pH measurement.

Measurements are taken at 3 different levels, numbered -1, 0, 1. Level -1 denotes atmospheric measurements (+ 28 m). Level 0 represents water surface measurements, while level 1 represents primary levels of immersion (-1.5 m).

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

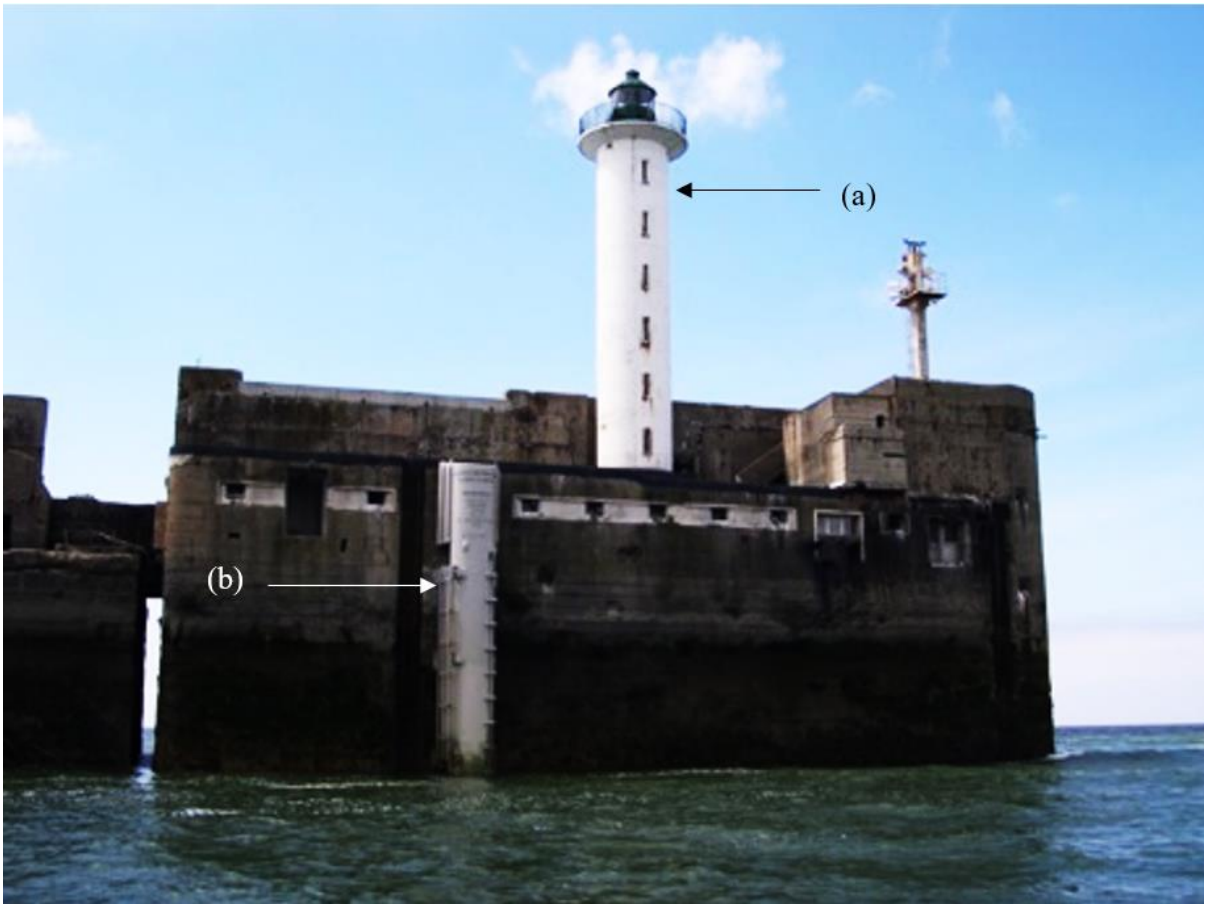


Figure 2 MAREL Carnot station consisting of the light house (a), and the MAREL Carnot tube (b) (photo © Ifremer).

3.3 Measured and Calculated Parameters

The MAREL Carnot multisensor station measures physiochemical and biological parameters in a continuous and autonomous mode. With a sampling frequency of 20 minutes, it is capable of providing high resolution data for conductivity ($S.m^{-1}$), water and air temperature (°C), pH, fluorescence (FFU), turbidity (NTU), dissolved oxygen concentration ($mg.L^{-1}$), **P**hotosynthetically **A**ctive **R**adiation or P.A.R (µmol of photons.s^{-1}.m^{-2}), wind direction (degree), and wind speed ($m.s^{-1}$), as well as sea level (m). On the other hand, nutrient concentration like nitrate, phosphate, and silicate are only measured once every 12 hours due to the limited amount of chemical reagents. As a result, taking measurements twice a day allows chemicals to last longer (3 months). In addition to these measurements, certain parameters are calculated such as oxygen saturation (%). Table 1 shows the different physiochemical parameters measured by MAREL Carnot station, along with their sensor and expert ranges.

