D3.6: Harmful algae bloom species product documentation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348.

Project no.	776348
Project acronym:	CoastObs
Project title:	Commercial service platform for user-relevant coastal water monitoring services based on Earth Observation
Instrument:	H2020-EO-2017
Start date of project:	01.11.2017
Duration:	36 months
Deliverable title:	
Due date of deliverable:	
Organisation name of lead contractor for this deliverable:	

Author list:

Name	Organisation
Luis González Vilas	UVIGO
Jesus M. Torres Palenzuela	UVIGO
Robyn Gwee	Water Insight
Steef Peters	Water Insight

Dissemination level		
PU	Public	X
CO	Confidential, restricted under conditions set out in Model Grant Agreement	
CI	Classified, information as referred to in Commission Decision 2001/844/EC	

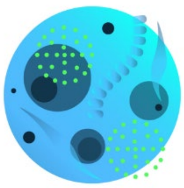
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

History			
Version	**Date**	**Reason**	**Revised by**
0.9	27-5-2019	Merge of HAB and *Phaeocystis* parts	Steef Peters
1	27-5-2019	Completed Draft	Luis González Vilas
2	30-5-2019	Revised Draft	Luis González Vilas
3	31-5-2019	Final version	Steef Peters

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

CoastObs Project

CoastObs is an EU H2020 funded project that aims at using satellite remote sensing to monitor coastal water environments and to develop a user-relevant platform that can offer validated products to users including monitoring of seagrass and macroalgae, phytoplankton size classes, primary production, and harmful algae as well as higher level products such as indicators and integration with predictive models.



phytoplankton

seagrass

harmful algal blooms

primary production

To fulfil this mission, we are in dialogue with users from various sectors including dredging companies, aquaculture businesses, national monitoring institutes, among others, in order to create tailored products at highly reduced costs per user that stick to their requirements.

With the synergistic use of Sentinel-3 and Sentinel-2, CoastObs aims at contributing to the sustainability of the Copernicus program and assisting in implementing and further fine-tuning of European Water Quality related directive.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Partnership

Water Insight BV. (WI)

The University of Stirling (USTIR)

Consiglio Nazionale Delle Ricerche (CNR)

Universite de Nantes (UN)

HZ University of Applied Sciences (HZ)

Universidad de Vigo (UVIGO)

Bio-Littoral (BL)

Geonardo Environmental Technologies Ltd. (GEO)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

TABLE OF CONTENTS

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 776348

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

FIGURES

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

TABLES

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

ABBREVIATIONS

List of abbreviations	
Abbreviation	Explanation
AOP	Apparent Optical Parameters
CDOM	Coloured Dissolved Organic Matter
EO	Earth Observation
FCM	Fuzzy C-Means
HABs	Harmful Algae Blooms
HPLC	High Performance Liquid Chromatography
MLP	MultiLayer Perceptron
MSI	Multispectral Instrument
NN	Neural Networks
OLCI	Ocean and Land Colour Instrument
RBF	Radial Basis Function
ROC	Receiver Operating Characteristic
SVM	Support Vector Machines
TSM	Total Suspended Matter

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Summary

The purpose of Work Package 3 is the development of a set of basic and innovative products based on Earth Observation (EO) data in order to improve monitoring of coastal water environments, including monitoring of harmful algae blooms (HABs). CoastObs task 3.4 aims to the development of a set of innovative products based on EO data for detection and monitoring of HABs.

HABs are an increasingly frequent phenomenon in coastal regions around the world. HABs can cause ecosystem damages, since they can be toxic to human health and other organisms, as well as have a strong impact on human activities such as fishing and aquaculture. For example, in Galician coast, a site of extensive mussel culture and CoastObs study area, HABs have had an important ecological, economic and social impact since they are often responsible for the closure of the mussel farming polygons. HABs detection and monitoring is traditionally based on direct observations, i.e. field samplings at fixed sampling stations. As compared to traditional sampling methods, satellite methods are faster, more cost-effective and produce map outputs providing a more synoptic view of the study area with a good temporal coverage.

The development of the EO-based products in task 3.4 is mainly focused on the three HAB taxa (namely *Pseudo-nitzschia* spp., *Alexandrium minutum*, *Phaeocystis* spp.), which are known to cause substantial problems in coastal waters, but also on the improvement of chlorophyll a (chl*a*) retrieval algorithms. Information about these taxa is of high importance and interest for the CoastObs users.

More specifically in task 3.4, we have developed a set of basic and innovative products based on Sentinel-2and Sentinel-3 data for the detection and monitoring of HABs in coastal waters:

- Regional algorithms for the retrieval of chl*a* concentration from Sentinel-3 images, which are aimed at the generation of accurate chl*a* maps allowing the study of the temporal and spatial distribution of the phytoplankton abundance in the area.
- Species indicators for two potentially toxic taxonomic groups: *Pseudo-nitzschia* spp. and *Alexandrium minutum*. Both species indicators aim at the production of bloom probability maps from Sentinel-2and Sentinel-3 images. Due to the complexity of the Galician Case 2 waters, algorithms were mainly based on machine learning methods relating satellite images to *in situ* data. In a later stage of the project, maps products will also be integrated into higher-level products to improve the detection and monitoring of HABs in Galicia.
- Although *Phaeocystis* spp. is not strictly considered to be a HAB species, it has been reported to cause serious ecological damages in the Southern Bight of the North Sea. A new product for the detection of *Phaeocystis* blooms has also been developed within

CoastObs Task 3.4. The algorithm takes into account salinity, irradiance and distance from coast, and is aimed at providing alerts for coastal zones or shellfish farms. In a later stage, it will be improved by combining it with a chlorophyll-a product or a dedicated optical algorithm using relevant spectral information.

Results show that map products based on EO data can provide useful information for improving detection and monitoring of HABs. All the products require further validation and testing. Within CoastObs, products will be also evaluated and improved according to the feedback provided by the final users.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

1 Introduction

CoastObs develops a user-relevant platform that offers user-relevant innovative and higher-level information products and coastal monitoring services. These services aim to be fully automated, commercial, reliable and sustainable. The validated basic and innovative products will be flexibly combined into higher-level products to fit the users' information needs.

Work Package 3 is aimed at the development of a set of services and products based on the combination of Earth Observation (EO) data with *in situ* information. Both basic products and a set of innovative products are being developed in accordance with users' needs in order to provide them with useful information to monitor coastal water environments and deal with the main problems affecting the different study areas.

Within WP3, task 3.4 is specifically aimed at the development of maps products based on Sentinel images that are expected to provide useful information for Harmful Algae Bloom (HABs) detection and monitoring in Galicia. Products include chl*a* concentration and species indicators for two taxonomic groups: *Pseudo-nitzschia* spp. and *Alexandrium minutum*. The task also concerns the further development of an algorithm or indicator for the presence of *Phaeocystis*, a nuisance algae species occurring in the North Sea.

1.1 Harmful algae blooms (HABs) in Galicia

The Rias Baixas are four large V-shaped coastal embayments located on the south-west coast of Galicia (NW Spain) along the northern boundary of the NW African upwelling system (Figure 1). The *rias* ecosystems are influenced by surface currents, strong tides, freshwater discharges from small rivers and the upwelling-downwelling dynamic on the adjacent continental shelf (Barton *et al*., 2015). Upwelling events associated with northern winds, occurring mainly between May and September, introduce deep, cold, nutrient-rich waters into the *rias* and increase significantly the productivity (Pitcher *et al*., 2010).

As a consequence of the high productivity, the area is rich in fish and shellfish resources and supports an intensive mussel culture using floating rafts (or *bateas*) organized in farming polygons. In fact, Galicia is the most important producer of aquaculture mussel of Europe and one of the world leaders (Labarta and Fernández-Reiriz, 2019).

HABs are a frequent and well-documented phenomenon in Galicia, where they cause an ecological damage, but also an important social and economic impact since they even force the closure of the mussel farming polygons (González Vilas *et al*., 2014; Rodríguez *et al*., 2011).

The HABs monitoring system in Galicia is based on field samplings in fixed stations and subsequent cell count of toxic species using optical microscopy (Anderson, 2009). Despite of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

direct observations are essential, different approaches incorporating satellite images have also been proposed. As compared to traditional sampling methods, satellite methods are faster, cheaper and produce map outputs providing more information about the spatial distribution of the HABs (Blondeau-Patissier *et al.*, 2014; Kudela *et al.*, 2017).

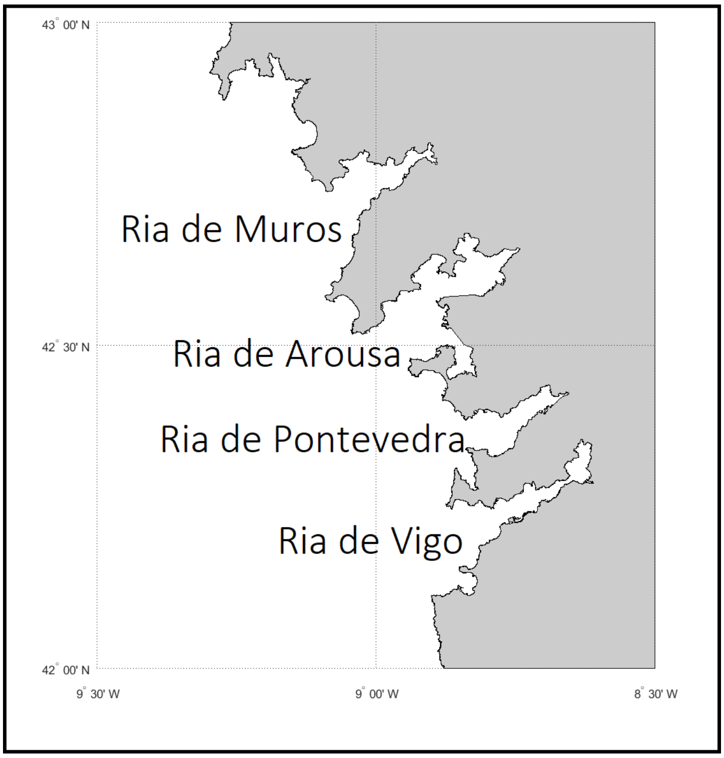


Figure 1 – Map showing the location of the study area, with the names of the four Rias Baixas.

After the successful results obtained using MERIS images between 2002 and 2012 (Spyrakos *et al.*, 2011), CoastObs exploits the use of the new Sentinel satellites to provide useful information for HABs detection and monitoring in Galicia.

Chapter 2 summarizes data and methods involved in the development of the algorithms. Products are presented in Chapter 3 and 4, while Chapter 5 includes a summary of their potential applications, considering strengths and limitations.

1.1.1 *Pseudo-nitzschia* spp.

Pseudo-nitzschia spp. is a diatom genus widely spread all over the oceans (Hasle, 2002). Some *Pseudo-nitzschia* spp. species produce domoic acid (DA), a neurotoxic amino acid which can cause deleterious effects to marine organisms and even humans (amnesic shellfish poisoning, ASP) when it is bio-concentrated via trophic transfer in the food web (Terseleer *et al.*, 2013).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Pseudo-nitzschia spp. is common in upwelling systems as the Iberian System (Gonzalez Vilas *et al*., 2014). In Galicia, HAB events associated with this genus imply not only a damage to the *rias* ecosystems and a risk to human health, but also an economic impact since they can even force the closure of the mussel production areas (Palma *et al*., 2010).

Some authors have already proposed regional-specific species indicators based on satellite data for *Pseudo-nitzschia* spp. For instance, Anderson *et al*. (2009) developed statistical models for *Pseudo-nitzschia* spp. abundance, particulate DA (pDA) and cellular DA (cDA) in Santa Barbara channel incorporating ocean colour (MODIS-Aqua and SeaWiFS) and sea surface temperature (AVHRR) data. In addition to chl*a* concentration and temperature, reflectances at 410 nm, 510 nm, 555 nm and 590 nm were also found to be related to the output parameters and included in the models. In Galicia, Spyrakos (2011) also found a relationship using regression analysis between pDA and MERIS bands at 510 nm, 560 nm and 620 nm.

Despite of the complexity of the Galician waters, *Pseudo-nitzschia* spp. often dominates the phytoplankton community with abundances exceeding of 10^5 cell L^{-1}, allowing for the possibility of being detected from satellite images.

1.1.2 *Alexandrium minutum*

Alexandrium minutum is a toxic dinoflagellate with many strains that produce paralytic shellfish toxins, and hence it implies a risk to ecosystems and human health (paralytic shellfish poisoning, PSP).

Although HABs of *Alexandrium minutum* are not so frequent as *Pseudo-nitzschia* spp. blooms on the Galician coast, it often co-occurs with the non-toxic dinoflagellate *Prorocentrum micans* and the copepod *Acartia clause* in the Ria de Vigo (Frangopoulos *et al*., 2000).

1.2 *Phaeocystis* blooms in the North Sea

Although *Phaeocystis* spp. is not considered to be a harmful algal bloom (HAB) species, its ability to form a mucoidal colony is considered to be an ecological nuisance as it has been found to, inter alia, potentially increase fish mortality, reduce growth and spawning of shell fish and impact coastal tourism during dense blooms due to both the odorous production of dimethylsulfide (DMS) and the accumulation of foam on coastal areas (Davidson and Marchant, 1992; Schoemann *et al*., 2005; Blauw *et al*., 2010). *Phaeocystis* has a polymorphic life cycle, alternating between a flagellate and colonial phase (Blauw *et al*., 2010). Colonies of colonial cells reside within a mucoidal matrix, which are commonly a main cause of foam formation along beaches. In the context of European coastal waters, *Phaeocystis globosa* has been found to particularly affect the Southern Bight of the North Sea (Lancelot *et al*., 2005; Astoreca *et al*., 2009). *Phaeocystis* blooms that have occurred along the Dutch coastal zones have been

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

intensively studied, resulting in the availability of extensive research on the phenology of this phytoplankton group (Teixeira *et al*., 2012; Peperzak and Poelman, 2018; van der Woerd *et al*., 2011 etc.). Of particular concern is the fact that *Phaeocystis* blooms occur along the Dutch coastal waters almost every year, with the uncertainty of how global warming might influence it in the future (Chen and Mynett, 2006).

Prior to the formation of a bloom under certain environmental conditions, *Phaeocystis* exists in the flagellate phase before transitioning to form colonial cells (Peperzak, 2002). The onset of a *Phaeocystis* bloom typically coincides with the depletion of silica – this has been hypothesized to be either associated with *Phaeocystis* being able to outcompete silica-limited diatoms, or the use of diatoms as a solid substrate for the growth of colonial cells (Peperzak, 2002; Riegman and van Boekel, 1996). Other factors like surface irradiance and nutrients also influence the initiation of the bloom. However, measuring the density of *Phaeocystis* cells and/or the growth rate can be a costly method for monitoring the evolution of a bloom on a reasonably high spatio-temporal scale. In 2004, a *Phaeocystis*-dominated bloom event occurred and degraded within the span of 10 days (Texeira *et al*., 2012). Since the initiation of a bloom can be difficult to predict, the duration of the bloom event further restricts the ability to acquire sufficient water samples to monitor the evolution of the event. Remote sensing offers itself as a cost-effective solution that is also able to improve the ability to acquire measurements at a higher spatio-temporal frequency, with the latest Sentinel-3 Ocean Land Colour Imaging (OLCI) spectrometer being able to achieve a spatial resolution of 300m and an average revisit time of 2-3 days. Subsequently, satellite imagery has been utilized for the detection of harmful algal blooms (HAB) by using chlorophyll-a as a proxy, however, this has was shown to be insufficient as a method for specifically detecting *Phaeocystis* blooms. Due to the influence of other algal pigments on the spectral reflectance, chl*a* alone is unable to serve as a reliable indication for the concentration of *Phaeocystis* (Kurekin *et al*., 2014), hence, resulting in the need to develop species-specific algorithms for the detection of these blooms (Astoreca *et al*., 2009).

In view of the need to develop an algorithm that will aid the detection of a *Phaeocystis* bloom within the project requirements of CoastObs Task 3.4, Chapter 6 consists of a literature review of the species to better understand the phenology of the marine phytoplankton and the various environmental variables driving the growth rate and initiation of a *Phaeocystis* bloom. The following chapters (Chapter 7 and Chapter 8) will comprise a review of the existing algorithms available for the detection of *Phaeocystis*, an assessment of the use of various environmental parameters for the modelling the evolution of a bloom and finally, the application of the algorithm onto satellite imagery.

2 Galicia HABs products: Methods

2.1 Datasets

2.1.1 *In situ* datasets

INTECMAR

The Technological Institute for the Control of the Marine Environment of Galicia (INTECMAR) is monitoring routinely the oceanographic conditions, marine biotoxins and HABs, chemical pollution, microbiology and pathology in Galician coastal waters. Its routine monitoring programme consists of a weekly sampling at 41 sampling stations distributed across the four Rias Baixas.

At each station, CTD profiles (i.e., conductivity, temperature and depth) between surface and a meter above bottom are obtained using a Seabird 25 or a Seabird 19plus loggers, which are equipped with a set of coupled sensors: a WETStar fluorimeter, a C-Star transmissometer, a spherical irradiometer LI-193SA, an oxygen sensor SBE 43 and an ultraviolet fluorimeter (UFV) Aquatracka. Measurements of coupled sensors allow the estimation of profiles of different *in situ* parameters, such as chl*a* concentration, transmittance, photosynthetically active radiation (PAR), oxygen saturation and dissolved aromatic hydrocarbons (DHA's).

In addition to *in situ* parameters, integrated water samples are collected at three depth ranges (0 m – 5 m; 5 m – 10 m and 10 m – 15 m) using PVC hoses. Samples are then analysed in the laboratory in order to determine chl*a* and inorganic nutrients concentrations (nitrate, nitrite, ammonium, phosphate and silicate). Specifically, chl*a* concentrations are spectrofluorometrically determined following the method proposed by Zapata *et al*. (2000).

Finally, phytoplankton data are also collected using phytoplankton tow nets (10 µm mesh) from surface to 15 meters depth. Samples are fixed with formaldehyde 4% and stored under dark and cool conditions. Total abundances (in cells L^{-1}) of different species, i.e. *Dinophysis acuminata*, *Dinophysis acuta*, *Gymnodinium catenatum*, *Alexandrium* spp. and *Pseudo-nitzaschia* spp., are counted using an inverted light microscope at 250x and 400x magnification (Utermöhl, 1958).

Dedicated CoastObs field campaigns

In the framework of CoastObs project, a field campaign was carried out in July 2018 in the Ria de Vigo using two research vessels. It consisted of five field trips on the same dates as Sentinel overpasses including a total of 58 sampling stations.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

At each station, water-leaving reflectance spectra were measured using two different sets of field radiometers: TriOS Ramses and Water Insight Spectrometer (WISP-3). Water samples were also collected from surface to a depth of 4 meters using a PVC hose. Finally, a set of parameters were measured *in situ* using different sensors on-board, including depth, temperature, pH, conductivity, turbidity, chl*a* fluorescence, coloured dissolved organic matter (CDOM), IOPs (absorption, backscattering and scattering) and underwater radiometric measurements.

Water samples were filtered in the laboratory to determine several biooptical parameters, including chl*a* concentration, CDOM, size-fractionated chlorophyll and total suspended matter (TSM). Specifically, chl*a* concentration was determined using a high performance liquid chromatography (HPLC) method with a reverse phase C_8, following the procedures described by Zapata *et al*. (2000) for pigment extraction and separation.

Some water samples were fixed with formaldehyde 4%, stored under dark and cool conditions and finally analysed using light microscopy by INTECMAR experts. So, phytoplankton cell counts of different species, including *Alexandrium minutum* and *Pseudo-nitzschia* spp., were obtained.

2.1.2 Sentinel-3 Images

Sentinel-3 mission is based on the heritage of ENVISAT MERIS, which has kept operational between 2002 and 2012 achieving excellent results in marine and coastal applications. Its Ocean and Land Colour Instrument (OLCI) covers a swath width of 1270 km providing the same spatial resolution (300 m) as MERIS, but with more spectral bands (21 instead of 15) ranging from 400 nm to 1020 nm. Since December 2018, a two satellite configuration (Sentinel-3A and Sentinel-3B) is available allowing a revisit time in Galicia of only one day (Donlon *et al*., 2012).

Forty-five Sentinel-3 cloud-free images over the Rias Baixas area acquired between April 2016 and November 2018 were available. Five images were acquired on the same dates as field data were collected during the field campaign conducted in 2018.

2.1.3 Sentinel-2 Images

The Multispectral Instrument (MSI) on-board Sentinel-2 operates in 13 spectral bands from 440 nm to 2220 nm providing a high spatial resolution: four bands at 10 metres, six bands at 20 metres and 3 band at 60 metres. Revisit time in Galicia varies between 3 and 5 days (Drusch *et al*., 2012). Four Sentinel-2cloud-free images over the Rias Baixas acquired in July 2018 were available on the same dates as the field campaign.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

2.2 Image pre-processing

2.2.1 Atmospheric correction

In the framework of CoastObs project, four different atmospheric correction (AC) algorithms were validated: Case–2 Regional processor (C2RCC), Case 2 Regional Coast Colour (aug-C2RCC), OLCI Level-2 Water Full Resolution (WFR) and Polymer. Initial validation was based on the comparison of image reflectance spectra for each AC with water-leaving spectra measured *in situ* using both TriOS and WISP-3 field radiometers.

In Galicia, five Sentinel-3 images and four Sentinel-2images over the Rias Baixas area were available on the same dates as field spectra were measured during the field campaign in July 2018. Initial validation results show that Polymer outperforms other algorithms, and hence it was selected in order to develop and validate both chlorophyll regional algorithms and species indicators.

POLYMER v4.6 was developed from an atmospheric correction processor for clear ocean (case-1) waters that is able to deal with sun glint (Steinmetz *et al.* 2011). Polymer applies a spectral optimization based on a bio-optical model and radiative transfer models to separate atmospheric (including glint) and water reflectance. As compared to alternative methods that extrapolate from near infra-red bands, it uses the full set of wavebands available. Output values are fully normalized water-leaving reflectances.

2.2.2 Masking

Masking is based on the pixel identification and classification tool IdePix, an open processor available in the STEP (Science Toolbox Exploitation Platform) Sentinel-3 and Sentinel-2 toolboxes. Pixels flagged as invalid, cloud (i.e. cloud_sure, cloud_buffer, cloud_shadow, cirrus_sure, cirrus_ambiguous), land or vegrisk are masked, while the remaining pixels are considered as water and hence included in further analyses.

2.2.3 FCM Clustering

Reflectance spectra extracted from atmospherically corrected Sentinel-3 images over the Galician Case 2 waters are affected by different optically active constituents. Hence, the use of a single chl*a* algorithm could result in unreliable chl*a* estimations if it is applied to data from different situations, for instance, clean waters and sediment-dominant waters after a storm period. Since concentrations of the different water constituents are not known *a priori*, several authors have proposed the application of clustering algorithms to reflectance data to establish clusters that could be related to different water types (Neil et al., 2019).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

We applied Fuzzy C-Means (FCM) algorithms to Sentinel-3 reflectance spectra linked to *in situ* measurements, as proposed by Gonzalez Vilas *et al*. (2014) using MERIS data. FCM results provide a framework for chl*a* retrieval algorithms, so that a different NN could be developed for each cluster. Note that although resulting clusters could be related to different optical water types, limited information about water constituents (mainly TSM and CDOM) prevented us from drawing relevant conclusions at this stage.

The FCM technique divides a dataset into a specified number of clusters assigning to each data point (i.e. reflectance spectra) a membership degree (ranging from 0 to 1) for each cluster. Therefore, unlike hard classifiers as k-means, each data point can potentially belong to more than one cluster. The method iteratively adjusts centroids and degrees of membership to optimize an objective function until the previously established optimization criteria are met. The resulting FCM algorithm is intended to minimize distances between data points and cluster centres (Moore *et al*., 2009).

FCM algorithm was developed using Matlab software selecting default optimization criteria: a maximum number of iterations of 100 and a minimum improvement in the objective function between two consecutive iterations of 1e-5. Hence, the algorithm only requires two input parameters: the number of clusters (*c*) and the weight exponent (*m*), which can be any real number greater than 1.

Since the number of expected clusters was unknown *a priori* because of the lack of *in situ* data to define optical water types, optimal values of *c* and *m* were established using a grid search procedure (Gonzalez Vilas *et al*., 2011), i.e. data were clustered several times varying *c* from 2 to 8 and *m* from 1.1 to 3, and the optimal combination was selected. Clustering results for each combination were evaluated using two functions: 1) the partition coefficient (*F*), a measurement of the overlap between clusters ranging from 0 (overlap between clusters, weak clustering) to 1 (without overlap between clusters, strong clustering); and 2) the compactness and separation index (*S*), ratio between the compactness (measuring the variance between clusters) and the separation (minimum distance between cluster centres). Hence, the optimal combination is expected to have a high *F* value (no overlap) and a small *S* value (compact and well separated clusters).

Once the best FCM algorithm was obtained, it can be applied to the training dataset or other independent dataset in order to assign a cluster value to each data point according to their highest membership degree.

Similarly, classification images can be obtained by assigning a cluster value to each valid (open water) pixel. If a single cluster-specific algorithm is applied to build a map, these images are useful for masking pixels belonging to other clusters. And if a different algorithm is available for

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

each cluster, membership grades for each cluster would also allow blending their results to create maps with soft transitions.

2.3 Generation of match-up databases

Retrieval algorithms are mainly based on machine learning methods requiring *a priori* information about the real output (chlorophyll *a* concentration or species indicator) corresponding to a set of input data (reflectance spectra and/or geometry values). Therefore, generation of a dataset of valid match-up data points associating Sentinel images and *in situ* data is needed for the development and validation of these algorithms.

Data from available Sentinel-2 and Sentinel-3 images were extracted and linked to the *in situ* databases. Due to the great temporal variability in the Galician *rias*, only data points extracted from images acquired on the same date as *in situ* data were considered as valid.

For each sampling point, the number of valid pixels in a 3x3 window centred at the *in situ* station location is first extracted. Valid pixels are open water pixels, i.e. pixels which were not masked (flagged as invalid, cloud, land or vegrisk) using the IdePix tool. Points are not included in the match-up database if the central pixel is not valid.

For Sentinel-3, the central pixel is used as match-up value to extract reflectance and geometry values, i.e. sun zenith, view zenith and difference between view and sun azimuths.

For Sentinel-2, the median of valid pixels (between 1 and 9) in the 3x3 window was computed in order to reduce the instrument noise. The median is preferred to the mean to reduce the effect of mixed and non-masked pixels with extremely high or low values.

In both cases, the number of valid pixels in the 3x3 window is considered as a quality indicator, ranging from 9 (highest quality) to 1 (lowest quality). Note that low quality values indicate that the sampling station is located near coast or cloud or foggy areas, so that reflectance values could be affected.

Moreover, the variation coefficient, ratio of the standard deviation to the mean, was also computed using only valid pixels (between 1 and 9) as a measure of relative variability of each band (reflectance or geometry).

More details about the match-up extraction process will be included in the validation report (D3.10) due to October 2018.

2.4 Retrieval algorithms

2.4.1 Neural Networks (NN)

Multilayer perpectron (MLP) neural networks (NN) have been widely used for retrieving the concentrations of water constituents on coastal waters from remote sensing data. MLP provide important advantages as compared to other statistical approaches, such as the ability to model multivariate, complex and nonlinear data without making any assumptions about data distribution (Lary *et al*., 2016). We have used NN for the retrieval of chlorophyll *a* concentration.

A MLP is a feedforward NN aimed at approximating a set of input data (reflectance spectra and geometry values) to the corresponding output (chl*a* concentration). It consists of a set of computational elements, called neurons or nodes, arranged in multiple layers and interconnected in a feedforward way: each node in a layer is only connected to the nodes of the immediately next layer, but has not connections to nodes in the same or previous layers (Haykin *et al*., 2009).

A typical MLP architecture is comprised of an input layer, one or more hidden layers and an output layer. The input layer only distributes the input variables into the network, without processing them. The nodes in the hidden and outputs layers transform their input signal by applying a bias and an activation function. Moreover, each connection is defined by a weight value.

The development of a MLP network require three phases: 1) design, 2) training and 3) validation.

Design

The design of a MLP implies to define the architecture of the algorithm: inputs and output, number of hidden layers (0, 1 or 2), number of nodes in each hidden layer and the activation functions to be applied. Note that the number of nodes of the input layer is equal to the number of variables in the input dataset and the output layer only has one node corresponding with the desired output. Non-linear activation functions (sigmoid or tangential) are preferred for the hidden layers, while the linear or identity functions are usually established in the output node.

Training

Once the MLP architecture is designed, the relationship between the input and the output depends on the bias and weight values. These values are established in the training phase by a supervised learning technique, which uses *a priori* information about the actual output corresponding to a set of input data. Biases and weights are iteratively adjusted to minimize an error function until the best approximation to the actual output is achieved, i.e. no more significant variations in the overall error are observed. The adjustment is based on a back-

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

propagation learning procedure: starts in the output layer, continues layer-by-layer in a looping pattern and ends in the input layer. Weights are adapted using different nonlinear optimization methods.

Validation

In the validation phase, the performance of the trained MLP is evaluated using a set of parameters which compares the observed output (*O*) and the obtained one using the algorithm (*M*):

- *Coefficient of determination* (R^2) between *O* and *M*

- *Mean prediction error (MPE):*

$$MPE = \frac{1}{N}\sum_{i=1}^{N} PE_i$$

$$PE_i = O_i - M_i$$

- *Variance of the predictor errors (VAR):*

$$VAR = \frac{1}{N-1}\sum_{i=1}^{N}(PE_i - MPE)^2$$

- *Root mean square error (RMSE):*

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} PE_i^2}{N}}$$

- *Relative RMSE:*

$$Rel.RMSE = \sqrt{\frac{RMSE}{\frac{1}{N}\sum_{i=1}^{N} ChlO_i}}100$$

R^2 measures the correlation between the observed and the predicted datasets. *MPE* is useful for determining if a NN tend to underestimate (high positive values) or overestimate (high negative values). *RSME* and *Rel. RMSE* are measurements of absolute and relative error, respectively, and *VAR* quantifies error variability.

Training strategy

Although the same dataset could be used in the training and validation phases, the general practice to avoid overfitting and ensure a greater generalization capability is divide the complete input dataset into three subsets: a training subset, a validation subset and a test subset.

The training and validation subsets are used in the training phase. The training subset is used for updating the weights and biases, while the validation subset is only used for computing an additional error measurement, which is monitored during the complete process of learning. Both validation and validation errors usually decreases during the initial phase of learning, but when the network begins to overfit the data, the validation error typically begins to rise. In order to avoid overfitting, the network weights and biases are saved at the minimum of the validation error.

The test subset is only used in the validation phase. Since it has not included in the training procedure, performance measurements computed from this independent dataset are useful for gaining insight into the generalization capability of the algorithm. Note that parameters from test subsets are expected to be worse than those attained using the training and validation subsets.

The best NN for a given problem (e.g. retrieval of chl*a* concentration) was selected in a two-step procedure.

In the first stage, the optimal design was selected using a trial-error procedure. Different NN models were trained and validated varying the input parameters, such as number of hidden layers and neurons or activation functions, choosing the best configuration according to the validation parameters computed for the train, validation and test subsets, and according to the following criteria: maximum R^2 and minimum *RMSE*.

The second stage deal with the problem of the variability of the results between different runs, since results depend on the initial value of the weights, which are randomly established. So, we performed 10,000 runs using NN with the optimal configuration and selected the best model using the same criteria as in the previous stage.

NN were trained using the Matlab implementation of Levenberg-Marquardt backpropagation method (Hagan *et al*., 1994), which has been reported to be the fastest algorithm for training moderate-sized feedforward neural networks (up to several hundred weights). Mean sum of square errors (*mse*) was selected as performance function, and other training parameters were set to default values: maximum number of epochs to train (1000), performance goal (0), initial mu (0.01), minimum performance gradient (1e-7) and maximum validation failures (6).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Input and output data were scaled between 0 and 1 using minimum and maximum values in order to improve the stability and performance of the NN models. Moreover, minimum and maximum values are useful for defining the application scope, so that pixels with input or output values out of range are excluded in the map generation process.

2.4.2 Support Vector Machines (SVM) for HABs detection

SVM are supervised learning algorithms that are usually used as binary (2-class) classifiers, although approaches for one-class classification, multi-class classification and regression have also been implemented. SVM are a robust technique with a strong theoretical basis, providing a better generalization capability and a lower computational overload than other classifiers. As neural networks, SVM do not make any assumption about data distribution and can model complex and nonlinear data. We have applied the SVM approach for 2-class classification to develop species indicators based on presence/absence and bloom/no bloom models for *Pseudo-nitzschia* spp. and *Alexandrium minutum*.

SVM binary classifier is based on the linear classifier, which uses a simple hyperplane to separate two classes, but operating in a feature space with a higher dimension than the input space. It is founded on the fact that linear separability is increased in this feature space, according to Cover's Theorem (Cover, 1965).

The linear classifier searches for the optimal hyperplane by maximizing the margin, i.e. the distance between the separating surface and the closest training data point of each class. Margin is maximized using the Lagrange method to solve a quadratic optimization problem constrained by linear restrictions that are satisfied when a perfect classification is achieved (Cortes and Vapnik, 1995). With the aim of avoiding overfitting if data are not perfectly separable, the so-called stack variables are introduced to relax linear restrictions and allow some classification errors, searching for a balance between maximizing the margin and minimizing the overall error. The cost parameter (C) controls the penalty for misclassification (Cristianini and Shawe-Taylor, 2000).

Computations in the high dimensional feature space are avoided using kernel functions, according to the Mercer's theorem (Mercer, 1909). Typical kernels are the linear function, the polynomial function and the radial basis function (RBF). We chose the RBF kernel because it has been reported to perform slightly better for datasets with a similar size (around 100-200 data points) and it only requires two parameters: gamma (γ) and the cost parameter (C), reducing the complexity of the model selection process as compared to other kernels (Camps-Valls and Bruzzone, 2005).

Binary SVM models were developed using the JAVA version of LIBSVM library, which implements a sequential minimal optimization type algorithm (Chang and Lin, 2011) to solve the constrained optimization problem expressed in terms of Lagrange multipliers using the kernel trick, i.e. to train the SVM model. In addition to the classical approach with a binary output (+1 or -1), LIBSVM also implements a probability output, i.e. probability estimates (between 0 and 1) are computed for each class.

Input data were linearly scaled between 0 and +1 before training the models in order to avoid a greater effect on the results of features with larger numeric ranges (Sarle, 1994).

The cost parameter (*C*) was weighted using a different weight for each class in order to deal with the imbalance problem, which may cause poorer results because of the classification bias towards the majority class. Hence, the accuracy of the minority class (i.e. with fewer training data points) is improved by applying a larger weight, at the cost of a possible increase of misclassification for the majority class. In practice, the percentage of data points in each class was set as weight of the other class.

Model evaluation

The models performance was evaluated using a set of measures that are commonly applied to binary classification problems (González Vilas *et al.*, 2014). All these measures are derived from the confusion matrix, a table with two rows and two columns which compares the model output with the observed output and reports the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) (Table 1).

Table 1 – Confusion matrix comparing the model output with the observed output.

		Observed	
		Class +1	Class -1
Model	Class +1	*TP*	*FP*
	Class -1	*FN*	*TN*

- *Overall accuracy (OA):* Percentage of data points that are correctly classified

$$OA=(TP+TN)/(TP+TN+FN+FP)$$

- *True positive rate (TPR)*: Individual accuracy for class +1

$$TPR= TP/(TP+FN)$$

- *True negative rate (TNR)*: Individual accuracy for class -1

$$TNR= TN/(TN+FP)$$

- *False negative rate (FNR)*:

$$FNR= FN/(TP+FN)$$

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

- *False positive rate (FPR)*:

$$FPR= FP/(TN+FP)$$

- *Kappa (κ):* It assesses the classification agreement removing the chance effect. It varies from 0 to 1.

$$\kappa= (OA-CA)/(1-CA)$$

$$CA= ((TP\cdot TN) + (FP\cdot FN))/(TP+TN+FP+FN)^2$$

As suggested by Cohen (Cohen, 1960), kappa value is interpreted as follows: values lower than 0 indicates no agreement; 0.01 to 0.20 as none to slight; 0.21 to 0.40 as fair; 0.41 to 0.60 as moderate; 0.61 to 0.80 as substantial and 0.81 to 1.00 as almost perfect agreement.

If the probability output (between 0 and 1) is used instead of the binary output (+1 or -1), it is necessary to select a threshold to convert probability values into a binary result by assigning +1 if probability is above this threshold or -1 otherwise. Hence, results of the matrix confusion and derived measures depend on the selected threshold.

In order to deal with this problem, the Area Under the Receiver Operating Characteristic (ROC) Curve (*AUC*) was also computed when probability outputs were available. The ROC curve plots *TPR* against *FPR* computed at different probability thresholds. Hence, *AUC* is a measure independent of the threshold that is useful for comparing different models. *AUC* values greater than 0.9 are considered excellent, from 0.8 to 0.9 very good, from 0.7 to 0.8 good, from 0.6 to 0.7 average and lower than 0.6 poor (Hosmer and Lemeshow, 2000).

Model selection

Although RBF SVM models only require two parameters (*C* and *γ*), their optimal values for a given problem are unknown a priori. The selection of the optimal parametric configuration was based on a simple grid-search approach, so that different models were trained and validated using different values of *C* and *γ* (but keeping the same scaling and weight values), and the model with the best performance was finally selected. The process was implemented in two consecutive phases in order to save computing time: a first search using a coarse grid with exponentially growing values ($C = 2^{-5}, 2^{-3}, \ldots, 2^{15}$; $\gamma = 2^{-15}, 2^{-13}, \ldots, 2^{3}$) and a second one using a finer grid with parameters varying linearly around the values obtained from the first search (Gonzalez Vilas *et al*., 2014).

Since using the same dataset for training and validation could lead to overfitting, the best model was selected according to performance measures computed from the results obtained in a cross-validation process. Cross-validation is a standard method that divides the training set into k subsets and then the model is trained k times using (k-1) subsets while the remaining one is retained for validation. Specifically, we applied leave-one-out cross validation, so that the model is trained N times using N-1 data points (being N the number of elements of the training set) and the remaining data point is retained for validation.