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

Table 1 Sensor and expert ranges of the various parameters measured by MAREL Carnot station

Parameter	Unit	Sensor Range	Expert Range
Fluorescence	FFU	0 - 500	0.03 - 120
pH	-	1 - 14	6.5 - 9.5
Practical Salinity	PSU	2 - 42	5 - 35
Electrical Conductivity	S/m	0 – 7	3 - 6
Sea Water Temperature	°C	-5 - 35	0 - 30
Air Temperature	°C	-20 - 40	-20 - 45
P.A.R (Photosynthetic Active Radiation)	$\mu mol.s^{-1}.m^{-2}$	0 – 5000	0 - 2500
Turbidity	NTU	0 – 500	0 - 270
Nitrate +Nitrite Concentration	µmol/L	0 - 100	0 - 100
Phosphate Concentration	µmol/L	0 - 100	0 - 10
Silicate Concentration	µmol/L	0 - 100	0 - 50
Dissolved Oxygen	mL/L	-*	-*
Dissolved Oxygen	mg/L	-*	0 - 20
Oxygen Saturation	%	0 - 150	0 - 150
Horizontal Wind Speed	m/s	0 - 40	0 - 40
Wind Direction Relative True North	degree	0 - 360	0 - 360
Observed Sea Level	m	- **	0 - 20

*: Due to complicated calculations involving several formulas and rules, the sensor range of dissolved oxygen is unavailable in mL/L. Consequently, it is also absent in mg/L
**: There is no specific sensor range for Observed Sea Level

3.4 Pre-processing of MAREL Carnot dataset

Figure 3 shows the different steps performed for the dataset before visualization. Briefly, the steps of pre-processing are Data Correction and NA transformation, Quality Control (QC) correction, sensor range correction, as well as time alignment. Below is a detailed explanation of each step.

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

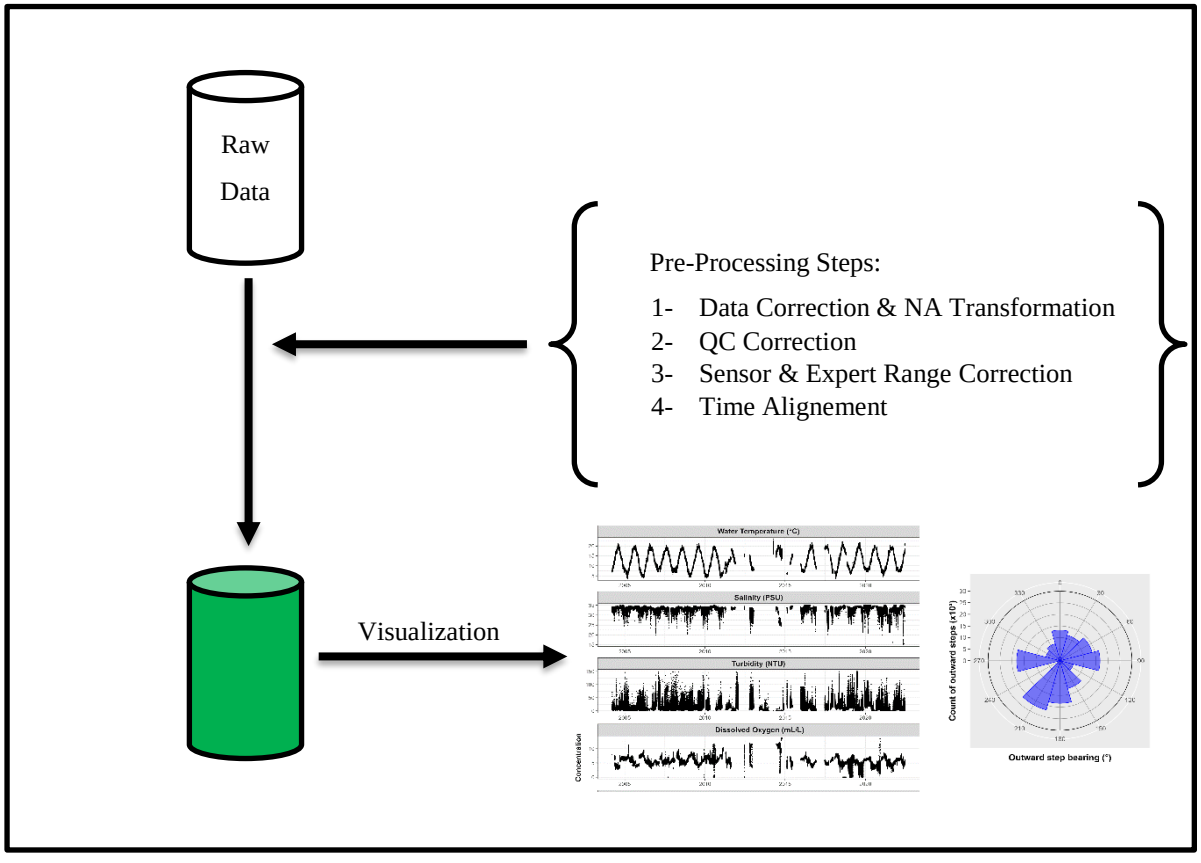


Figure 3 Schematic representation of the pre-processing steps and visualization of the dataset

3.4.1 Data Correction and NA Transformation

First, we checked the data to see if it contains major errors. After the year 2020, we corrected an offset in the PAR (Photosynthetically Active Radiation) variable. Then, we discovered some data for nutrients present at Level 2. Since MAREL Carnot doesn't contain Level 2 data, we knew that these nutrients belong to Level 1, but were wrongly introduced into Level 2. Hence, we made sure values are similar at Level 1 and Level 2, and remove all values from Level 2.

In addition, missing values in datasets are typically represented as NA, which stands for *Not Available*. However, in some cases, NA values are replaced with other numbers such as 77.77, 7777, 999, 999.999, 9999.99... Etc. A dataset may also include values like Inf, which stands for infinity, and Nan, which stands for Not A Number. Because these types of observations can affect or even obstruct further processing steps, we convert them into something feasible, which is NA.