The parametric configuration of the SVM model resulting in the highest kappa (with binary output) or *AUC* (with probability output) computed for the leave-one-out cross-validation output dataset was selected as the optimal one in the grid search procedure.

Once the optimal values of C and γ are selected, the final SVM model is trained using the complete training dataset. This model can be used for validation from an independent dataset or generation of map products.

3 HABs products for Galicia case study: Regional chlorophyll *a* algorithms

3.1 Theoretical basis

Chl*a* concentration is one of the most important basic products derived from optical satellites. In fact, it has been processed in an operational way for open ocean areas since the 1980's using empirical algorithms based on ratios (e.g. between blue and green bands) or combinations of different spectral bands. Despite of empirical algorithms provide reliable results in Case 1 waters (open ocean), the strong absorption in the blue wavelength region in typical Case 2 waters (i.e. water with high concentrations of other water constituents in addition to chl*a*) prevent an accurate retrieval of chl*a* concentration (Gregg and Roussseaux, 2014).

Therefore, other approaches have been proposed for the retrieval of chl*a* concentration in Case 2 waters, including semi-analytical algorithms based on the properties of the reflectance peak near 700 nm; or neural networks (NN), a supervised learning technique that uses *in situ* data to train the algorithms. For instance, NN were successfully applied in the MERIS Case-2-Regional Processor (C2R) (Doerffer and Schiller, 2008).

Retrieval of chlorophyll a (chl*a*) concentrations in the optically complex waters of the Galician *rias* using optical satellite images has proven to be a challenging task. Optical properties of the water are the result of rapid changes in the temporal and spatial distribution of phytoplankton abundance and composition related to regional characteristics, as upwelling and freshwater inputs from small rivers. Moreover, the relatively low chl*a* concentrations in the area (between 0 and 10 mg m-3, mainly lower than 3 mg m^{-3}) hinders the reliability of most of Case 2 algorithms, which usually cover a much wider range of chl*a* concentration. As a consequence, development of specific regional algorithms was required in order to obtain reliable chl*a* maps from ENVISAT MERIS (Gonzalez Vilas *et al*., 2011), and is also proposed for Sentinel-3.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

3.2 Algorihtm development

Since Sentinel-3 OLCI is based on the heritage of MERIS, development of chl*a* algorithms was based on the same methodology proposed by Gonzalez Vilas *et al.*, 2011.

This approach consists of three steps: 1) atmospheric correction; 2) definition of the scope of the algorithms using masking and FCM clustering techniques; and 3) application of NN algorithms for the retrieval of the chl*a* concentration.

FCM Clustering

FCM clustering algorithm was developed using a match-up database linking Sentinel-3 reflectance spectra with INTECMAR sampling stations (see section 2.3). Note that Polymer atmospheric correction was first applied to all the images (see section 2.2.1).

The match-up database was filtered to include only data points with a quality indicator of 9 (9 valid pixels in the 3x3 window, see section 2.2.3). Moreover, data points with negative reflectance values in the 3x3 window and with high variation coefficients for a given band (higher than percentile 95) were also removed. The final database consisted of 457 data points derived from 36 images acquired between April 2016 and November 2018.

Results of the grid search procedure to select the optimal number of clusters (*c*) and weight exponent (*m*) are shown in Table 2 (see section 2.2.3)

Table 2 – Summary of clustering results applying FCM algorithm. Partition coefficient (F) and compactness and separation index (S) were computed for a range of conditions (m, weighting exponent; c, number of clusters).

	m									
c	1.1		1.5		2		2.5		3	
	F	*S*	*F*	*S*	*F*	*S*	*F*	*S*	*F*	*S*
2	**0.99**	**0.17**	0.50	1864	0.79	0.14	0.70	0.15	0.64	0.18
3	0.98	0.19	0.88	0.16	0.70	0.16	0.56	0.19	0.48	0.26
4	0.97	0.27	0.83	0.21	0.62	0.18	0.46	0.25	0.38	0.32
5	0.97	0.31	0.80	0.25	0.57	0.24	0.40	0.27	0.31	0.44
6	0.96	0.66	0.79	0.22	0.52	0.22	0.35	0.42	0.27	1.68
7	0.96	0.46	0.76	0.33	0.49	0.31	0.31	0.57	0.22	10.77
8	0.96	0.39	0.73	0.42	0.45	0.46	0.28	0.88	0.20	3.33

The best FCM algorithm was defined with two clusters and a weighting exponent of 1.1 ($F = 0.99$; $S = 0.17$). A cluster was assigned to each one of the 457 data points according to their highest membership degree, resulting in a 40% ($n = 185$) of data points belonging to cluster#1 and a 60% ($n = 272$) to cluster#2 (Table 2). Selection of the corresponding cluster was unequivocal (membership degree for that cluster greater than 0.99) in a 93% of the data points.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Table 3 summarizes basic statistical information about geometry and chl*a* concentration associated to each cluster. These parameters were computed using only data points in which reliable chl*a* information was available (*n = 373*). Both chl*a* concentrations and geometry values are overlapped in both clusters. In fact, according to the results from t- test, there are not significant differences in geometry and chl*a* between both clusters.

Table 3 – Number of data points belonging to each cluster without (nTot) and with (nChla) chla concentration data. Geometry and chla statistics (mean ± standard deviation and range) are shown for each cluster.

Cluster	nTot	nChl*a*	Sun Zenith (º)	View Zenith(º)	Azimuth Diff.	chl*a* (mg m^{-3})
#1	185	140	53.52±15.08 26.34-71.16	36.83±13.58 4.37 - 54.46	91.80±87.15 17.38 - 237.43	1.96±1.80 0.04-9.22
#2	272	233	39.92±10.91 24.26-69.72	26.08±17.79 0.69 - 54.46	74.84±80.29 17.02 - 237.44	2.32±2.56 0.04-9.66

Therefore, differences between both clusters could be related to other water constituents, i.e. TSM and CDOM. Unfortunately, the lack of TSM and CDOM information prevent a complete characterization of possible water types associated to these clusters.

However, a valuable clue can be found in the mean reflectance spectra shown in Figure 2. Cluster#1 shows higher values than cluster#2 at lower wavelengths (between 400 nm and 560 nm), which could be related to more turbid waters supporting a higher sediment load.

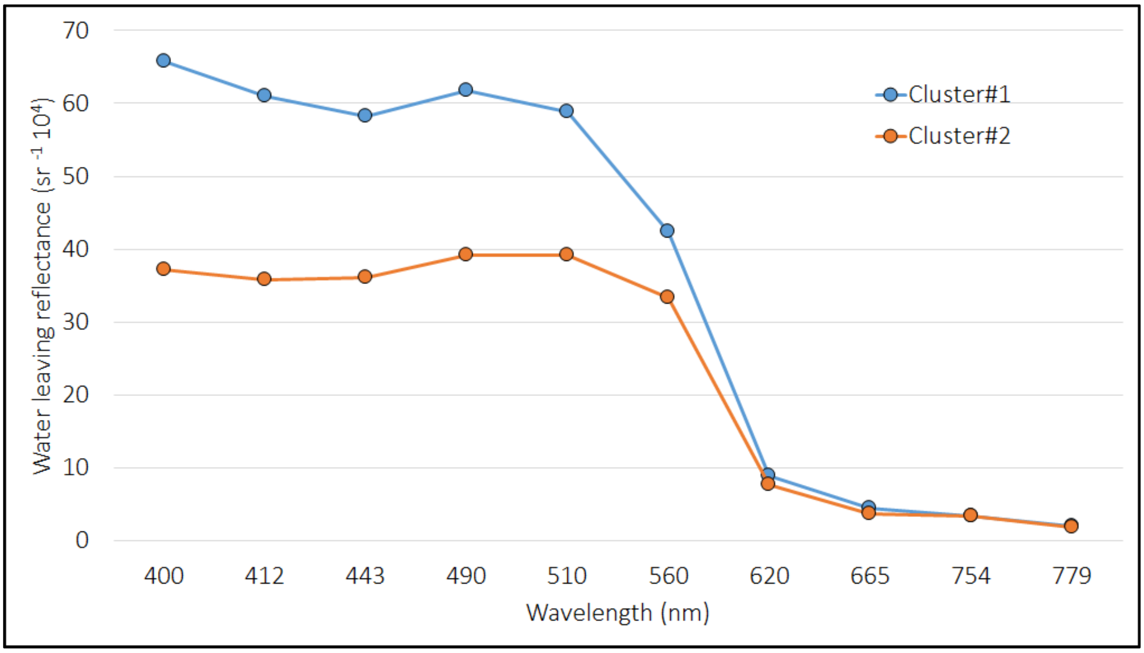


Figure 2 – Mean reflectance spectra for each cluster.

Classification images were obtained from the 36 images included in the match-up extraction. Both clusters are present in most of the images, although cluster#1 is dominant (more than 90% of pixels of the image) in six images and cluster#2 in four images.

NN Algorithms

The retrieval of chl*a* concentration is based on MLP NN (see section 2.4.1). FCM results defined the scope of the NN, so that a different algorithm could be applied to each cluster and results could be merged to obtain a final chl*a* map.

Note that NN are supervised algorithms and hence require *in situ* chl*a* data paired with Sentinel-3 images. NN were developed using the same match-up database as the FCM algorithm, but including only data points with chl*a* concentrations measured *in situ* by INTECMAR in its monitoring program. After applying filtering using the quality indicator and the variation coefficient, 373 data points were available (140 assigned to cluster#1 and 233 to cluster#2). Data derived from 35 images between April 2016 and November 2018.

Both clusters cover the complete range of variation observed in the dataset (from 0.04 mg m^{-3} to 9.66 mg m^{-3}), with average values around 2 mg m^{-3} (Table 3). According to the typical chl*a* pattern recorded in the Rias Baixas area (Nogueira *et al*., 1997, González Vilas *et al*., 2011), chl*a* concentrations tend to be lower than 1 mg m^{-3} (mainly in winter) and rise up to maximums of 8 mg m^{-3} during upwelling events (especially in spring and autumn). In our dataset, a similar pattern is observed, with a 45.6% of the data points showing concentrations lower than 1 mg m^{-3}. Concentrations greater than 8 mg m^{-3} recorded in 8 data points (2.1 % of the total) were due to exceptional situations, such as the *bloom* of *Alexandrium minutum* observed in summer 2018.

A different NN model was developed for each cluster: NNRB-Cl#1 and NNRB-Cl#2. Input and output datasets were first linearly scaled between 0 and 1 using gain and offset values computed from minimum and maximum values shown in Table 4.

The basic architecture was selected in a trial and error procedure (see section 2.4.1). Both NN models consist of an input layer, two hidden layers and an output layer. The input layer includes 13 input nodes: 10 reflectance values corresponding with Polymer bands ranging from 400 nm to 579 nm and three geometry values: sun zenith, view zenith and difference between sun and view azimuth (Table 4). The output layer has a unique node associated to the chl*a* concentration. In both models, hyperbolic tangent function was selected as activation function for each hidden layer, while a simple identity function was applied to the output node. In terms of design, the main difference between both algorithms is the number of nodes in the hidden layers: 7 and 4 in NNRB-Cl#1; 5 and 10 in NNRB-Cl#2.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Table 4 – Input and output variables included in each model, showing the minimum and maximum values used for scaling. Reflectances values are multiplied by 10^4.

Variable	NNRB-Cl#1		NNRB-Cl#2	
	Minimum	Maximum	Minimum	Maximum
Rw400	151.20	320.12	41.93	185.30
Rw412	147.75	294.09	44.22	167.90
Rw443	140.81	283.48	47.89	163.30
Rw490	153.17	328.78	59.17	163.31
Rw510	136.83	340.44	60.91	171.37
Rw560	80.51	307.14	63.66	171.17
Rw620	10.67	81.33	9.11	51.92
Rw665	2.14	50.73	1.43	30.82
Rw754	3.27	24.05	4.24	26.63
Rw779	0.54	13.15	1.09	15.89
Sun Zenith	26.34	71.16	24.26	69.72
View Zenith	26.34	54.46	0.69	54.46
Azimuth Difference	26.34	237.43	17.02	237.44
Chl*a* concentration	0.04	9.22	0.04	9.66

Each cluster-specific NN model was developed using only data points with an unequivocal assignation to the corresponding cluster, i.e. with a membership grade higher than 0.99. Hence, mixed pixels were excluded.

Note that the complete dataset was divided into three subsets (training, validation and test). Table 5 shows a summary of the performance measures computed from these subsets (see section 2.4.1 for details).

Table 5 – Summary of the performance measures computed from the training, validation and test datasets using the NN models developed in the project.

Model	Dataset	*N*	R^2	*MPE*	*VAR*	*RMSE*	*RMSE%*
NNRB-Cl#1	Training	97	0.93	0.00	0.01	0.10	21.1
	Validation	19	0.40	-0.01	0.08	0.27	34.1
	Test	13	0.49	-0.05	0.20	0.44	43.3
NNRB-Cl#2	Training	161	0.97	0.00	0.01	0.10	20.5
	Validation	32	0.65	-0.14	0.13	0.38	41.5
	Test	22	0.78	-0.21	0.13	0.40	42.9

Parameters from the validation and test subsets are excepted to be worse than those obtained using the training set, since data were not included in the learning procedure. Good results in the test subset are indicative of a good generalization capability.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

NNRB-Cl#1 shows very good results in the training subset. According to the test subset, it shows a linear trend between the observed and predicted values but with a low R^2 (0.49), an absolute error (*RMSE*) lower than 0.5 mg m^{-3} and a tendency to underestimate the chl*a* concentrations (*MPE* = -0.05).

NNRB-Cl#2 outperforms NNRB-Cl1# providing more reliable results, with a smaller difference between the training and test subsets evidencing a better generalization capability. A clear linear trend is observed between the observed and predicted values (R^2 = 0.78 in the test set). According to the test result, the model also tends to underestimate values (negative *MPE* values) and shows similar absolute and relative errors (lower than 0.5 mg m^{-3} and 45%), but a lower error variability.

Despite of the promising results, both algorithms could be improved by a further validation. With this aim, a field campaign will be conducted in June 2019 in the Ria de Vigo within CoastObs. Validation results will be included in the deliverable D3.10 (Validation report) due in October 2019.

3.3 Generation of chl*a* maps

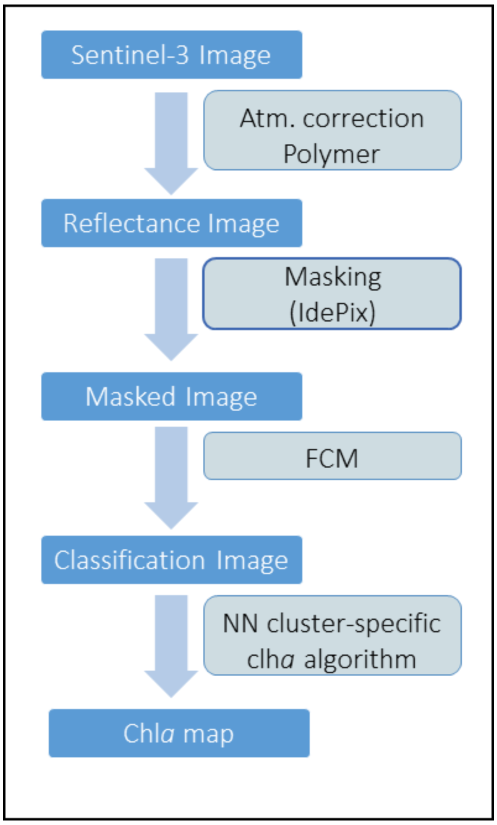


Figure 3 – Chain processing for generating a chla map starting from a Sentinel-3 image.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Figure 3 summarizes the chain processing for the generation of chl*a* maps from Sentinel-3 images.

FCM Intermediate Products

Two intermediate map products derived from the 2-cluster FCM algorithm are included in the processing chain for the generation of chl*a* maps. Both products are built consecutively:

1) Grade maps for each cluster showing the membership degree to that cluster for each non-masked (sea) pixel.

2) Classification image showing the cluster for each non-masked (sea) pixel, which is assigned as the cluster with the maximum membership degree in the grade maps.

Figure 4 shows an example of grade image showing the membership degree computed for cluster#2 and the resulting classification image.

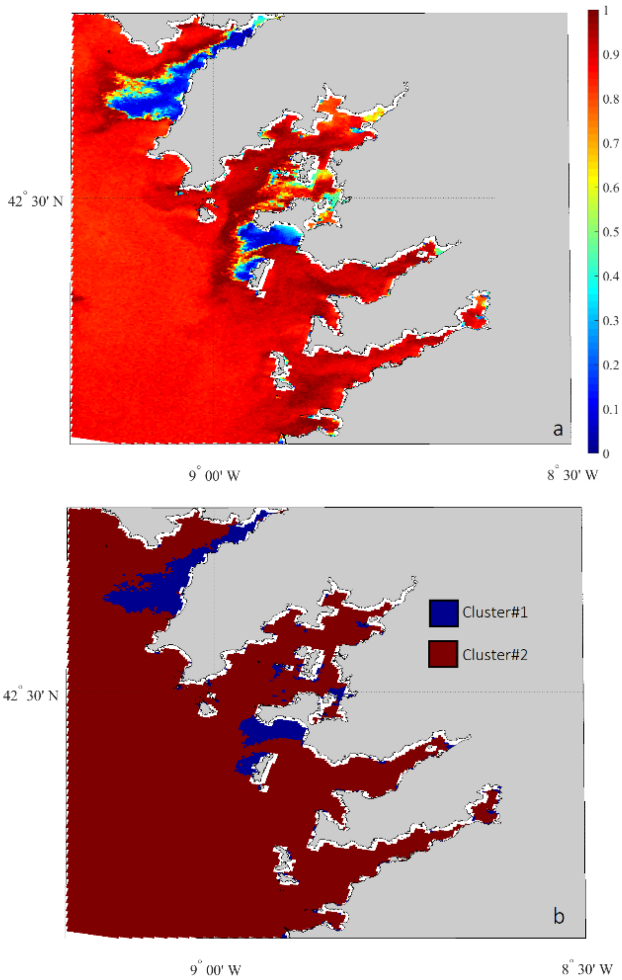


Figure 4 – FCM results derived from a Sentinel-3 image on 19 June 2018. a) Grade image showing the membership degree to cluster#2. b) Classification image showing the cluster value assigned to each pixel.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Note that red areas with grades near 1 are classified as cluster#2 while blue areas are identified as cluster#1. Grade images are useful for merging chla concentrations derived from different NN cluster-specific algorithms while classification images define the scope of application of a single algorithm.

NN Product

NN cluster-specific algorithms are only applied to non-masked pixels belonging to the corresponding cluster in the classification image.

For a given pixel, the final output value is computed following the next steps:

1) Scale input variables. If the input value for a given variable is out of the range, pixel is masked and the following steps are not performed.
2) Compute values of the nodes in the first hidden layer.
3) Compute value of the nodes in the second hidden layer.
4) Compute value of the output node, i.e. the neural network output.
5) Compute final output (i.e. chl*a* concentration) from the scaled value obtained as neural network output.

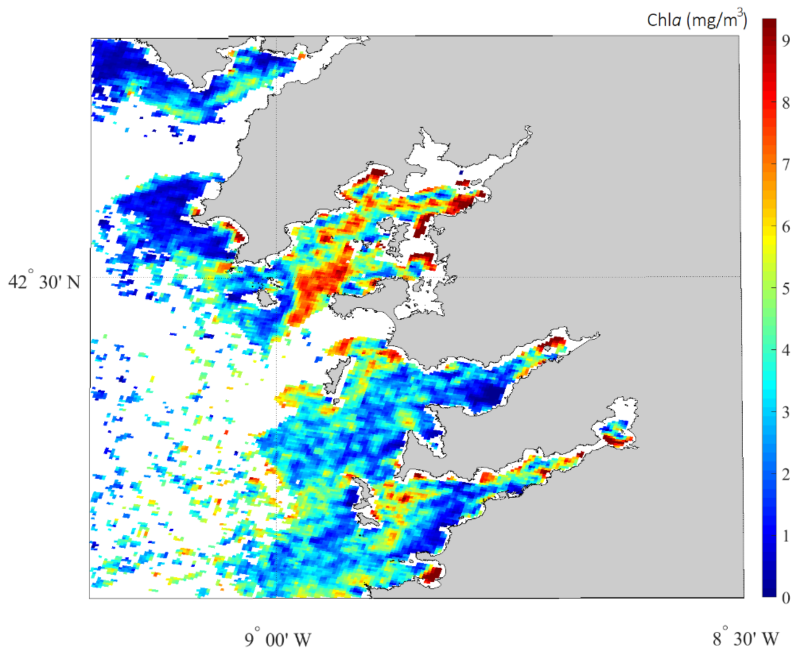


Figure 5 – Chla map on 19 June 2018 derived from the application of NNRB-Cl#2.

Figure 5 shows an example of final chl*a* map on 19 June 2018. Chl*a* concentrations were computed using NNRB-Cl#2, which provides more reliable results. Classification image (Figure 4) was used for defining the application scope, so that pixels belonged to cluster#1 were masked.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

4 Galicia HABs products: Species indicators

4.1 Theorethical basis

Chl*a* concentration is common to almost taxonomic groups and thus it is a good estimator of phytoplankton biomass. Unfortunately, it does not provide information about the species or their toxicity. Species indicators are algorithms aimed at the direct detection of a specific species or taxon from satellite colour images in order to generate abundance, bloom/no bloom or presence/absence maps. These indicators are founded on the fact that massive proliferations of some species may cause a distinctive water colour, and hence show a characteristic spectral signature which could be detected in the images. In addition to the typical limitations of optical remote sensing (e.g. cloud cover), Kudela *et al.* (2007) highlight the main problems associated with the development of these algorithms: 1) images are limited to a discrete number of spectral bands so that distinctive spectral features of the species could be out of their spectral range; 2) spectral signatures, especially on complex coastal waters, are the result of the interaction of different species, not only the target species, and other inorganic and organic components. In order to deal with these drawbacks and gain a deeper insight into the species behaviour, indicators proposed by different authors usually integrate ancillary data.

As with chl*a* concentration, application of regional-specific algorithms being able to capture the complexity of the study area identifying typical situations are expected to provide better results. Therefore, machine learning methods based on a supervised learning using *in situ* data are preferred to empirical or semi-analytical methods.

Within CoastObs, we have developed species indicators for two toxic taxonomic groups causing HABs in Galicia: *Pseudo-nitzschia* spp. and *Alexandrium minutum*. Development of higher-level products based on the integration of ancillary data with these algorithms is planned in a later stage of the project.

4.2 Species indicator for *Pseudo-nitzschia* spp.

4.2.1 Algorithm development

The general approach is based on three steps: 1) atmospheric correction; 2) masking; 3) application of a SVM model for estimating the probability of a bloom of *Pseudo-nitzschia* spp.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Dataset

Only *Pseudo-nitzschia* spp. total abundances provided by INTECMAR were available for the development of the algorithm. Therefore, both non-toxic and toxic species are included in the dataset, making still more difficult to define distinctive spectral features. Unfortunately, data about species identification or toxicity (cDA or pDA) were not available.

A total of 834 data points with *Pseudo-nitzschia* spp. abundance derived from INTECMAR stations sampled on the same dates as Sentinel-3 images were available between May 2016 and November 2018. After removing data points associated with pixels which were masked (mostly cloudy pixels) or showed a quality indicator lower than 9 and/or negative reflectances in the 3x3 extraction window, the final match-up dataset linking reflectance and geometry values with *Pseudo-nitzschia* spp. abundances consisted of 383 data points. This final dataset covers the complete temporal coverage (between May 2016 and November 2018) with data from 34 different images.

Table 6 shows the abundance distribution for the complete dataset and considering clusters derived from FCM algorithm explained in section. Note that cluster datasets are built including only data points with an unequivocal assignation to the corresponding cluster (membership degree greater than 0.99), so that 27 mixed data points were excluded. Threshold for defining bloom and no bloom situations is set in 10^5 cell L^{-1}.

Table 6 – Abundance distribution of Pseudo-nitzschia spp. for the complete, cluster#1 and cluster#2 datasets.

Abundance (cells L^{-1})	Complete Dataset	Cluster#1	Cluster#2
0	118 (30.81 %)	79 (56.03 %)	30 (13.95 %)
10^3-10^4	90 (23.50 %)	35(24.82 %)	48 (22.33 %)
10^4-10^5	108 (28.20 %)	24 (17.02 %)	75 (34.88 %)
10^5-10^6	57 (14.88 %)	3 (2.13 %)	52 (24.19 %)
> 10^6	10 (2.61 %)	0	10 (4.65 %)
No bloom (< 10^5)	316 (82.51 %)	138 (97.87 %)	153 (71.16 %)
Bloom (> 10^5)	67 (17.49 %)	3 (2.13 %)	62 (28.84 %)
Total	383	141	215

Approximately 30% of the data points in the complete dataset show zero abundance, evidencing a total absence or abundances below the detection limit. Anyway, zero values could lead to unreliable results if approximation methods, such as regression or neural networks, are applied to estimate abundances. Hence, development of binary models (presence/absence or bloom/no bloom) are preferred to abundance algorithms.

In terms of bloom/no bloom, the complete dataset is clearly unbalanced, with less than 20% of data points identified as bloom. As a consequence, we opted for SVM methods, which have proved to be a valuable tool for working with unbalanced datasets (see section). In fact, Gonzalez Vilas *et al.* (2014) developed bloom/no bloom SVM models for predicting *Pseudo-*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

nitzschia spp. on the Galician *rias* from a set of environmental variables using a dataset with approximately 15% of blooms.

A remarkable fact is that the cluster#1 dataset only has 3 data points (2.13 %) identified as blooms, so that the development of a cluster-specific bloom/no bloom model is unfeasible.

Therefore, there raised two options: 1) development of a bloom/no bloom model using the complete dataset; and 2) development of a bloom/no bloom model specific for cluster#2, considering all the data points belonging to cluster#1 as "no bloom"

Both options were tested, but the bloom/no bloom SVM model developed using the complete dataset has proven to be more robust and has provided more reliable results.

SVM bloom/no bloom model

Table 7 shows the input variables included in the bloom/no bloom model: 10 reflectance (Polymer) values between 400 nm and 779 nm and three geometry values (sun zenith, view zenith and difference between sun and view azimuths). Input data were scaled between 0 and 1 using the minimum and maximum values also shown in Table. These ranges define the application scope of the SVM model.

The best bloom/no bloom model was obtained as explained in section, applying first a leave-one-out cross validation for selecting the optimal parameters and then training the complete training dataset with this optimal parametric configuration to obtain the final SVM model. Probability output was preferred to binary output since it provides more information.

Table 7 shows the performance measures computed from the leave-one-out cross-validation and the application of the final SVM model to the complete training dataset. As we work with probability outputs (between 0 and 1), a threshold needs to be set in order to compute performance measures derived from the confusion matrix (*OA, TPR, TNR, FNR, FPR*, κ, see section). Results shown in the table were computed using the optimal threshold, i.e., threshold maximizing the sum *TPR + TNR.*

Table 7 – Performance measures computed from the leave-one-out cross-validation process and from the training dataset using the final SVM bloom model.

	Leave-one-out cross-validation	Training dataset (Final SVM Model)
OA	0.90	0.95
TPR	0.90	0.93
TNR	0.90	0.96
FNR	0.10	0.07
FPR	0.10	0.04
κ	0.87	0.94
AUC	0.95	0.97
Optimal threshold	0.50	0.50

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Results from cross-validation evidence a robust model, with the same individual accuracy for both classes despite of the unbalance in the dataset, an excellent *AUC* value and an almost perfect agreement (κ = 0.87) (see section for *AUC* and κ interpretation). Moreover, it shows a good generalization capability, with similar results from both cross-validation and training datasets. The model is able to predict more than 90% of bloom or no bloom situations correctly, with a false alarm percentage lower than 5%.

4.2.2 Generation of maps products

Figure 6 summarizes the chain processing for the generation of a bloom probability map or a bloom/no bloom map starting from a Sentinel-3 image.

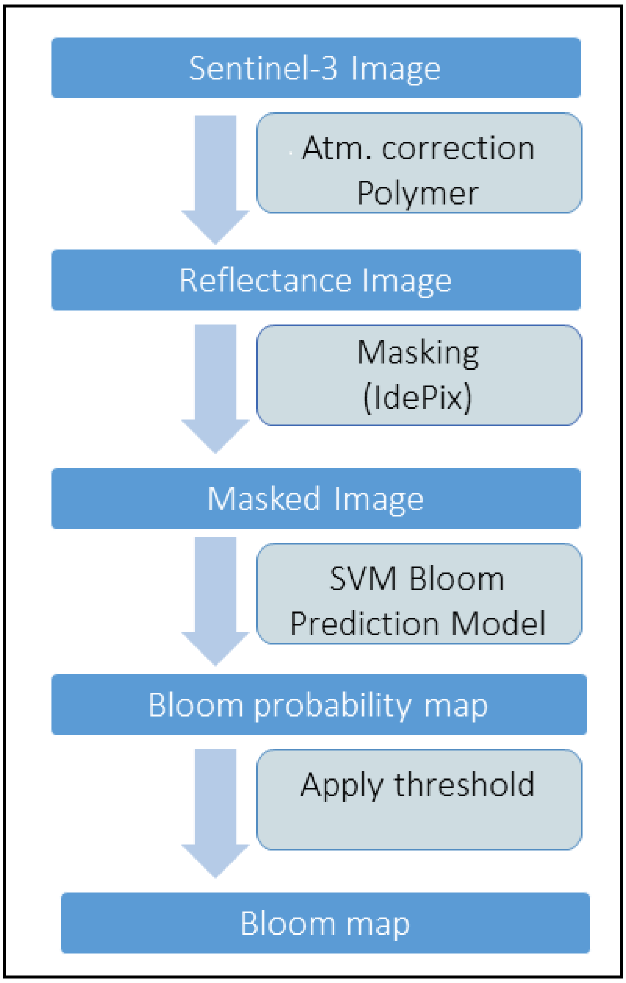


Figure 6 – Chain processing for generating a Pseudo-nitzschia spp. bloom/no blooom map starting from a Sentinel3 image.

For a given non-masked pixel, the application of the SVM bloom prediction model consists of the next steps:

1) Scale input variables
2) Compute kernel values for each support vector by applying the RBF kernel function

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

3) Compute decision values for each class (bloom or no bloom)
4) Compute the pairwise probability matrix
5) Compute the probability estimation for each class. Probability for bloom class is used for building bloom probability maps.
6) Apply a threshold of 0.5 to decide if the given pixel is classified as bloom (probability value higher or equal than threshold) or no bloom (probability value lower than threshold). Results are useful for generation bloom/no bloom maps.

In practice, two kinds of maps products are generated, as we can see in the example in Figure 7: a bloom probability map (derived in step 5) and a bloom/no bloom map (obtained in step 6).

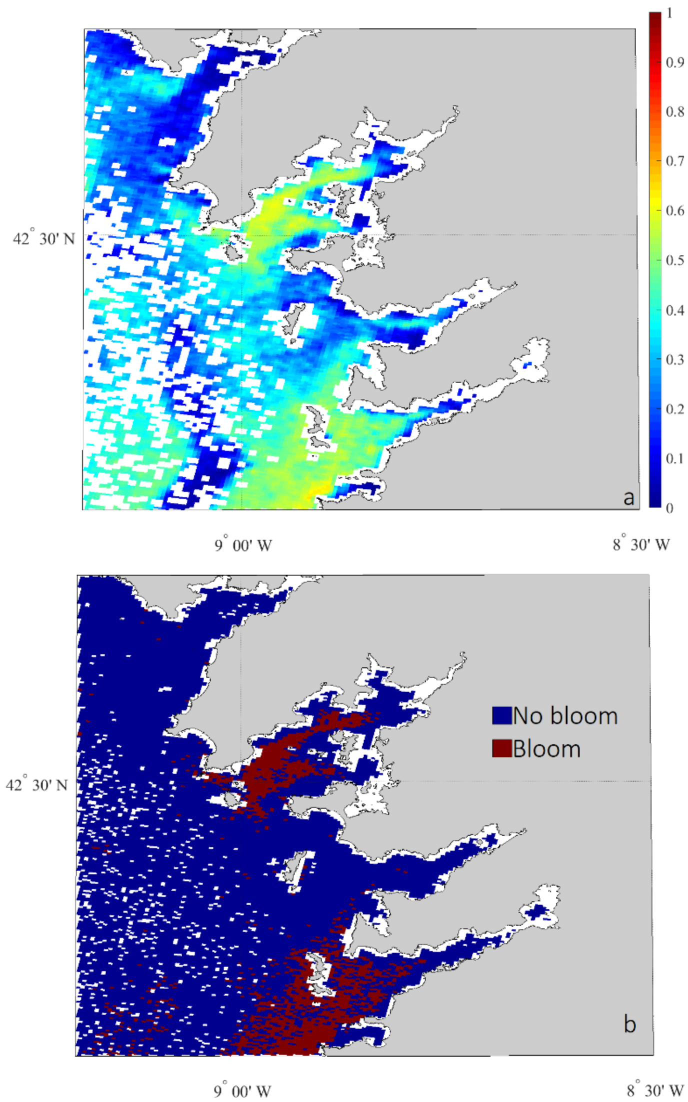


Figure 7 – Results from SVM bloom prediction model for Pseudo-nitzschia spp. on 1 August 2018. A) Bloom probability map. B) Bloom/No Bloom map.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

4.3 Species indicators for *Alexandrium minutum*

In summer 2018, between the end of May and the beginning of August, a HABs of *Alexandrium minutum* was detected in the southern Rias Baixas, i.e. Vigo and Pontevedra. High concentrations (higher than 10^6 cells L^{-1}) were recorded. In fact, the bloom was even visible to the naked eye, appearing as brown-red patches (Figure 8).

Figure 8 – Detail of a brown patch due to a bloom of Alexandrium minutum in the Ria de Vigo on July 5, 2018.

Despite of the limited image availability because of fog and clouds, species indicators for Alexandrium minutum were developed for both Sentinel-3 and Sentinel-2.

4.3.1 Sentinel-3 species indicator

4.3.1.1 Algorithm development

Dataset

This product was developed using *Alexandrium minutum* abundances measured from water samples collected during the field campaign conducted in the Ria de Vigo within CoastObs and the INTECMAR routine monitoring program (including the four Rias Baixas). The complete *in situ* database includes 385 data points from 20 different dates between 28 May and 1 August, 2018. Bloom (abundances greater than 10^5 cells L^{-1}) and significant presence (abundances from 10^4 cells L^{-1} to 10^5 cells L^{-1}) of *Alexandrium minutum* were detected in 24 and 52 sampling stations, respectively, all of them located in the *rias* of Vigo and Pontevedra. In this way, 36.5 % of 208 sampling stations in these rias were affected by A. minutum, but it was not detected in the *rias* of Arousa and Muros (summing up 177 sampling stations).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Reflectance values were extracted from atmospherically corrected (Polymer) Sentinel-3 images and associated with the *in situ* dataset considering only stations sampled on dates with images, summing up a total of 177 data points. The final valid match-up dataset, after filtering masked pixels and/or pixels with a quality indicator lower than 9, includes only 60 data points derived from 5 images. Global abundance distribution, as well as for southern and northern *rias* 8 (Figure 1), is shown in Table 8.

Table 8 – Abundance distribution of Alexandrium minutum for the complete study area, southern rias (Vigo and Pontevedra) and northern rias (Muros and Arousa).

Abundance (cells L^{-1})	Global	Vigo and Pontevedra	Muros and Arousa
0	42 (70.00 %)	3 (14.29 %)	39 (100 %)
$10\text{-}10^4$	5 (8.33 %)	5(23.81 %)	0
$10^4\text{-}10^5$	6 (10.00 %)	6 (28.57 %)	0
$>10^5$	7 (11.67 %)	7 (33.33 %)	0
Absence ($< 10^4$)	47 (78.33 %)	3 (14.29 %)	39 (100 %)
Presence ($> 10^4$)	13 (21.67 %)	18 (85.71 %)	0
Total	60	21	39

Data points with zero abundances (abundances below the detection limit) represent 70% of the data points in the global dataset, prevent us from obtaining reliable results using approximation functions to estimate abundances. On the other hand, only 7 data points (11.67 %) were identified as bloom (abundances greater than 10^5 cells L^{-1}) and hence SVM presence/absence models were preferred to bloom/no bloom models. A threshold of 10^4 cells L^{-1} was selected, so that a ratio 20:80 between the minority class (presence) and the majority class (absence) was observed.

As observed in the complete *in situ* dataset, there is a clear spatial pattern. While northern *rias* (Muros and Arousa) were not affected by *A. minutum*, presence was observed in 85.71 % of the samplings in the southern *rias* (Vigo and Pontevedra) with more than 30% of blooms.

SVM presence/absence model

Input variables include 10 reflectance values between 400 nm and 779 nm. Geometry values were excluded due to the limited availability of images. Table shows the minimum and maximum values used for scaling the input data between 0 and 1.

The best presence/absence model was obtained as explained in section. First, the optimal parametric configuration was selected by applying a leave-one-out cross validation, and secondly the final SVM model is obtained by training the complete training dataset using the optimal values of C and γ. Probability output was preferred to binary output.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Performance measures computed from the leave-one-out cross-validation process and from applying the best model to the training dataset are shown in Table 9. Threshold-dependant measures (all except *AUC*) were computed using the optimal threshold (i.e. threshold maximizing the sum *TPR + TNR*).

Table 9 – Performance measures computed from the leave-one-out cross-validation process and from the training dataset using the final presence SVM model.

	Leave-one-out cross-validation	Training dataset (Final SVM Model)
OA	0.80	0.87
TPR	0.77	0.85
TNR	0.81	0.87
FNR	0.23	0.15
FPR	0.19	0.13
κ	0.77	0.85
AUC	0.87	0.94
Optimal threshold	0.19	0.20

Despite of the low size of the dataset, the SVM model is quite robust, showing similar accuracies for both presence and absence classes (around 0.8), a good *AUC* of 0.87 and a substantial agreement (κ = 0.77) computed from the cross-validation dataset. Results from the training dataset are only slightly better than the obtained ones by cross-validation, indicating a good generalization capability. The final model is able to predict more than 85% of presence or absence situations correctly, with less than 15% of false alarms.