3.4.2 Quality Code Correction

Coastal CORIOLIS, or simply CORIOLIS, is a data portal for all in situ data platforms in Coastal French waters, including

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

MAREL Carnot (https://data.coriolis-cotier.org). CORIOLIS quality control procedure provide the users with the quality of each measurement as a Quality Code (QC). This code is normally given following the completion of quality control procedures, which are a part of the CORIOLIS harmonized method. During this process, the data is automatically verified using fundamental statistics (minimum, maximum, median, and standard deviation), and is subsequently validated or modified by an expert using more sophisticated methods and based on his environmental expertise. Table 2 shows the significance of the quality code utilized with MAREL Carnot dataset. Thus, all data with QC =4 (Bad data) were deleted, and replaced with *NA* values.

Table 2 Significance of the quality code (QC)

Quality Code	Significance
0	No Quality Code was performed
1	Good data
2	Probably Good data
3	Probably bad data that are potentially correctable
4	Bad data
5	Value Changed
6	Not Used
7	Not Used
8	Interpolated Value
9	Missing Value

It is worth noting that the QC procedure is not always accurate, experts cannot verify all measurements, and a false value may be found under a different code, like QC=2 or QC=3. For instance, an oxygen reading of 0 is wrong but may not be consistent with QC=4. As a result, we sought an additional method for automatically eliminating a sizable portion of the potential false data in addition to QC. Hence, correction is usually performed using both "sensor" and "expert" ranges.

3.4.3 Sensor and Expert Range Correction

The sensor range is an interval of correct values defined by the manufacturer, while the expert range is an interval of correct values defined by a field expert. Indeed, the sensor ranges were obtained from the information provided by the sensor suppliers, whereas the expert ranges were derived from expert judgment based on specific knowledge acquired in the studied area through previous research activities. For all of the parameters, only the values that fall within the sensor and expert ranges are kept. Values that fall outside of the ranges are replaced with missing data (*Not Available* or NA).

The sensor and expert ranges for MAREL Carnot are represented in Table 1. Indeed, the expert range is more precise than the sensor range. For instance, the sensor may give us a salinity value of 38, but our specialists know that salinity can only reach 35 in the Eastern English Channel, so the sensor's result is qualified as false and must be corrected. Indeed, it is worth mentioning that scientists willing to use this dataset for any research objectives shall perform this additional pre-processing

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

step in order to achieve higher levels of accuracy and precision.

3.4.4 Time Alignment

Before statistical methods can be applied to the dataset, it must have an identical time interval between each measurement. However, the measurements of the various sensors are not taken at the same time, resulting in a time lag that can range from few seconds to several minutes. In addition, the series may contain duplicates in some cases. Thus, to eliminate potential replicates and synchronize the dataset, we perform a temporal alignment using the average time interval of the measurements. The alignment protocol calculates the average time step. This time step creates a regular/no-replicates time variable. Based on the parameter to be regularized and the goals to be attained, the maximum/minimum/or average of all subsets of our dataset matching to each regular interval of our ideal time variable is then returned.

In this paper, the maximum value is chosen for all variables except oxygen, where the minimum value is used. This is because during phytoplankton blooms, the amount of oxygen in the water drops. Hence, it is more interesting to use the minimum values to highlight this feature of HABs.

4 Results and Discussion

To summarize, several pre-processing procedures were performed on the displayed dataset, including data correction and NA transformation, quality code correction, sensor range correction, as well as temporal alignment. Table 3 represents the descriptive statistics for the main physiochemical parameters measured by MAREL Carnot from 2004 until 2022. The results show a high percentage of missing data, denoted as NA, or *Not Available*. In fact, missing data is a major problem in time series. It is primarily due to sensor failure, communication problems, or sensor maintenance disability.

Table 3 Statistical summary (minimum, first quartile, median, mean, third quartile, maximum, and percentage of NA) of the parameters measured by MAREL Carnot station

Parameters (Units)	Min	Q1	Median	Mean	Q3	Max	percentage of NA
Air Temperature (°C)	-6.18	7.82	11.90	11.71	16.11	35.00	44.58
Water Temperature (°C)	3.60	8.73	12.80	12.94	17.30	23.50	23.73
Salinity (PSU)	10.04	33.07	33.64	33.32	34.09	35.00	28.45
Turbidity (NTU)	0.00	4.63	8.91	14.62	17.30	259.70	25.25
P.A.R (µmol.s^{-1}.m^{-2})	0.00	0.00	25.00	282.26	373.20	2489.03	47.81
Wind Direction (Degree)	0.00	92.00	198.00	177.99	246.00	359.90	48.14
Horizontal Wind Speed (m.s^{-1})	0.00	5.77	9.08	9.71	13.12	39.96	48.23
Fluorescence (FFU)	0.03	0.52	1.08	3.04	2.48	116.59	20.93
Nitrate + Nitrite (µmol.L^{-1})	0.01	5.81	13.22	17.44	24.06	99.54	65.50
Phosphate (µmol.L^{-1})	0.00	0.48	0.71	0.95	0.96	10.00	66.00
Silicate (µmol.L^{-1})	0.00	2.10	4.44	5.55	7.79	49.04	64.73
Dissolved Oxygen (mL.L^{-1})	0.00	4.80	5.60	5.67	6.46	13.92	28.76

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

Parameters (Units)	Min	Q1	Median	Mean	Q3	Max	percentage of NA
Oxygen Saturation (%)	0.00	83.78	91.68	88.89	97.30	198.72	54.78
pH	6.50	7.92	8.10	8.14	8.38	9.33	52.66
Conductivity ($S.m^{-1}$)	3.00	3.60	3.94	3.99	4.41	4.96	54.10
Sea Level (m)	5.43	8.08	10.20	10.28	12.32	15.84	57.29

Figure 4 and Figure 5 show the signals collected from MAREL Carnot station from 2004 until 2022. We noticed that some signals have visible cycles, such as water and air temperature as well as photosynthetic active radiation (PAR). In addition, the signals contain episodic or continuous missing values over several time periods. For instance, a large number of missing values can be found around the year 2014 almost in all signals. This is due to station and sensor alterations that occurred during that time, particularly, the replacement of several sensors with a multi-parameter probe (Lefebvre & Schmitt, 2016). Similarly, signals of air temperature, PAR, wind speed as well as sea level have been lost for several years. However, conductivity signals have only been available since 2015.