Within the bloom prediction product, this model is only an intermediate step that is aimed at identifying presence areas where an abundance estimation algorithm will be applied in the next stage. As a probability output is obtained, application of a threshold is required to generate a binary (presence/absence) output. Instead of using the optimal threshold, we opted for the minimum presence threshold (*MPT*), i.e., the threshold allowing for identifying all the presence data points (*TPR = 1*) at the cost of a decrease in the absence accuracy (*TNR*) and an increase of false alarm rate (*FPR*). MPT is always lower than the optimal threshold. In this case, it takes a value of 0.13 (instead of 0.2), with the false alarm rate increasing from 0.13 to 0.30.

Abundance estimation algorithm

A basic abundance estimation algorithm based on multiple regression was developed using only data points with abundances higher than zero (*n = 18*), so that data points below the detection limit were excluded.

Despite of the low size of the dataset, a significant regression (R^2 = 0.79; F = 5.26; p<0.01) was found using seven reflectance bands as independent variables and abundances (log-10

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

transformed) as dependent variable. Regression parameters and performance measures are summarized in Table 10.

Table 10 – Regression parameters and performance measures for the abundance estimation algorithm based on multiple regression.

Input variable	Coefficient	t statistic
Intercept	1.43	0.89
Rw400 x 10^4	-0.05	-1.32
Rw412 x 10^4	0.04	0.69
Rw443 x 10^4	-0.11	-1.50
Rw490 x 10^4	0.15	3.46**
Rw620 x 10^4	-0.35	-3.02*
Rw665 x 10^4	0.30	2.19
Rw754 x 10^4	0.21	3.03*
Performance measures		
R^2		0.79
MPE		0.00
VAR		0.21
RMSE		0.45
RelRMSE		30.90 %

According to the results, reflectances at 490 nm ($p<0.01$), 620 nm ($p<0.05$) and 754 nm ($p<0.05$) show a significant correlation with the abundance of *A. minutum,* so that it increases with increasing reflectances at 490 nm and 754 nm but with decreasing values at 754 nm.

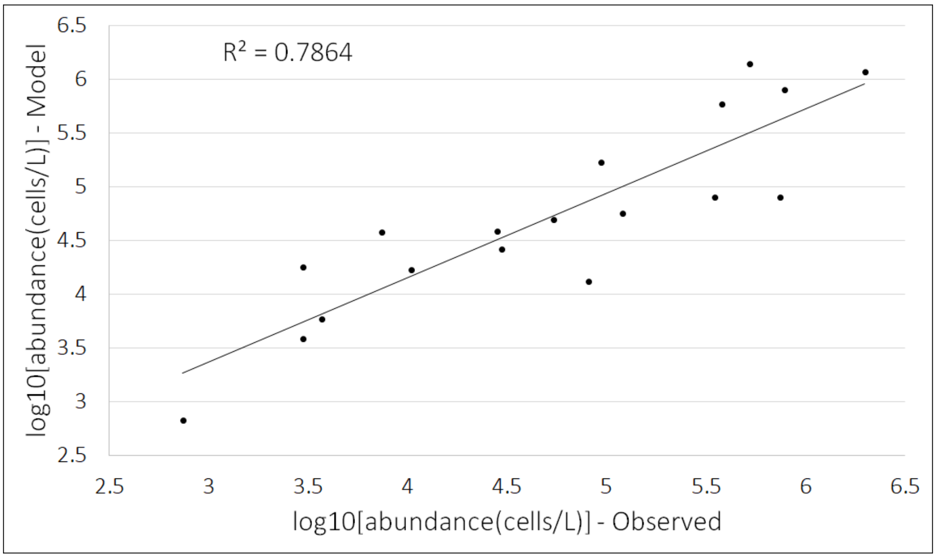


Figure 9 – Relationship between the abundance observed and predicted using the abundance estimation algorithm.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Figure 9 shows the relationship between the observed and predicted abundances. A clear linear trend is observed since the model provides a good fit without overestimating or underestimating the abundances (*MPE = 0*).

4.3.1.2 Generation of map products

Figure 10 summarizes the chain processing for the generation of a bloom probability map starting from a Sentinel-3 image.

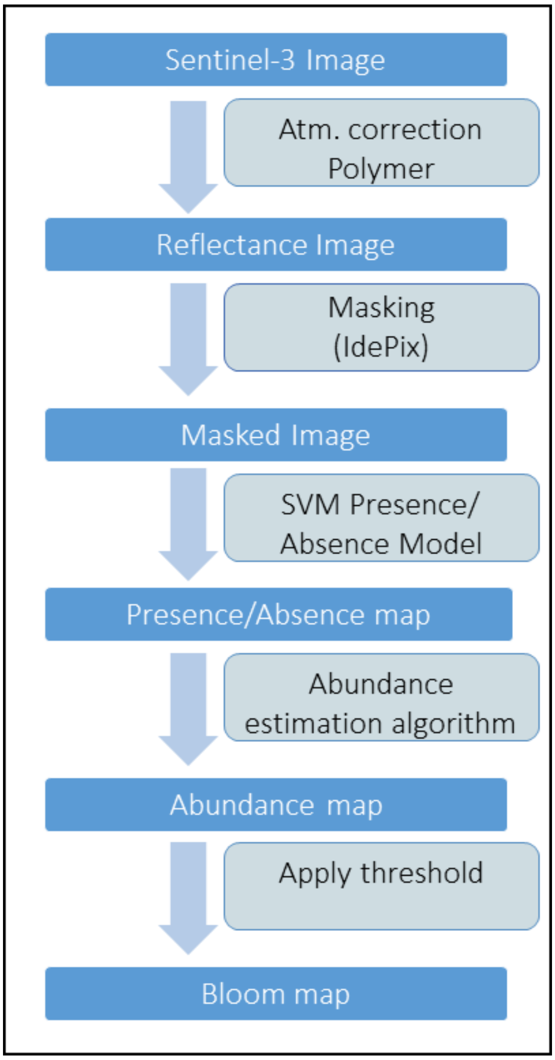


Figure 10 – Chain processing for generating a Alexandrium minutum bloom/no blooom map starting from a Sentinel-3 image.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

For a given non-masked pixel, the methodology consists of the next steps:

1) Application of the SVM presence/absence model to obtain the presence probability for *Alexandrium minutum* (see section 4.2.3 for more details).
2) Application of a threshold of 0.13 (MTP threshold) to classify the pixel as presence (presence probability higher or equal than the threshold) or absence (presence probability value lower than the threshold)
3) Application of the abundance estimation algorithm based on multiple regression for predicting abundances of *Alexandrium minutum*.
4) Identification of the pixel as bloom or no bloom. The pixel is classified as bloom only if it was classified as presence in step 2 and if the abundance is higher than a threshold of 6 (abundance units are log-transformed, it corresponds with 10^6 cells/L). If not, it is identified as no bloom

A different map product is obtained in each step: 1) presence probability map; 2) presence/absence map; 3) abundance map; and 4) bloom/no bloom map. Figure 11 shows an example with the different maps products derived from the Sentinel-3 image on.

Note that the presence probability map (Figure 11a) only predict presence (abundances higher than 10^4 cells/L), so that high probability values do not imply the presence of a bloom. If fact, this map is an intermediate product aimed at defining presence areas (Figure 11b) where the abundance algorithm (Figure 11c) is expected to provide more reliable results.

The bloom/no bloom product (Figure 11d) combines information from the SVM presence probability model and the abundance estimation algorithm, and hence it is supposed to provide a more reliable result.

In any case, this species indicator was developed using a limited number of images and it should be applied with caution, since it is only valid for a limited range of environmental conditions occurring between May and August, 2018. In a later stage of the project, the species indicator will be improved by a further validation and the integration of additional data in a more reliable higher level product.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

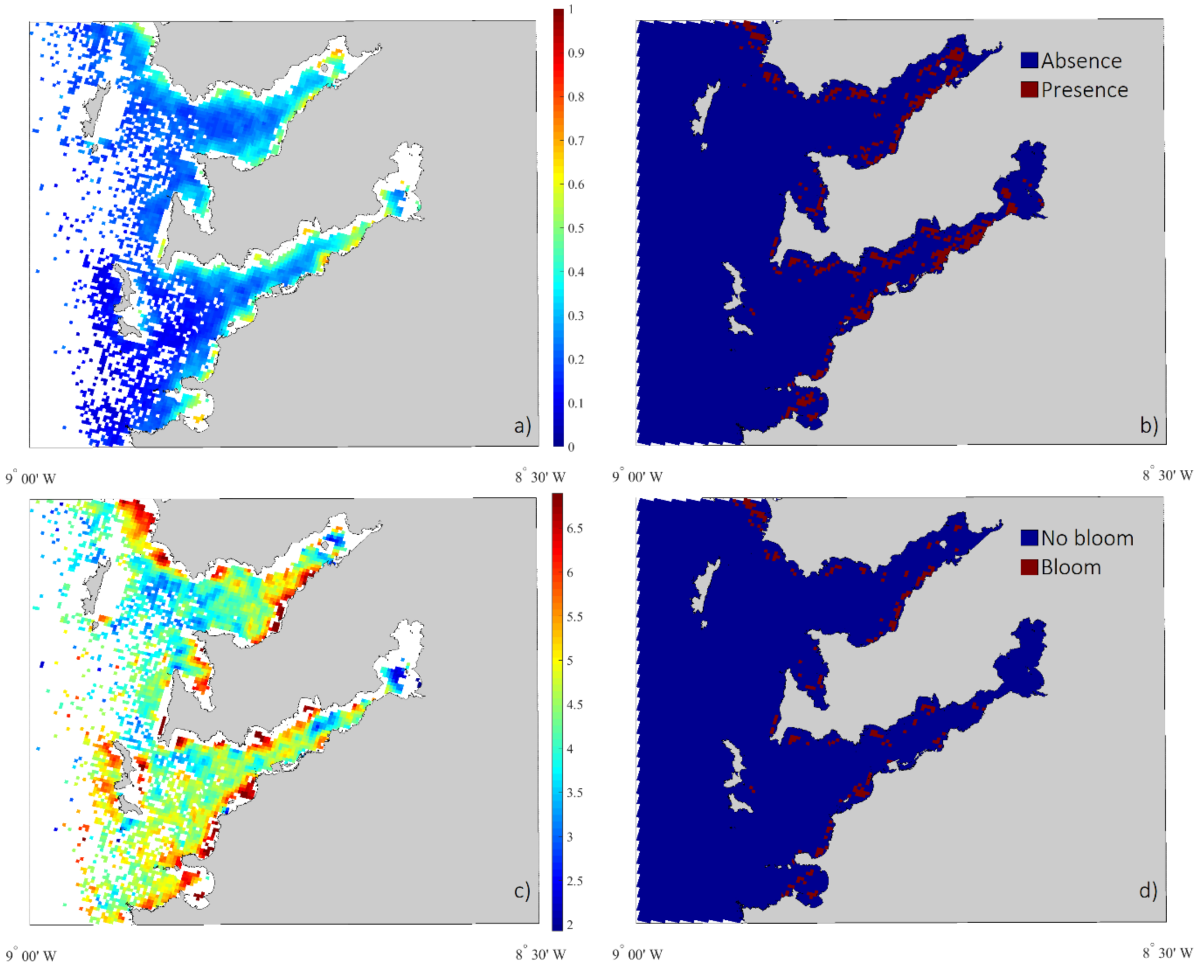


Figure 11 – Results from Sentinel-3 species indicator for Alexandrium minutum on 4 July 2018. a) Presence probability map. b) Presence/Absence map. c) Abundance map (in log10[abundance(cells/L)]). d) Bloom/No bloom map.

4.3.2 Sentinel-2 species indicator

4.3.2.1 Algorithm development

A simple algorithm for detecting blooms of *Alexandrium minutum* from high-resolution Sentinel-2 images based on detecting its characteristic spectral signature was developed.

The algorithm is mainly based on reflectance spectra collected using TriOS field radiometer on 17 July 2018, which were linked with *Alexandrium minutum* abundance data measured from water samples collected on the same stations.

Then, 6 available reflectance spectra were grouped and averaged for three classes (Figure 12): bloom (abundances higher than 10^5 cells/L, 2 spectra), no bloom with clean waters (3 spectra) and no bloom with sediment-dominated waters (1 spectrum).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

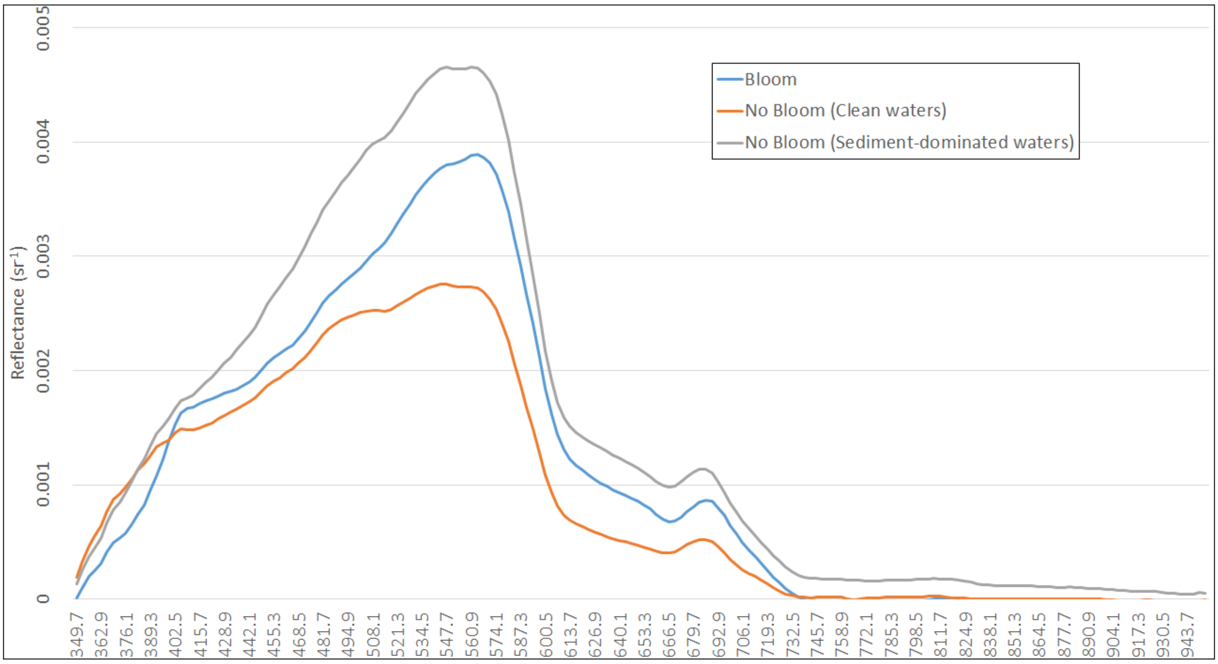


Figure 12 – Average field spectra for three classes defined using in situ data.

Average field spectra for each class were first simulated for the Sentinel-2 bands using the corresponding spectral response function, assuming a Gaussian distribution around the band centres.

Then, using a match-up database linking Polymer data and spectra collected during the field campaign conducted in the Ria de Vigo in July 2018, a linear relationship ($R^2 = 0.64$) was found between Polymer normalized reflectances and *in situ* reflectances.

Using this relationship, simulated Sentinel-2 spectra were finally adapted to Polymer normalized reflectances (Figure 13).

The membership degree to each class can be computed for a given open water pixel in the same way as it is done with clusters centres derived from FCM algorithm. Hence, two kinds of products could be derived from Sentinel-2 images: bloom images showing the pixels belonging to the bloom class, or bloom probability images, showing the membership degree (between 0 and 1) of the bloom class.

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 776348

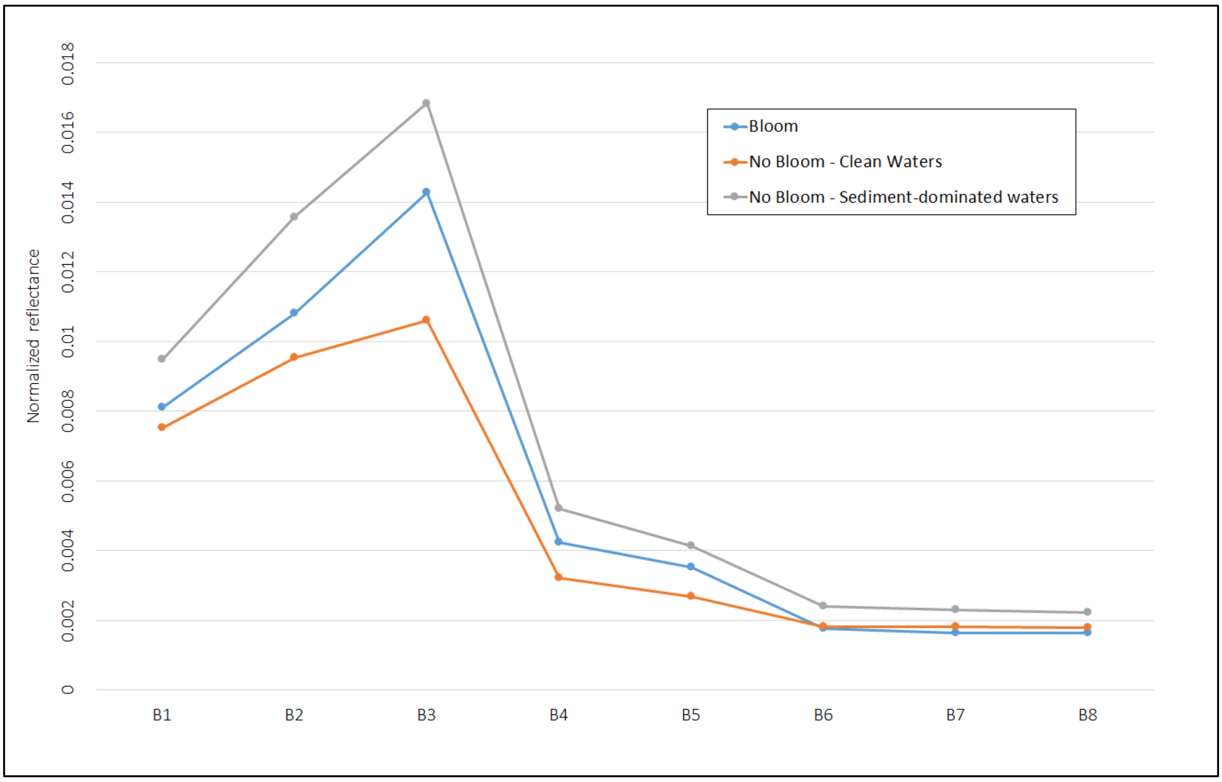


Figure 13 – Simulated Sentinel-2 Polymer reflectance spectra for three classes defined using field spectra and in situ data.

4.3.2.2 Generation of maps products

Figure 14 summarizes the chain processing for the generation of a bloom/no bloom or a bloom probability map starting from a Sentinel-2 image.

Map product generation is based on the same procedure explained in section 3.3 to build FCM intermediate map products, but using the simulated spectral classes defined from field spectra (Figure 13) as clusters centres in order to produce grade maps and classification images.

In this way, two kinds of maps products are generated:

1) Bloom probability map: It is the grade map showing the membership degree to the class “bloom” for each pixel.

2) Bloom/ No Bloom map: It is the classification image showing the class with the highest membership degree for each pixel.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 776348

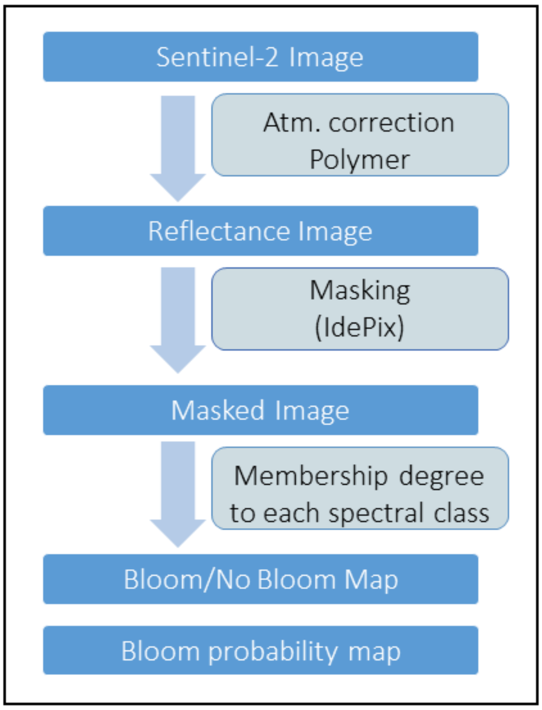


Figure 14 – Chain processing for generating a Alexandrium minutum bloom/no bloom or bloom probability map starting from a Sentinel-2 image.

Figure 15 shows an example of a bloom probability map on 17 July 2018. As compared to the Sentinel-3 species indicator, it provides a more accurate mapping of the interior of the *rias*. Since patches of *Alexandrium minutum* were mainly observed in areas near the coastline where data from Sentinel-3 could be less reliable, information is very relevant.

Unfortunately, species indicator from Sentinel-2 has some disadvantages: images are noisier, making more difficult the development of algorithms, longer processing times and a poorer temporal coverage.

Moreover, as Sentinel-3 species indicator, algorithms were developed using a limited number of images and it should be applied with caution since it could show a poor generalization capability.

Sentinel-2 species indicator will be improved by a further validation and it could be also integrated in a higher-level product.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

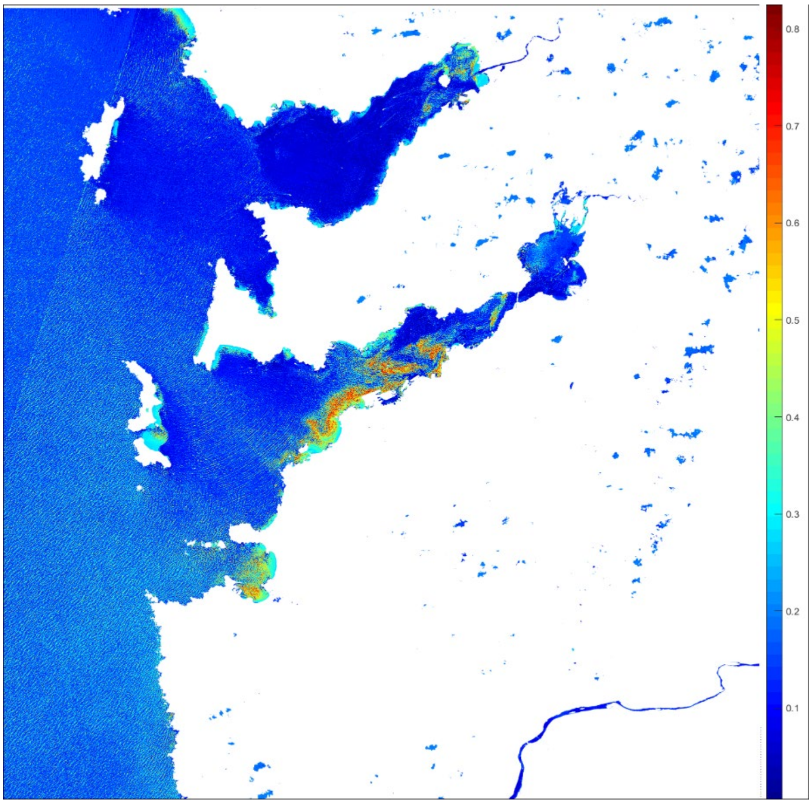


Figure 15 – Bloom probability map on 17 July 2018.

5 Potential use of Galicia HABs products

5.1 Temporal and spatial distribution of phytoplankton abundance

Since chlorophyll is common to almost all phytoplanktonic groups, chl*a* concentration is a good estimator of phytoplankton abundance. In fact, chl*a* maps derived from MERIS images have already proven to be a useful tool for analysing phytoplankton distribution and detecting high biomass "patches" in the Galician *rias* and the adjacent continental shelf during a upwelling cycle in 2008 (Spyrakos *et al.*, 2018).

The spatial structure of the phytoplankton distribution revealed by chl*a* maps is affected by regional characteristics, i.e., surface currents, freshwater inputs and specially the upwelling-dowelling cycle, which depends on the prevailing winds in the platform area. A physical-biological coupling is often observed, so that high chl*a* concentrations coupled with low sea surface temperatures are associated with summer and spring upwelling events.

Figure 16 shows an example of the increase of chl*a* concentration because of an upwelling event. On 29 August 2016 (Figure 16a), generally low concentrations were observed. A week later, on 5 September 2016 (Figure 16b), patches of high chlorophyll were observed mainly in the outer parts of the *rias* and in the adjacent continental shelf. Both images were processed using NNRB-Cl#2.

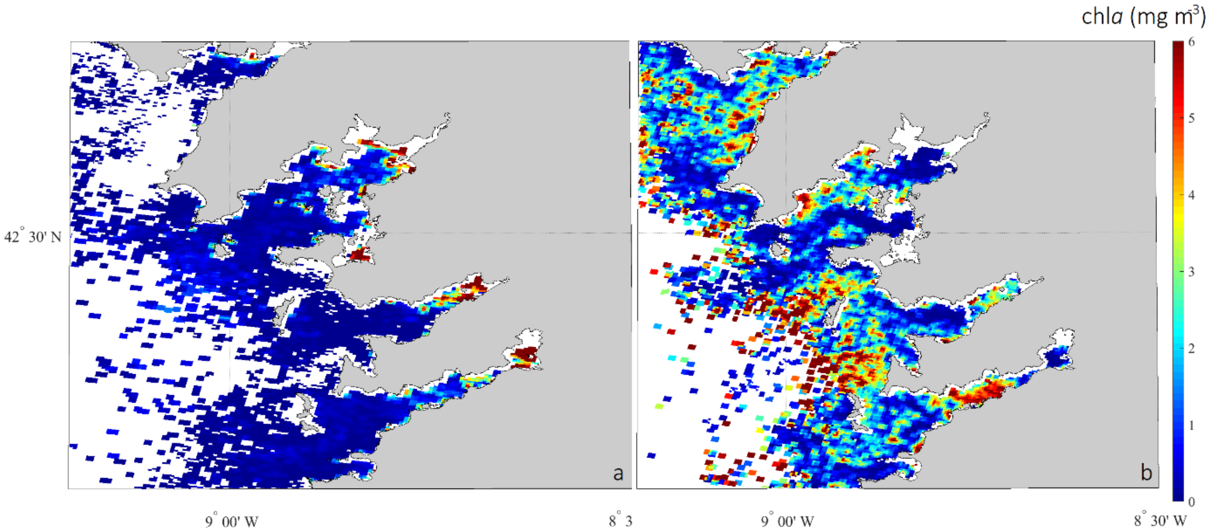


Figure 16 – Chla maps derived from Sentinel-3 images using NNRB-Cl#2 algorithm. a) 29 August 2016. b) 5 September 2016.

Chl*a* maps may also provide useful information about HABs. For instance, "high biomass" patches linked with upwelling events are usually related to the dominance of diatoms, including the potentially toxic *Pseudo-nitzschia* spp. Chl*a* maps could also be useful for tracking blooms of *Gymnodynium catenatum*, a paralytic shellfish toxin (PST) producer that usually follows a northward progression from the Portuguese coast to the Galician *rias*, showing the highest abundances at the end of upwelling events.

Despite the fact that data can only be obtained under cloud-free conditions, chl*a* maps derived from Sentinel-3 images provide a good spatial resolution (300 m), allowing an accurate mapping of the interior the Rias Baixas and the adjacent continental shelf, and an excellent temporal coverage with a daily image available since December 2018 thanks to a two-satellite configuration (Sentinel 3a and b). As compared to field studies, with a limited spatial coverage and temporal frequency, chl*a* maps have already proven to be able to record dynamic changes in chl*a* distributions that can be missed by *in situ* monitoring.

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 776348

5.2 HABs detection (higher-level products)

Species indicators are mainly aimed to the detection of HABs, i.e. massive proliferations (or bloom) of specific toxic species from satellite images. As compared to direct observation methods based on sampling stations, species indicators are faster, cheaper and produce map outputs providing a more synoptic view of the study area (Blondeau-Patissier *et al.*, 2014; Kudela *et al.*, 2017). Sentinel-3 images provide a good spatial resolution and an excellent temporal coverage, with a daily image, although map product generation is limited by the cloud cover. However, as explained in section 4.1, results reliability is hindered by the limited spectral resolution of the images and specially by the complexity of the coastal waters.

Considering the fact that HABs of specific toxic species are usually associated with certain environmental conditions, the combination of the species indicators with additional data, acquired on the previous and/or same day as the image, could lead to a significant improvement in the reliability and accuracy of the HABs detection.

HABs higher-level products integrating map products (chl*a* maps and species indicators) and auxiliary data are currently being developed within CoastObs in Task 3.8. Results will be presented in D3.8 (due to August 2019).

The *Pseudo-nitzschia* spp. species indicator already provides reliable results allowing to distinguish between days clearly affected by blooms and “no bloom” days (Figure 17), but without providing information about toxicity (*Pseudo-nitzschia* spp. include both toxic and non-toxic species). The higher-level product integrating upwelling indices and nutrient concentrations in the previous days is aimed to improving bloom detection and determining (to some extent) if the detected bloom is toxic, i.e. DA producer.

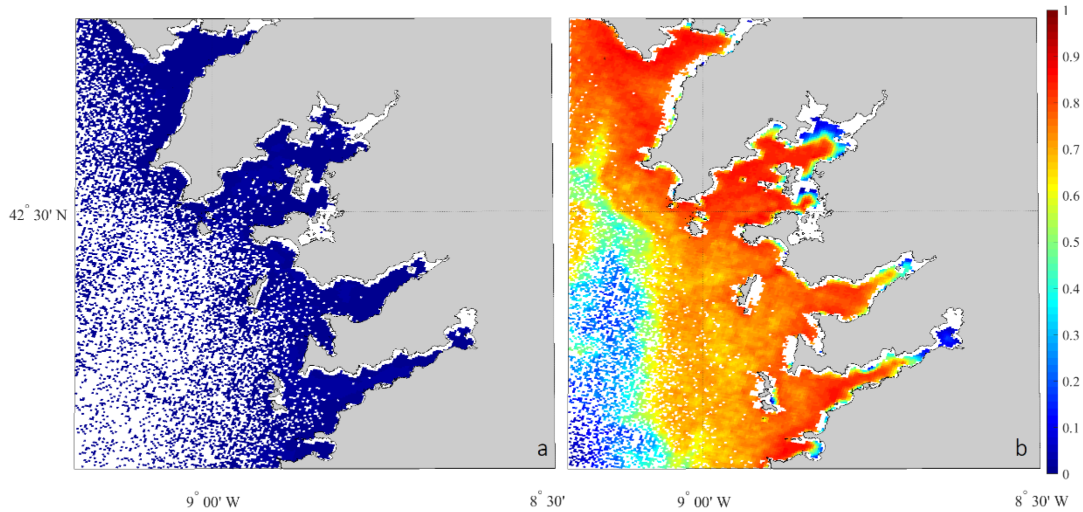


Figure 17 – Pseudo-nitzschia spp. bloom probability maps derived from SVM model. a) 4 December 2017. b) 20 June 2016.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 776348

Regarding *Alexandrium minutum*, development of the bloom in summer 2018 was associated with anomalously high sea surface temperatures (SST), so that a higher-level product including SST data could significantly improve the accuracy of the species indicator, reducing false alarms.

5.3 EO products for mussels farms

INTECMAR monitoring program in Galicia is mainly focused on maintaining shellfish safety, making decisions about closure and/or reopening of the mussel production areas. This service is essential and irreplaceable, but it shows some limitations in terms of temporal and spatial coverage. It is only based on weekly samplings in a limited number of stations, and thus important information could be missed considering that the Rias Baixas is a highly dynamic environment.

Therefore, map products derived from satellite images could complement the existing monitoring program by providing information on a daily basis about the spatial distribution of phytoplankton abundance and/or HABs with a higher spatial resolution.

In order to compare and/or combine the map products with information from the monitoring program or other sources, data can be extracted for a given sampling station (see section 2.3) or using spatial averages for specific areas (e.g. *rias*, mollusc production areas or rafts polygons). For instance, Figure 18 shows average chl*a* values derived from maps processed using the NNRB-Cl#2 algorithm for the mollusc production areas defined by the Spanish government for two dates on July 2018.

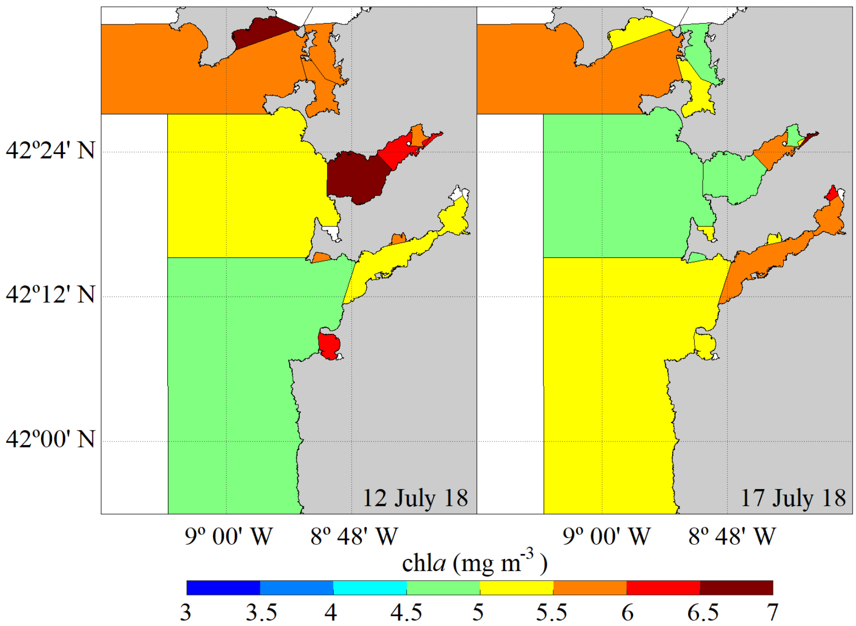


Figure 18 – Averaged chla values derived from NN chla maps for the mollusc production areas on 12 and 17 July 2018.

6 *Phaeocystis* North Sea: Environmental drivers

6.1 Temperature

Phaeocystis can survive in a broad range of temperatures and has even been found as an intact colony at -0.6°C during the winter period in Dutch coastal waters (Riegman and van Boekel, 1996). In the case of the North Sea, *P. globosa* has a temperature range of -4°C to +20°C. Due to the broad temperature range, the concluding hypothesis was that the initiation of the blooms was not well correlated with the temperature of the water (Riegman and van Boekel, 1996). Despite this, temperature still has an effect on the specific growth rate of *Phaeocystis* (Schoemann *et al*., 2005; Peperzak *et al*., 1998; Sun *et al*., 2018), although arguably, light is a more important factor influencing the timing of the blooms for *Phaeocystis* (Riegman and van Boekel, 1996).

6.2 Light

Compelling evidence gathered in literature suggests that although *Phaeocystis* blooms are not controlled by temperature, the amount of daily irradiance available for photosynthesis appears to control the timing of the blooms. Peperzak (2002) found that at surface irradiances of above 38 E/m^2/day, *Phaeocystis* cells were able to begin forming colonies. This was observed in enclosure experiments, where diatoms dominated the phytoplankton population with a daily irradiance of 60 Wh/m^2/day (Peperzak, 2002). Such enclosure experiments reflect an artificial environment and may not reproduce in reality. In 1992, a delay in the timing of a *Phaeocystis* bloom in Marsdiep, Netherlands was attributed to the fact that in the days prior, the daily irradiances recorded in that area were less than approximately 38 E/m^2/day (Brussaard, 1995). This also strengthens the hypothesis that the timing of the blooms is controlled by the amount of daily surface irradiance. High growth rates of more than 1 division per day are found when this daily irradiance threshold is exceeded (Jahnke *et al*., 1989). The growth rate, however, is not exponentially infinite as *Phaeocystis* is also affected by photoinhibition, which is reached at values of 91-180 $E/m^2/s$ (Blauw *et al*., 2010). Yet, light is not the only factor that plays an important role in the formation of a *Phaeocystis* bloom, as nutrients also play a crucial role in this process.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

6.3 Nutrients

Recent decrease in phosphate loading into coastal areas has led to an increasingly favourable environment for *Phaeocystis* (Verity *et al*., 2006; Lancelot *et al*., 2007). For *P. globosa*, under light-saturated conditions, the ratio of the assimilation rate of nitrogen to phosphate was found to be 9.8, which is lower than the average ratio of 11.1 for other algae (Hecky and Kilham, 1988). The ability for *Phaeocystis* to outcompete other algae due to its relatively high nitrogen uptake rate also means that this species is less likely to proliferate in the event of excess in phosphate. Since, in the Southern North Sea, nutrient loading is positively correlated with salinity (Desmit *et al*., 2015), salinity presents itself as a potential indicator of the likelihood of a *Phaeocystis* bloom, as done in Blauw *et al*. (2010). This, of course, should be used under the assumption that the nitrogen to phosphor ratio of the total nutrient load from rivers exceeds the Redfield ratio requirement. While salinity can be used as an indicator of nutrient availability, salinity itself is an environmental parameter that also controls the growth rate of *Phaeocystis.*

6.4 Salinity

To examine the effect of salinity on *Phaeocystis*, Peperzak (2002) acquired seawater from the North Sea and Eastern Scheldt Estuary and diluted it to different levels of salinity, thus excluding the proportional change in nutrients typical of an estuarine environment. Results from Peperzak (2002) indicate a very strong correlation between salinity and the growth rate of *Phaeocystis globosa* (r^2 = 0.95) under experimental conditions. At salinities of less than 15 psu, *Phaeocystis* cells started to die, while the maximum growth rate was achieved at 29 psu. Considering that the salinity varies between 26-32 psu along the Dutch coastal waters (Peperzak, 2002), *Phaeocystis globosa* in the Southern North Sea is at least considered to be both euryhaline and eurythermal, much like the rest of the other species of *Phaeocystis* (Schoemann *et al*., 2005).