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

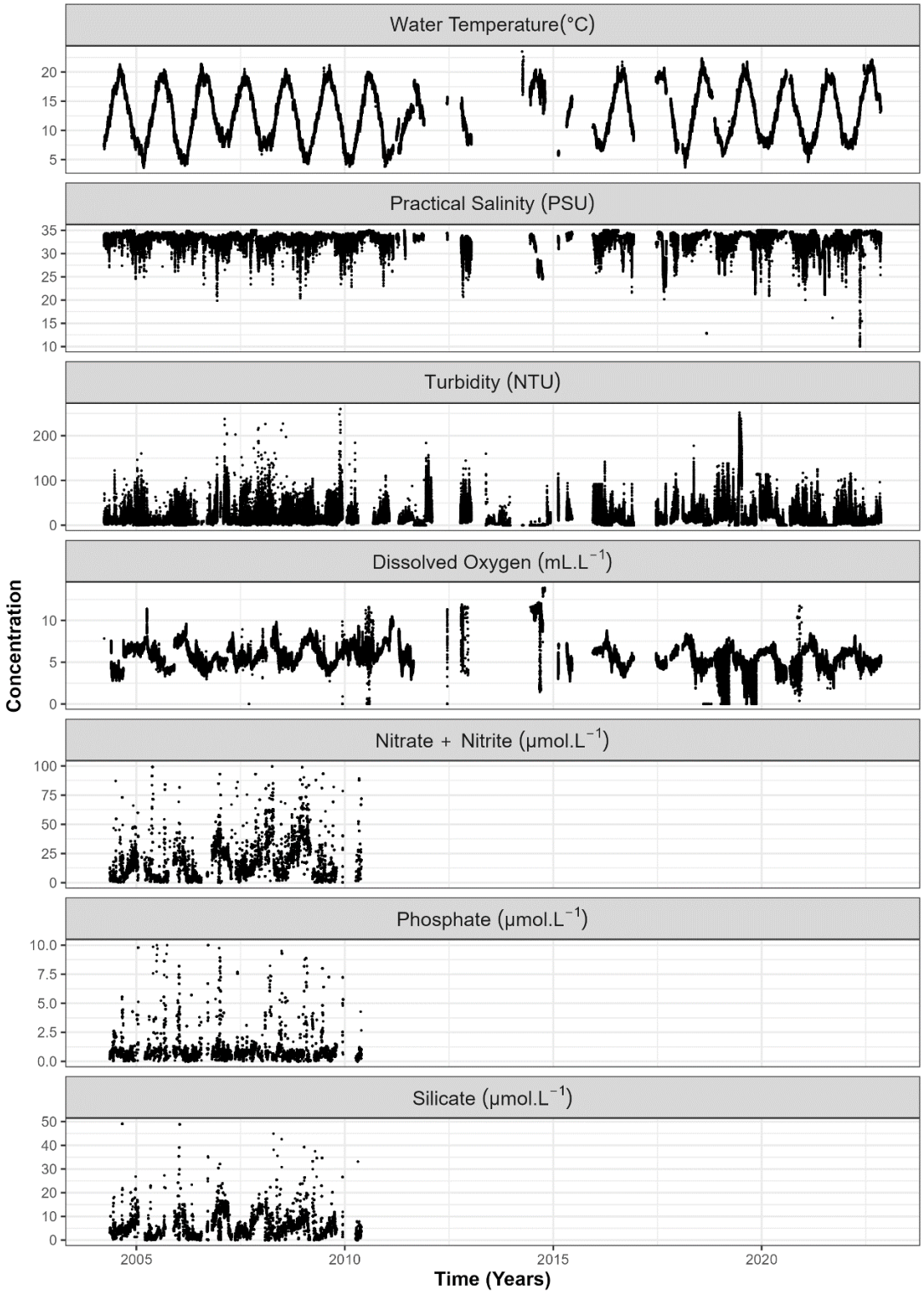
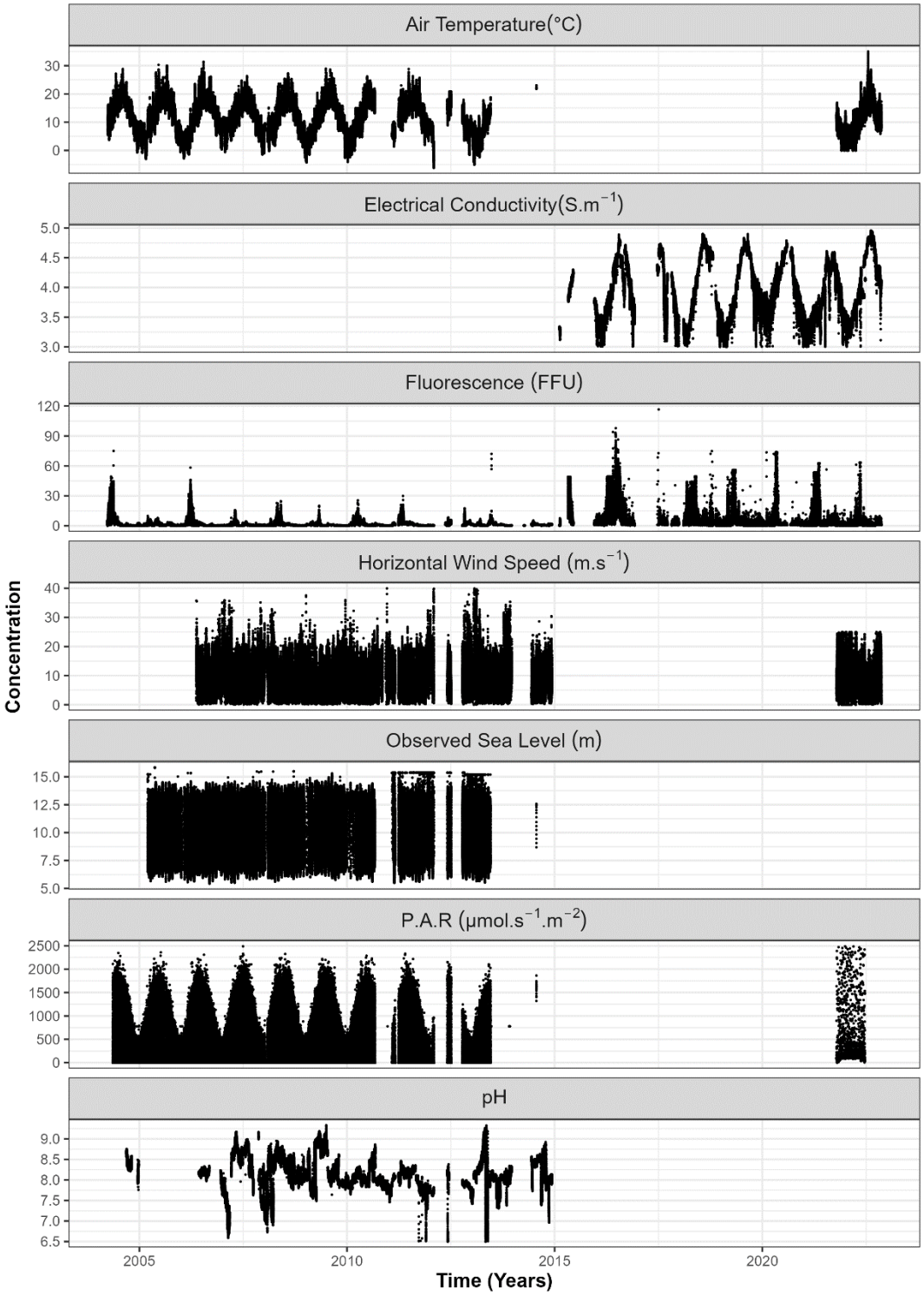


Figure 4 Signals collected from the MAREL Carnot station during the period 2004-2022

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023



240 **Figure 5 Signals collected from the MAREL Carnot station during the period 2004-2022**

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

Indeed, high frequency fluorescence data from MAREL Carnot station can contribute to the calibration of satellite observations. In addition, MAREL Carnot perform measurements of nutrients that are not provided by current spacecraft techniques. Hence, it is much more effective than satellites at monitoring water quality (Lefebvre & Schmitt, 2016). Nonetheless, nutrient signals such as phosphate, nitrate, and silicate are only available until 2010. This is caused by a previous sensor failure and the inability to replace it.

Figure 6 shows a wind rose for the frequency, wind speed and direction relative true north collected by MAREL Carnot station, after removing all NA values.

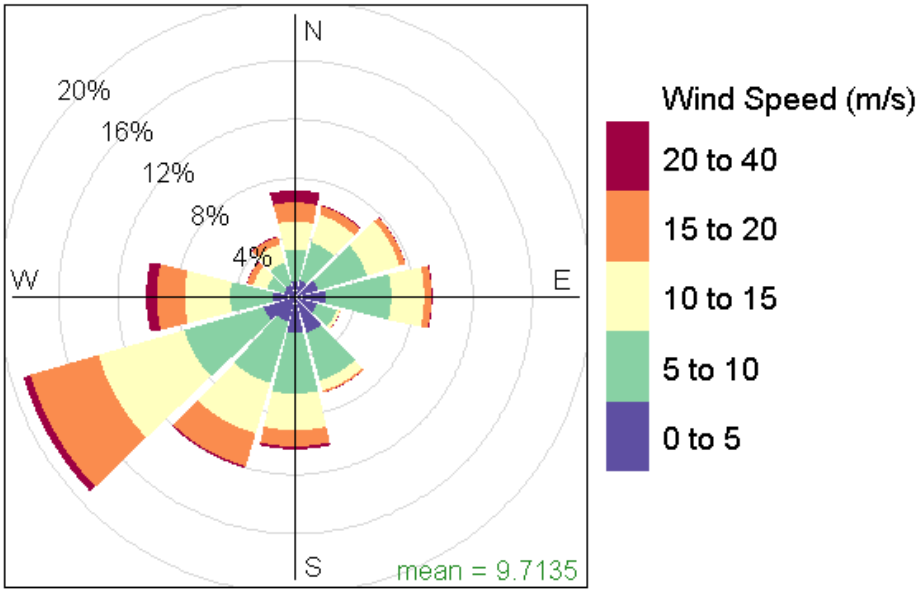


Figure 6 Wind Rose representing the wind speed and wind direction measured by MAREL Carnot station

Indeed, scientists from several disciplinary backgrounds have utilized MAREL Carnot data to accomplish a wide range of research objectives. In the following paragraphs, we will go over some of the most significant findings from several research efforts. The scientific community that is interested in the MAREL Carnot dataset may find this evaluation useful in determining which topics may or may not require further study based on the results of this evaluation. In general, this dataset allows researchers to investigate the dynamics of phytoplankton as well as detect blooms caused by human activities and/or climate change.