6.5 Phenological link with diatoms

As mentioned in the introduction, the initiation of a *Phaeocystis* bloom tends to coincide with the depletion of silica. This is in part due to the silica limitation caused by diatoms uptake, resulting in a slowing down of growth rate and the ability for *Phaeocystis* to be able to outcompete the diatoms in such an event (Peperzak, 2002). This population dynamic has also been observed to occur in a 51-year model simulation of North Sea hydrodynamics, resulting in a clear distinction between the diatom peak followed by the *Phaeocystis* bloom, especially in in Regions of Freshwater Influence (ROFI), which includes the Dutch coastal areas (van Leeuwen *et al*., 2015). In the Belgian Coastal Zone (BCZ), this diatom-*Phaeocystis* succession has been

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

observed to occur every year between 1989-2003, though the magnitude of the interannual variability fluctuates depending on other environmental factors such as nutrient loading from the Scheldt (Gypens *et al.*, 2007; Breton *et al.*, 2006).

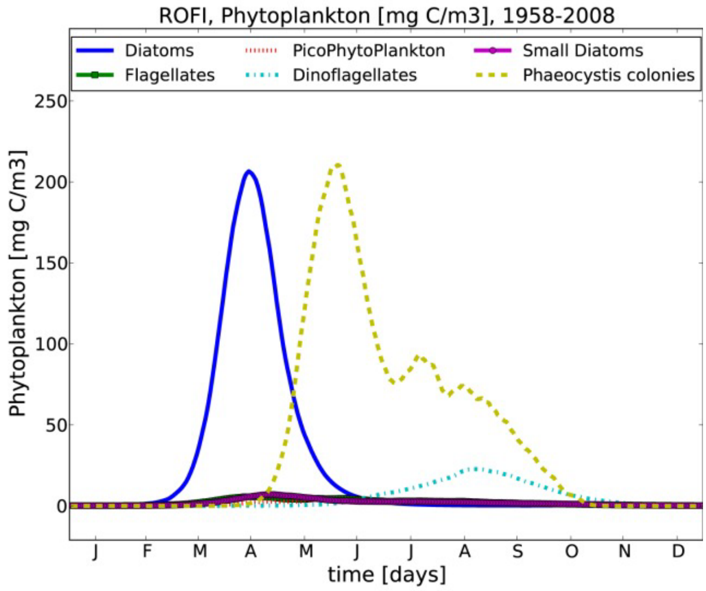


Figure 19 – Phytoplankton dynamics across a 51-year average in the Region of Freshwater Influence (ROFI) (taken from van Leeuwen et al., 2015).

Although the diatom-*Phaeocystis* succession is not an environmental variable, this intrinsic link with the diatom population dynamic allows for preliminary qualitative prediction of a bloom occurrence.

In summary, an assessment of the main environmental variables affecting *Phaeocystis* growth rates and the timing of a bloom highly suggests that light is the most important factor in both the initiation of a *Phaeocystis* bloom and the growth rates of the species. Although temperature is commonly an important factor driving phytoplankton growth rates, the eurythermal *Phaeocystis* can tolerate a wide range of temperatures, thus, this factor more likely dictates a range of environments that this phytoplankton group can survive in, rather than being an influential factor contributing to variability in growth rates. From the literature review, *Phaeocystis* is known to have a strong correlation with the salinity of the water column, with an optimal salinity of approximately 29 psu. Salinity thus presents itself as an alternative indicator of the likelihood of a *Phaeocystis* bloom. In addition, *Phaeocystis* has the ability to assimilate DIN better than the average phytoplankton group, making it a competitive species in nitrogen replete environments such as the Southern North Sea, where the Rhine river brings in a much higher load of nitrogen than phosphor into the North Sea. Lastly, knowledge of the repeatedly occurring diatom-*Phaeocystis* succession is able to contribute to a qualitative prediction of a bloom event occurring. In the next chapter, the existing algorithms available for satellite detection of *Phaeocystis* are examined.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

7 A review of existing algorithms for *Phaeocystis* detection and future research

7.1 Introduction

While it has been possible to detect and estimate the abundance and productivity of phytoplankton in both the open ocean and coastal zones using satellite imagery (see Eleveld *et al*., 2007; Ma *et al*., 2014), distinguishing phytoplankton groups in satellite-based observations is notoriously difficult due to the influence of other algal pigments on the reflectance spectrum (Blondeau-Patissier *et al*., 2014). In fact, the total absorption coefficient of seawater at at a particular wavelength λ can be summarized as a function of the absorption of water aw, phytoplankton and non-algal particles apart and coloured-dissolved organic matter aCDOM (Babin *et al*., 2003), as shown in the following equation.

$$a_t(\lambda) = a_{part}(\lambda) + a_{CDOM}(\lambda) + a_w(\lambda)$$

[Equation 1]

Figure 20 shows an example of the absorption spectra for some of the phytoplankton groups. It becomes immediately clear why detecting specific HAB species can be complicated and difficult to do so. For example, it might not be so simple to distinguish the absorption coefficients of dinoflagellates from premnesiophytes, the group that *Phaeocystis* falls under, due to the similar absorption characteristics in the optical range (Figure 20).

Since *Phaeocystis* is an ecological nuisance that affects the Dutch coastal zones every year during the spring bloom period, it has become increasingly important to identify the spatial and temporal extent of these blooms. Unfortunately, unlike coccolithophores, *Phaeocystis* lacks a strongly distinguishable absorption or scattering characteristic, making the detection of this specific phytoplankton group more challenging. In this chapter, we review the reflectance characteristics of *Phaeocystis* and assess some of the existing algorithms available for the identification and/or quantification of *Phaeocystis* in the Southern North Sea.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

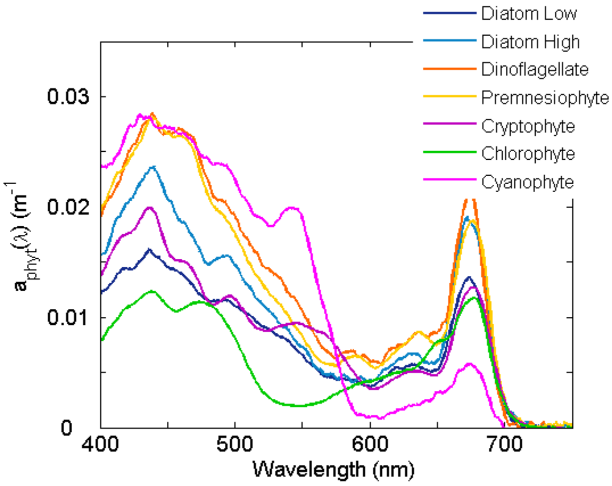


Figure 20 – Absorption feature of different phytoplankton groups (taken from Ocean Optics Web Book [Accessed: 21/02/2019].

7.2 Estimating the abundance of *P. globosa* using the 467nm wavelength

Pure strains of two diatom species, *Thalassiosira rotula* and *Ditylum brightwellii*, and a strain of *Phaeocystis globosa* isolated from the BCZ were grown and cultured under two different light intensities by Astoreca *et al*. (2009). An optical characteristic of *P. globosa* was identified at 467nm, where absorption was higher than that of diatoms due to the absorption of the C3 chlorophyll pigment at that range.

Based on this information, Astoreca *et al*. proposed the utilization of this specific wavelength for the detection of *Phaeocystis*, therefore, quantifying the absorption of chlorophyll C3 (a_{c3}) in the following equation:

$$a_{c3}(\lambda_{c3}) = a_t(\lambda_{c3}) - a_t(\lambda_1)^{[1-w]} * a_t(\lambda_2)^w$$

[Equation 2]

Where $w = \frac{\lambda_{c3}-\lambda_1}{\lambda_2-\lambda_1}$, and λ_1 and λ_2 are 450 nm and 480 nm respectively.

λ_1 and λ_2 represent two blue bands on either side of the chlorophyll C3 absorption bands, forming an exponential interpolation. To apply the algorithm onto water-leaving radiance, substitution of the equation results in the following:

$$a_{c3}(\lambda_{c3}) = \left(\frac{1}{\rho_w(\lambda_{c3})} - \frac{1}{\rho_w(\lambda_1)^{(1-w)}} * \frac{1}{\rho_w(\lambda_2)^w}\right) * a_w(\lambda_{NIR}) * \rho_w(\lambda_{NIR})$$

[Equation 3]

Where λ_{NIR} = 700 nm and a_w(700nm) = 0.57 m^{-1}

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Therefore, based on the absorption feature at 467nm, four bands at 450nm, 467nm, 480nm and 700 nm are used in this algorithm for the detection of *Phaeocystis*. This algorithm can be applied to hyperspectral remote sensing instruments such as the TriOS RAMSES, WISP-3 or WISPstation.

A reasonably good correlation between chlorophyll c3 concentrations derived from a_t and field chlorophyll C3 concentrations was found ($r^2 = 0.72$, $p\text{-value} < 0.0001$), while a high regression coefficient was found ($r^2 = 0.94$, $p\text{-value} < 0.0001$) between chlorophyll c3 concentrations derived from at and field *Phaeocystis* cell numbers. When applying the algorithm onto water-leaving radiance, however, the regression coefficient is then calculated to be 0.56 and 0.57 for regressing chlorophyll C3 concentrations derived from [Equation 2] against field chlorophyll C3 concentrations and field *Phaeocystis* cell numbers respectively.

One major shortcoming of using this algorithm is that although it may be applied to in-situ hyperspectral remote sensing instruments, it cannot be applied to past Envisat-MERIS or current Sentinel-3 imagery. In Envisat-MERIS, there are no bands covering the 467nm range, nor do they have the bandwidth required to reach it. Furthermore, diatoms have an absorption feature at 465nm (Astoreca *et al*., 2009), causing further complications if the spectral resolution is not high enough. The Sentinel-3 OLCI instrument inherits the bands from Envisat-MERIS. Despite the addition of other bands, these still do not sufficiently cover the 4 bands required, and do not have a narrow enough bandwidth to exclude the influence of other absorption features. Therefore, this algorithm is unsuitable for applying onto satellite imagery in order to have a synoptic view of the spatio-temporal extent of a *Phaeocystis* bloom. In the next section, methods for *Phaeocystis* detection involving satellite imagery will be reviewed to explore the feasibility of using such methods that do not involve wavelength-specific algorithms.

7.3 Alternative algorithms for *Phaeocystis* detection

A shortcoming of Astoreca *et al*.’s method (2009) is the lack of the 467nm band in dedicated ocean colour sensors such as MERIS or Sentinel-3. One way of circumventing this issue is to use multispectral specific algorithms such as that proposed by Lubac *et al*. (2008), which focuses on using the 442 nm, 490nm and 510nm bands for the detection of *Phaeocystis*. By using a band ratio of remote sensing reflectance $R_{rs}(490nm)/R_{rs}(510nm)$ and $R_{rs}(442.5nm)/R_{rs}(490nm)$, regressing the first band ratio against the second band ratio respectively results in a significantly different linear fit between *P. globosa* and diatoms (Lubac *et al*., 2008). A sensitivity analysis performed on the ratios suggests that the use of this method is reliable regardless of the bloom intensity, however, the presence of Coloured Dissolved Organic Matter (CDOM) presents a complicating factor as the behaviour of these ratios are more influenced by the composition of phytoplankton instead (Lubac *et al*., 2008). This could be a potential issue if the method was

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 776348

applied onto the Southern North Sea, where there is a high flux of CDOM from the rivers Seine, Scheldt, Meuse and Rhine (Kurekin *et al*., 2014). This method, therefore, presents a possibility for use within multispectral images including SeaWIFS, Envisat MERIS and Sentinel-3, provided the regions being assessed are dominated by *P. globosa* during the period of observation.

Wavelength or band-dependent methods, while offering a relatively straightforward method of quantifying or detecting *Phaeocystis*, can present a multitude of problems associated with the inherent optical properties of the water and the lack of sufficient spectral resolution. Classification methods provide an alternative to those methods, such as the one developed by Miller *et al*. (2006), constituting a HAB classifier that was trained on large datasets including inherent optical properties (i.e. absorption, backscattering) and manual labelling of HAB pixels. Building upon that classifier, Kurekin *et al*. (2014) further developed the method for the specific identification of *P. globosa* and *Karenia mikimotoi*, both HAB species, using a Linear Discriminant Analysis (LDA). The results of the classifier included a successful classification of more than 88% of pixels for both species, along with the potential to linearly transform the components of the (LDA) for the estimation of cell abundances of each species. One advantage of this method is that it does not require the calculation of chlorophyll-a, allowing for greater accuracy as species like *P. globosa* do not have a linear relationship with that parameter. However, it requires manual labelling of HAB pixels, in addition to a large amount of training datasets required in order for the classifier to have a maximum probability of correctly classifying each pixel.

In retrospect, the main issue is that *Phaeocystis*, or its Southern North Sea type *P. globosa,* can be incredibly difficult to detect or quantify using optical methods. This is due to the fact that chlorophyll-a concentrations are not always related to the abundance of this species, along with the presence of the chlorophyll c3 pigment and other carotenoids that can influence the reflectance spectra due to their varying absorption characteristics. Classifier methods like Kurekin *et al*.'s (2014) or Miller *et al*.'s (2006) offer an alternate method for the detection of *Phaeocystis* blooms but come with laborious prerequisites before the classifier can be deemed successful. In the next chapter, we explore the available data for various environmental parameters considered important for the growth of *Phaeocystis* and attempt to create a model that can successfully detect or quantify *P. globosa* in the Dutch coastal zone.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

8 Development and testing of a *Phaeocystis* detection indicator

8.1 Introduction

In Chapter 6, we reviewed the environmental parameters that would typically affect the growth rate and bloom dynamics of *Phaeocystis*. Based on those parameters, a multiple-regression model was developed and applied to satellite imagery in order to capture the spatial evolution of a *Phaeocystis* bloom in the Southern North Sea. An analysis of the available data revealed a strong correlation a combination of irradiance history, distance from coast and salinity levels. Although the model was calibrated using data from the Dutch coast of the North Sea, it should not be used as a quantification model for *Phaeocystis* abundance, but rather, an indicator of the likelihood of occurrence during a particular day of the year. In addition, the model shows good agreement with an actual recorded occurrence of a *Phaeocystis* bloom in 2008 (Peperzak and Poelman, 2008).

8.2 Methodology

8.2.1 Algorithm development

Given the importance of salinity, irradiance and nutrients affecting the growth rate of *Phaeocystis*, a simple classification algorithm was created in order to try to capture the spatial extent of *Phaeocystis* bloom occurrence. In-situ *Phaeocystis* cell abundance and salinity data from the Rijkswaterstaat were used (https://www.informatiehuismarien.nl), while downwelling shortwave radiation, i.e. solar surface irradiance, was taken from the Surface Solar Radiation Data Set – Heliosat (SARAH) using measurements taken from Meteosat Visible Infra-Red Imager (MVIRI).

For modelling purposes, daily average salinity levels of the North Sea were taken from Copernicus Marine Environment Monitoring Service (CMEMS) using the Atlantic-European North West shelf reanalysis based upon the Forecasting Ocean Assimilation Model 7km Atlantic Margin Model (FOAM AMM7). Although the model was calibrated to in-situ surface salinity values, since high spatio-temporal resolution of measured salinity values are not available, only surface salinity values from the FOAM AMM7 model were used in the final model output for detecting *Phaeocystis* blooms. In a preliminary analysis of the influence of salinity on *Phaeocystis*, it was found that there was not a particularly strong relationship present between the data. Instead, by creating a parameter that takes into account the maximum threshold value

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

for *Phaeocystis* abundance and the optimal value for *Phaeocystis* growth, this allowed for a better constrain of the relationship between salinity and *Phaeocystis*. In earlier tests of the model without any transformation on the data, the use of salinity alone resulted in no contribution to improving the correlation between model result and measured data, hence the need to explore alternative transformations of the parameter since salinity was described by Peperzak (2002) to be a very important influencing factor in *Phaeocystis* growth. This is described in the following equation:

$$e^{\frac{Sal_M - Sal_{meas}}{Sal_M - Sal_{opt}}}$$

[Equation 4]

Daily surface irradiance was also taken from SARAH and the irradiance history was calculated by averaging a certain number of days prior to the date of analysis. Exact day, 3-, 5-, 10- and 14-day averages were made in order to determine the most appropriate amount of irradiance history to use as input for the model. Distance from coast was calculated by performing a proximity analysis of the North Sea basin with respect to the coast, thus generating a map containing pixel values representing the distance to the nearest coastal pixel in kilometres (Figure 21).

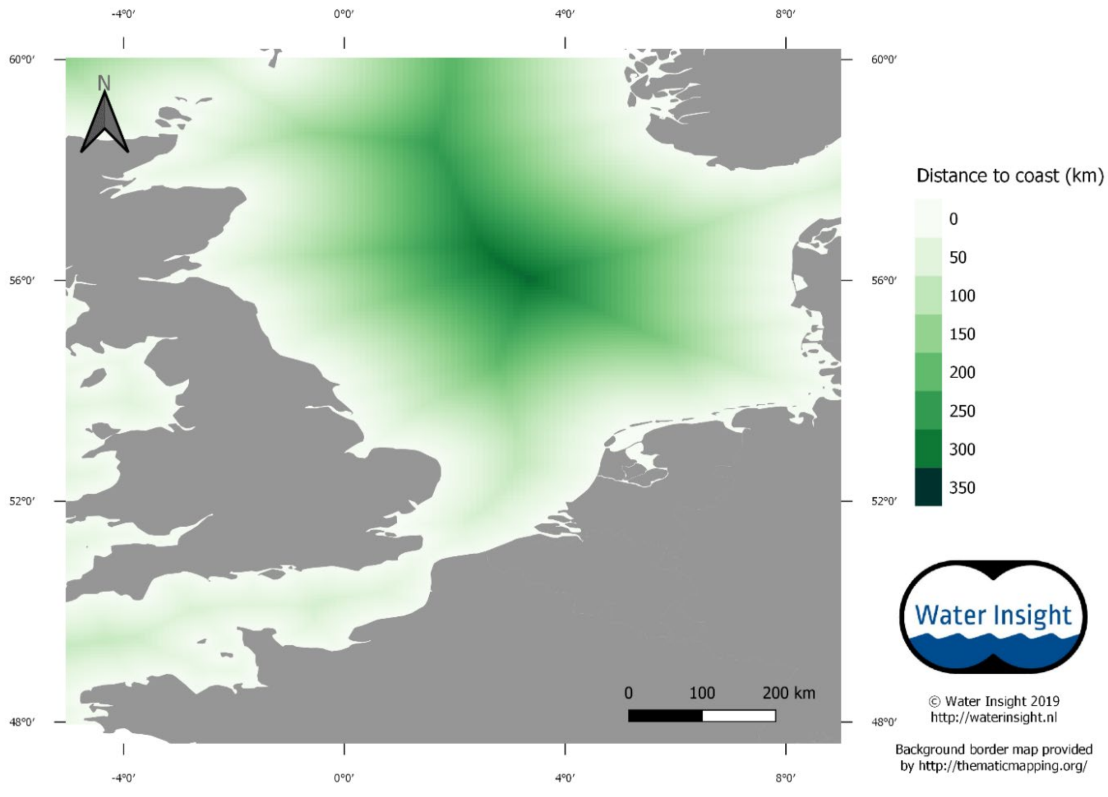


Figure 21 – Map of distance to the nearest coast generated using proximity analysis.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

$$\left[a\left(\frac{1}{D_c}\right)+b\left(e^{\frac{Sal_M-Sal_{meas}}{Sal_M-Sal_{opt}}}\right)+c(Irr)\right]^{d(M_p)}\left.\begin{matrix}< 30: \textit{Little to no likelihood of occurrence}\\ > 30: \textit{Likely chance of occurrence}\\ > 100: \textit{Very high likelihood of occurrence}\end{matrix}\right.$$

[Equation 5]

Table 11 –Table of parameters with set values and parameter description.