For instance, Rousseeuw et al., (2015) developed an unsupervised Hidden Markov Model for monitoring the marine environment, specifically for detecting algal blooms and understanding phytoplankton dynamics. In his unsupervised Hidden Markov Model, uHMM parameters were estimated using spectral clustering rather than the commonly used iterative Expectation Maximization. The results obtained using MAREL Carnot dataset showed that the proposed system is efficient to detect the main productive and non-productive periods, as used for the purposes of the EU Water Framework Directive to assess good environmental status, and refine knowledge about phytoplankton bloom dynamics in a temperate ecosystem,

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

temporarily dominated by a harmful algae, *Phaeocystis globosa*. Thus, the suggested uHMM system successfully characterizes phytoplankton dynamics from new incoming data (in near real-time), and will enable researchers to gain a better understanding of the main controlling or forcing parameters (e.g., nutrient pressure, light availability, turbidity), the environmental status (e.g., phytoplankton biomass), and the direct and/or indirect effects of such blooms (e.g., oxygen concentration). Most importantly, the ability of uHMM to establish environmental states represents a clear potential to better understand what a good environmental condition is, as defined and applied for the needs of the WFD, MSFD, or other regional sea conventions such as OSPAR. Even though uHMM was only applied to the MAREL Carnot dataset, it could contribute to the processing of huge multivariate time series generated by high resolution platforms, which are increasingly used for the integrated observation of pelagic ecosystems and biogeochemical cycles in oceans (Rousseeuw et al., 2015).

On the other hand, Grassi et al., (2019) suggested a Multilevel Spectral Clustering (M-SC) to split multivariate time series from general patterns to extreme events without *a priori* knowledge. The results obtained from MAREL Carnot dataset have shown that we can extract knowledge on dynamics of events or environmental states. In addition, it was shown that M-SC allows unsupervised labelling of time series, which is a basic part of machine learning and is needed to build an event prediction system and come up with sampling strategies that work close to real time (Grassi et al., 2019). As a result, scientists will be able to create a HAB early warning expert system to warn shellfish farmers, and prevent both public health risks and commercial losses in the shellfish farming business.

Nevertheless, datasets, notably MAREL Carnot, are typically incomplete and contain a significant amount of missing data due to sensor failures, communication/transmission difficulties, or poor weather conditions for manual measurements or maintenance. Phan et al., (2018) proposed Dynamic Time Warping method to fill in successive missing values of univariate time series as well as low uncorrelated multivariate time series (Phan et al., 2017).

Furthermore, to compare alternative approaches to studying and understanding HABs, all researchers must have access to the same datasets. For instance, researchers may now assess the effectiveness of new and old machine learning algorithms in understanding the dynamics and forecasting harmful algal blooms. Thus, a comparative study of clustering approaches applied to spatial or temporal pattern discovery gave promising results in the segmentation of both UCI databases and marine time series compared to other approaches (Grassi et al., 2020). Therefore, we may conclude that the MAREL Carnot dataset is beneficial not just for marine ecologists, but also for machine learning specialists and data scientists. It is worth mentioning that all the above algorithms are available and published on the Comprehensive R Archive Network (CRAN).

Another study performed by Derot et al., (2020) analysed how forecasts of phytoplankton blooms are impacted by different sampling frequencies. They applied Random Forest (RF) and sliding window strategy on 12 parameters derived from MAREL Carnot dataset. This research demonstrated that the sampling frequency has a direct impact on the forecast performance of a Random Forest (RF) model as high-frequency datasets might provide useful information to the RF. Furthermore, this type of model sets the groundwork for the creation of a numerical decision-making tool that could help mitigate the impact of algal blooms, and can recreate interactions that closely resemble the real biological processes (Derot et al., 2020).

Moreover, MAREL Carnot dataset is useful for studying turbulence. In fact, many fields in the marine environment fluctuate

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

over a wide variety of geographical and temporal scales. To study their dynamics and estimate their variations at all scales, high frequency measurements are needed (Huang & Schmitt, 2014). Hence, Derot et al., (2015) investigated phytoplankton biomass during bloom by applying Empirical mode decomposition (EMD) on fluorescence dataset from MAREL CARNOT. Results revealed that bloom events include considerable internal variations. In other words, blooms are not smooth and "mountain-like", but exhibit high frequency oscillations due to turbulent advection and complex population dynamics (Derot et al., 2015). Similarly, Huang & Schmitt., (2014) analysed time dependent intrinsic correlation analysis of temperature and dissolved oxygen time series using empirical mode decomposition. The anti-correlation between temperature and oxygen showed that higher temperatures may favor larger phytoplankton growth rate, and hence, with a time delay, a lower percentage of oxygen (Huang & Schmitt, 2014). In addition, Zongo & Schmitt, (2011) demonstrated that pH fluctuations in marine waters are strongly influenced by turbulent hydrodynamical transport, and may be considered as a turbulent active scalar (Zongo & Schmitt, 2011).

Overall, the MAREL Carnot station provides automatic, continuous, and long-term observation of various physiochemical and biological parameters that allow for monitoring the general quality of marine environment, detecting harmful algal blooms (HAB) and understanding phytoplankton dynamics. Hence, MAREL Carnot dataset aligns with objectives of SRN (Suivi Régional des Nutriments in French, Regional Nutrients Monitoring Program), especially by assessing the influence of continental inputs on the marine environment, and their implication on possible eutrophication, and can assist in estimating the effectiveness of development and management policies in the marine coastal zone (Alain & David, 2022). In other words, MAREL Carnot is the first sampling station for the SRN transect. Thus, it assists in understanding phytoplankton dynamics by determining recurrent, extreme and rare events.

Furthermore, MAREL Carnot dataset can contribute to both REPHY (Observation and Surveillance Network for Phytoplankton and Hydrology in coastal waters) (https://doi.org/10.17882/47248), and REPHYTOX (Monitoring Network for Phycotoxins in marine organisms) (https://doi.org/10.17882/47251). Actually, the goal of REPHY is to measure the biomass, abundance, composition, and hydrological parameters of marine phytoplankton in coastal and lagoon waters. REPHYTOX is designed to find and track three types of toxins that can build up in bivalve mollusks and cause DSP (Diarrheic Shellfish Poisoning), PSP (Paralytic Shellfish Poisoning), and ASP (Amnesic Shellfish Poisoning) (Belin et al., 2021). Monitoring carried out by MAREL Carnot in parallel with REPHY and REPHYTOX permits continuous adaptation to the objectives, developing analysis strategies with extensive and complex data, thereby ensuring sustainability, which were challenges faced by REPHY and REPHYTOX before.

Thus, the contribution of MAREL Carnot to improve assessment based on low frequency renders it important to achieve the objectives of WFD (Water Framework Directive) and MSFD (Marine Strategy Framework Directive). Besides, the lighthouse's sensors provide valuable data for meterological research and may improve local weather forecasts by measuring variables including wind speed, wind direction and air temperature. Additionally, depending on the goals, some of the parameters determined by our station can actually be useful for fisheries research when making the link between the different trophic levels.

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

While MAREL Carnot has made substantial progress toward automating marine ecosystem monitoring, there are still some significant challenges to overcome. In fact, it can be interrupted by rough sea conditions, such as strong tidal currents and storms. In addition, biofouling presents a major problem for sensors in the coastal environment, which explains why only a few moored autonomous systems have been deployed in the coastal environment despite their ease of maintenance (Blain et al. 2004). Also, new EOV (Essential Ocean Variables) and EBV (Essential Biodiversity Variables) might be added with time. This adds a further obstacle, as it may be necessary to install brand-new sensors for the updated parameters. Above all, the major challenge will continue to be the issue of missing values, especially when it comes to data that has been missing for a long time, as in the case of nutrients, where nitrate, phosphate, and silicate observations have been missing since 2010 due to sensor failure.

In future work, we plan to use a multi-scale, multi-source, multi-criteria, and multi-parameter approach to characterize and predict harmful algal blooms in the Eastern English Channel caused by *Phaeocystis globosa* and *Pseudo-nitzschia spp.*. We will do this by combining high frequency datasets from MAREL Carnot, satellite, and modeling data with low frequency datasets from other sources. This integrated observing system will be used to identify environmental states present in the region, and develop an early warning system that can anticipate harmful algal blooms.

4.1 Data Availability

The raw data are present on the official Coriolis website. "These data were collected and made freely available by the Coriolis project and programmes that contribute to it (http://www.coriolis.eu.org)." The dataset after quality control procedures are present on the SEANOE website (DOI: 10.17882/39754) (Lefebvre, 2015). In fact, our data are made available according to the FAIR approach (Findable, Accessible, Interoperable, Reusable).

5 Conclusion

In conclusion, this high frequency dataset is useful in many scientific fields, such as phytoplankton ecology, data science, and turbulence. It can be used to describe and predict harmful algal blooms in the Eastern English Channel, which is important to warn shellfish farmers and prevent economic losses and health problems. It can also be used with satellite, modeling, and low-frequency in situ data to achieve better knowledge and understanding of the marine ecosystem. Most importantly, this data set has been shown to be useful for fulfilling the goals set by the Water Framework Directive (WFD), the Marine Strategy Framework Directive (MSFD), and the Oslo and Paris Conventions (OSPAR). This would help researchers in the future get better results, which would lead to more scientific progress.

Competing Interests

The authors declare that they have no conflict of interest

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

Author Contribution

Raed Halawi Ghosn wrote the paper. Alain Lefebvre led the conceptualization, the writing of the paper and he also led the funding acquisition, the scientific coordination of MAREL Carnot related activities since 2002.We highly appreciate the effort of each of Émilie Poisson-Caillault, Guillaume Charria, Armel Bonnat, and Michel Repecaud for their contribution in data pre-processing. We would also like to sincerely thank each of Jean-Vallery FACQ, Loïc QUÉMÉNER, Vincent DUQUESNE, and Camille BLONDEL for all their effort in providing technical information, and maintaining in operational conditions.

Acknowledgement

This PhD project is funded by the Office Français de la Biodiversité (OFB) and by Ifremer (grant agreement OFB.21.0578). This work has been financially supported (i) by the European Union (ERDF), the French State, the French Region Hauts-de-France and Ifremer, in the framework of the project CPER MARCO 2015-2021, (ii) by the European Union's Horizon 2020 research and innovation programme under grant agreement No 654410 in the framework of the project JERICO S3, and (iii) by the Artois Picardy Water Agency.

Bibliography

Belin, C., Soudant, D., & Amzil, Z. (2021). Three decades of data on phytoplankton and phycotoxins on the French coast: Lessons from REPHY and REPHYTOX. *Harmful Algae*, *102*(December 2019), 101733. https://doi.org/10.1016/j.hal.2019.101733

Blain, S., Guillou, J., Tréguer, P., Woerther, P., Delauney, L., Follenfant, E., Gontier, O., Hamon, M., Leildé, B., Masson, A., Tartu, C., & Vuillemin, R. (2004). High frequency monitoring of the coastal marine environment using the MAREL buoy. *Journal of Environmental Monitoring*, *6*(6), 569–575. https://doi.org/10.1039/b314073c

Brylinski, J. M., Brunet, C., Bentley, D., Thoumelin, G., & Hilde, D. (1996). Hydrography and phytoplankton biomass in the Eastern English Channel in spring 1992. *Estuarine, Coastal and Shelf Science*, *43*(4), 507–519. https://doi.org/10.1006/ecss.1996.0084

Delegrange, A., Lefebvre, A., Gohin, F., Courcot, L., & Vincent, D. (2018). Pseudo-nitzschia sp. diversity and seasonality in the southern North Sea, domoic acid levels and associated phytoplankton communities. *Estuarine, Coastal and Shelf Science*, *214*(April), 194–206. https://doi.org/10.1016/j.ecss.2018.09.030

Derot, J., Schmitt, F. G., Gentilhomme, V., & Zongo, S. B. (2015). Long-term high frequency phytoplankton dynamics, recorded from a coastal water autonomous measurement system in the eastern English Channel. *Continental Shelf Research*, *109*, 210–221. https://doi.org/10.1016/J.CSR.2015.09.015

Derot, J., Yajima, H., & Schmitt, F. G. (2020). Benefits of machine learning and sampling frequency on phytoplankton bloom

forecasts in coastal areas. *Ecological Informatics*, *60*(May), 101174. https://doi.org/10.1016/j.ecoinf.2020.101174

Dickey, T. D. (2003). Emerging ocean observations for interdisciplinary data assimilation systems. *Journal of Marine Systems*, *40–41*, 5–48. https://doi.org/10.1016/S0924-7963(03)00011-3

Grassi, K., Caillault, E. P., & Lefebvre, A. (2019). Multilevel spectral clustering for extreme event characterization. *OCEANS 2019 - Marseille, OCEANS Marseille 2019*, *2019-June*(June). https://doi.org/10.1109/OCEANSE.2019.8867261

Grassi, K., Poisson-Caillault, É., Bigand, A., & Lefebvre, A. (2020). Comparative study of clustering approaches applied to spatial or temporal pattern discovery. *Journal of Marine Science and Engineering*, *8*(9). https://doi.org/10.3390/JMSE8090713

Huang, Y., & Schmitt, F. G. (2014). Time dependent intrinsic correlation analysis of temperature and dissolved oxygen time series using empirical mode decomposition. *Journal of Marine Systems*, *130*, 90–100. https://doi.org/10.1016/J.JMARSYS.2013.06.007

Karasiewicz, S., & Lefebvre, A. (2022). Environmental Impact on Harmful Species Pseudo-nitzschia spp. and Phaeocystis globosa Phenology and Niche. *Journal of Marine Science and Engineering*, *10*(2). https://doi.org/10.3390/jmse10020174

Le Moal, M., Gascuel-Odoux, C., Ménesguen, A., Souchon, Y., Étrillard, C., Levain, A., Moatar, F., Pannard, A., Souchu, P., Lefebvre, A., & Pinay, G. (2019). Eutrophication: A new wine in an old bottle? *Science of the Total Environment*, *651*, 1–11. https://doi.org/10.1016/j.scitotenv.2018.09.139

Lefebvre, A. (2015). MAREL Carnot data and metadata from Coriolis Data Centre. SEANOE. https://doi.org/10.17882/39754

Lefebvre, A., & Devrecker, D. (2022). *How to learn more about hydrological conditions and phytoplankton dynamics and diversity in the eastern English Channel and the southern bight of the North Sea ? the SRN data set (1992-2021). July*, 1–21. https://doi.org /10.5194/essd-2022-146

Lefebvre, A., & Schmitt, F. G. (2016). *In Mesures à haute résolution dans l'environnement marin côtier* (CNRS Edition). CNRS.

Phan, T. T. H., Bigand, A., & Caillault, É. P. (2018). A New Fuzzy Logic-Based Similarity Measure Applied to Large Gap Imputation for Uncorrelated Multivariate Time Series. *Applied Computational Intelligence and Soft Computing*, *2018*. https://doi.org/10.1155/2018/9095683

Phan, T. T. H., Caillault, E. P., Lefebvre, A., & Bigand, A. (2017). Which DTW method applied to marine univariate time series imputation. *OCEANS 2017 - Aberdeen*, *2017-Octob*, 1–7. https://doi.org/10.1109/OCEANSE.2017.8084598

Riegman, R., & Van Boekel, W. (1996). The ecophysiology of Phaeocystis globosa: A review. *Journal of Sea Research*, *35*(4), 235–242. https://doi.org/10.1016/S1385-1101(96)90750-9

Ross Brown, A., Lilley, M. K. S., Shutler, J., Widdicombe, C., Rooks, P., McEvoy, A., Torres, R., Artioli, Y., Rawle, G., Homyard, J., Tyler, C. R., & Lowe, C. (2022). Harmful Algal Blooms and their impacts on shellfish mariculture follow regionally distinct patterns of water circulation in the western English Channel during the 2018 heatwave. *Harmful Algae*, *111*(June 2020), 102166. https://doi.org/10.1016/j.hal.2021.102166

https://doi.org/10.5194/essd-2023-8
Preprint. Discussion started: 31 January 2023

Rousseeuw, K., Poisson Caillault, E., Lefebvre, A., & Hamad, D. (2015). Hybrid hidden markov model for marine environment monitoring. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(1), 204–213. https://doi.org/10.1109/JSTARS.2014.2341219

Zongo, S. B., & Schmitt, F. G. (2011). Scaling properties of pH fluctuations in coastal waters of the English Channel: pH as a turbulent active scalar. *Nonlinear Processes in Geophysics*, *18*(6), 829–839. https://doi.org/10.5194/NPG-18-829-2011