Model Parameters	*Set value*	*Term description*
D_c	Pixel-basis	Distance from coast (km)
Sal_M	35	Maximum threshold value for *Phaeocystis* growth rate. Value set at 35psu based on Peperzak (2002)
Sal_{meas}	Pixel-basis	Measured salinity value
Sal_{opt}	29	Optimal salinity value for *Phaeocystis* growth rate. Value set at 29psu based on Peperzak (2002)
Irr	Pixel-basis	Downwelling shortwave radiation (*W/m^2*)
M_p	Based on month of the year, generated using []	Parameter describing the likelihood of a bloom occurrence based on the average *Phaeocystis* cell abundance based on the month

A multiple regression model was developed, and each parameter was assigned a constant value that was calibrated according to the in-situ measured data (*a*, *b*, *c*, *d* in Equation 5). Furthermore, some of the parameters have been parameterized with threshold values to further constrain the model result.

A simple parameter describing the likelihood of bloom occurrence for each month was calculated by creating a 5-order polynomial based on the month number and the percentage of blooms occurring in the dataset. Samples are classified as blooms if *Phaeocystis* cell abundance exceeds more than 80000 cells per litre. This was performed in order to capture the approximate behaviour of bloom occurrence throughout the year, although arguably, relatively less intense blooms may still occur with less than 80000 cells per litre.

The equation is represented in the following:

$$M_p = -0.0002x^5 + 0.0077x^4 - 0.0852x^3 + 0.3812x^2 - 0.563x + 0.5233$$

[Equation 6]

8.3 Results & Discussion

Table 12 – Statistical information about the relationship between various environmental parameters and *Phaeocystis* abundance.

Parameter vs *Phaeocystis* abundance	Statistical information
Salinity	r^2 = -0.212, p-value = 0.031
Exponent-transformed salinity	r^2 = 0.214, p-value = 0.030
5-day average irradiance	r^2 = 0.261, p-value = 0.0152
Distance from coast	r^2 = -0.1369, p-value = 0.091

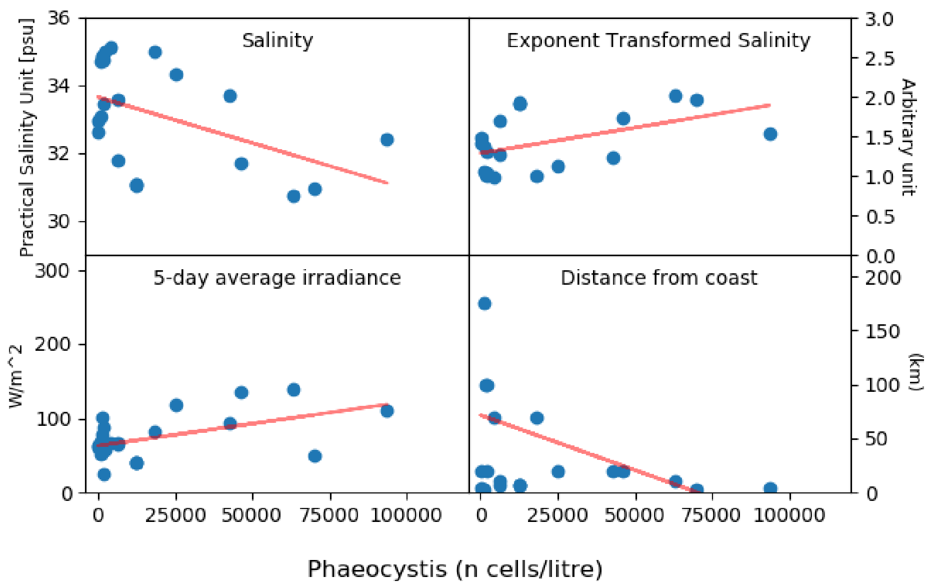


Figure 22 – Plot of various environmental parameters versus Phaeocystis abundance.

To test the effectiveness of the model, the model was applied onto datasets from the year 2008, where a known intense *Phaeocystis* spring bloom occurred, as recorded by Peperzak and Poelman (2008) from stations BG8 and OS4 in the Scheldt Estuary. Figure 23 illustrates a simple classification map of the period prior, during and after the spring bloom recorded by Peperzak and Poelman. According to their results, a strong intense spring bloom began around the 20/04/2008, peaked around 24/04/2008 and finally terminated approximately around

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

20/05/2008 onwards. In Figure 23, between 01/04 – 07/04, relatively large areas of the Southern North Sea had a low probability of *Phaeocystis* occurrence (shown in light green patches). Risk of a medium chance of occurrence increased from 03/04 onwards in patches originating from the West. By 11/04, most of the pixels in the North Sea had a medium chance of occurrence. Of particular interest is the Dutch coastal region and the Wadden Zee, which remains largely at high risk of *Phaeocystis* occurrence throughout April. The spatial pattern of the potential high occurrence is not uniform throughout the period – for most of the Dutch coast and the Wadden Zee, high occurrence pixels are rare with the exception of the Voor Delta and Wadden Zee region in the early part of April. From around 9 April onwards, the red pixels begin expanding further off the coast, with patches of low probability pixels (green) begin shrinking. In Peperzak and Poelman's study (2008), the highest peak was reached on the 24/04, although there was no measurement taken on 23/04, the largest spatial extent of the *Phaeocystis* bloom is reached on 23/04 in Figure 23. Although the intensity of the bloom cannot be inferred from the map, the bloom was recorded to terminate around 20/05 in the study. In In Figure 6, it can be observed that most of the North Sea now has a low probability of occurrence for *Phaeocystis*. In the case of the Scheldt Estuary, this is now reduced to a completely low probability of occurrence, with the exception of the northern part of the "Voor Delta".

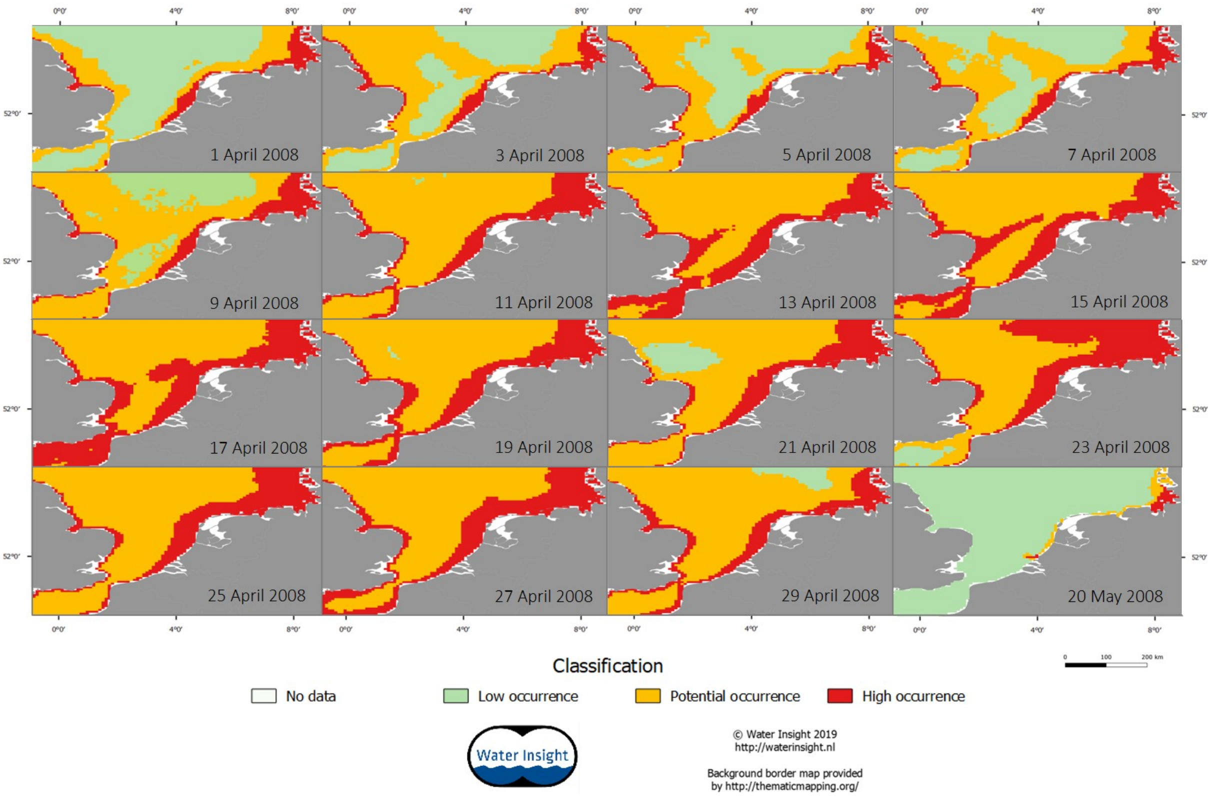


Figure 23 – Classified maps covering the period before, during and after the spring bloom in 2008.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

An important assumption in the model made is that the calibration against data taken from Dutch waters is representative of the Southern North Sea. This may not necessarily hold true, since different species of *Phaeocystis* exist and therefore, may have slightly different threshold values for each of the environmental parameters that might influence their growth. Furthermore, although temperature is a typically major factor influencing growth rates of phytoplankton, changes in temperature do not influence *Phaeocystis* growth rates. Since this was widely reported in scientific literature, it was not included in the study. However, a potential way of constraining the model is adding temperature as an environmental parameter for the purpose of having a threshold, since the North Sea ecotype does not bloom below temperatures of 8-9°C (Riegman and van Boekel, 1996). Additionally, although nutrients influence *Phaeocystis* growth rates strongly, this parameter is not included in the model. Despite nutrients being strongly related to the distance from coast (see de Vries *et al*., 1998), a usage of such a proxy may be inaccurate since nutrients could also be hydrodynamically influenced. Hydrodynamic influence is also not explored in the model, since the model result shows where the *Phaeocystis* bloom is expected to occur, but not where it could potentially occur when taking into account the hydrodynamics of the North Sea. Another way of improving the model is by altering the irradiance parameter such that it takes into account the 100 W/m^2 threshold for *Phaeocystis* bloom occurrence, similar to what was done for salinity. Currently, the model depends on salinity values from the FOAM AMM7 model. There remains the possibility of using remotely sensed salinity values as input for the model, therefore, this option should be explored in the future. Lastly, since salinity had the highest pixel resolution amongst the other input parameters, the pixel resolution of the model output is therefore approximately 7km. This makes it difficult to estimate the likelihood of occurrence in smaller, dynamic areas such as the Eastern Scheldt (Oosterschelde).

9 Conclusions

9.1 Galicia HABs products

HABs map products derived from colour images in Galicia are a useful tool for studying the spatial and temporal distribution of phytoplacton abundance and for the detection and monitoring of HABs of specific species, complementing the existing monitoring program based on field samplings.

HABs map products will be validated using *in situ* data acquired between June and September 2019, including data from the INTECMAR monitoring program and from the dedicated CoastObs field campaign that will be conducted in June 2019. Note that algorithms could also

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

be improved during the validation phase. Validation results will be included in a validation report (D3.10) due to October 2019.

CoastObs products and services are being developed in accordance with a set of criteria established by the final users. In Galicia, final users include the Cooperative of Fishing Ship Owners of Vigo (ARVI) and the Regulatory Council of Mussel from Galicia. Therefore, all the products will be also evaluated and improved according to the user's feedback. User evaluation will start in summer 2019 and will last to the end of the project in October 2020. Results will be included in the service assessment report (D5.4).

9.2 *Phaeocystis* detection

Following a literature review of the environmental parameters that potentially affects *Phaeocystis* growth rates, we present a newly developed algorithm for *Phaeocystis* bloom detection that takes into account salinity, irradiance and distance from coast. Thus far, the model created in this exercise works as a proof of concept by comparing it to data recorded from the *Phaeocystis* spring bloom in 2008. This should be rigorously tested for other years and with scientific literature in order to ensure the full applicability and workability of this algorithm. This new algorithm has potential in being able to provide alerts for coastal zones or shellfish farms when the risk of *Phaeocystis* blooms occurring is high, pending further testing. Future development will include the delineation of *Phaeocystis* blooms either based on a combination of the new algorithm and a chlorophyll-a product or a dedicated optical algorithm using relevant spectral information.

10 References

Anderson, D.M. (2009) Approaches to monitoring, control and management of harmful algal blooms (HABs). Ocean & Coastal Management, 52(7), 342-347.

Astoreca, R., Rousseau, V., Ruddick, K., Knechciak, C., Van Mol, B., Parent, J. and Lancelot, C. (2009) Development and application of an algorithm for detecting *Phaeocystis* globosa blooms in the Case 2 Southern North Sea waters. Journal of Plankton Research 31, 287-300.

Babin, M., Stramski, D., Ferrari, G., Claustre, H., Bricaud, A., Obolensky, G. and Hoepffner, N. (2003) Variations in the light absorption coefficients of phytoplankton, nonalgal particles, and dissolved organic matter in coastal waters around Europe. Journal of Geophysical Research, 108.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Barton, E.D., Largier, J.L., Torres, R., Sheridan, M., Trasviña, A., Souza, A., Pazos, Y. and Valle-Levinson, A. (2015) Coastal upwelling and downwelling forcing of circulation in a semi-enclosed bay: Ria de Vigo. Progress in Oceanography, 134, 173-189.

Blauw, A., Los, F., Huisman, J. and Peperzak, L. (2010) Nuisance foam events and *Phaeocystis globosa* blooms in Dutch coastal waters analyzed with fuzzy logic. Journal of Marine Systems, 83, 115-126.

Blondeau-Patissier, D., Gower, J., Dekker, A., Phinn, S. and Brando, V. (2014) A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans. Progress in Oceanography, 123, 123-144.

Breton, E., Rousseau, V., Parent, J., Ozer, J. and Lancelot, C. (2006) Hydroclimatic modulation of diatom/*Phaeocystis* blooms in nutrient-enriched Belgian coastal waters (North Sea). Limnology and Oceanography, 51, 1401-1409.

Brussaard, C., Riegman, R., Noordeloos, A., Cadée, G., Witte, H., Kop, A., Nieuwland, G., van Duyl, F. and Bak, R. (1995) Effects of grazing, sedimentation and phytoplankton cell lysis on the structure of a coastal pelagic food web. Marine Ecology Progress Series, 123, 259-271.

Camps-Valls, G. and Bruzzone L. (2005) Kernel-based methods for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 43, 1351-1362.

Chang, C.-C.L. and Lin C.-J. (2011) LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2, 27:21-27:27.

Chen, Q. and Mynett, A. (2006) Forecasting*Phaeocystis* globosablooms in the Dutch coast by an integrated numerical and decision tree model. Aquatic Ecosystem Health & Management, 9, 357-364.

Cohen, J.A. (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cortes, C. and Vapnik, V. (1995) Support-vector networks. Machine Learning, 20, 273-297.

Cover, T.M. (1965) Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Transactions on Electronic Computers, Ec14, 326-335.

Cristianini, N. and Shawe-Taylor, J. (2000) An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge (UK): Cambridge University Press.

Davidson, A.T. and Marchant, H.J. (1992) The biology and ecology of *Phaeocystis* (Prymnesiophyceae). In: Round, F.E., Chapman, D.J. (Eds.), Progress in Phycological Research, vol. 8. Biopress, Bristol, pp. 2 – 45.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Desmit, X., Ruddick, K. and Lacroix, G. (2015) Salinity predicts the distribution of chlorophyll a spring peak in the southern North Sea continental waters. Journal of Sea Research, 103, 59-74.

Doerffer, R. and Schiller, H. (2008). MERIS regional coastal and lake case 2 water project—Atmospheric correction ATBD. GKSS Research Center 21502 Geestacht, Germany. Version 1.0, 18 May 2008.

Donlon, C., Berruti, B., Buongiorno, A., Ferreira, M-H., Féménias, P., Frerick, J., Goryl, P., Klein, U., Laur, H., Mavrocordatos, C., Nieke, J., Rebhan, H., Seitz, B., Stroede, J. and Sciarra, R. (2012) The Global Monitoring for Environment and Security (GMES) Sentinel-3 mission. Remote Sensing of Environment, 120, 37-57.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F. and Bargellini, P. (2012) Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sensing of Environment,120, 25-36.

Eleveld, M. A., van der Woerd, H. J. and Beck, H. (2007) Improving the estimation of North Sea primary production: MERIS CHL and Kd in VGPM. In *ESA Special Publication SP-636, Proc* (pp. 1-6). Montreux (CH): ESA ENVISAT.

Frangoulos, M., Guisande, C., Maneiro, I., Riveiro, I. and Franco, J. (2000) Short-term and long-term effects of the toxic dinoflagellate *Alexandrium minutum* on the copepod *Acartia clausi*. Marine Ecology Progress Series, 203, 161–169.

González Vilas, L., Spyrakos, E., Palenzuela, J.M.T. and Pazos, Y. (2014) Support Vector Machine-based method for predicting Pseudo-nitzschia spp. blooms in coastal waters (Galician rias, NW Spain). Progress in Oceanography, 124, 66-77.

González Vilas, L., Spyrakos, E. and Palenzuela, J.M.T. (2011) Neural network estimation of chlorophyll a from MERIS full resolution data for the coastal waters of Galician Rias (NW Spain). Remote Sensing of Environment, 115, 524-535.

Gregg, W.W. and Rousseaux, C.S. (2014) Decadal trends in global pelagic ocean chlorophyll: A new assessment integrating multiple satellites, in situ data, and models. Journal of Geophysical Research C: Oceans, 119(9), 5921-5933.

Gypens, N., Lacroix, G. and Lancelot, C. (2007) Causes of variability in diatom and *Phaeocystis* blooms in Belgian coastal waters between 1989 and 2003: A model study. Journal of Sea Research, 57, 19-35.

Hagan, M.T. and Menhaj, M. (1994) Training feed-forward networks with the Marquardt algorithm, IEEE Transactions on Neural Networks, 5 (6), 989-993.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Haykin, S. (2009) Neural Networks and Learning Machines. 3rd ed. Upper Saddle River (NJ, USA): Pearson Prentice Hall.

Hasle, G. R. (2002) Are most of the domoic acid-producing species of the diatom genus Pseudo-nitzschia cosmopolites? Harmful Algae 1, 137-146.

Hecky, R. and Kilham, P. (1988) Nutrient limitation of phytoplankton in freshwater and marine environments: A review of recent evidence on the effects of enrichment1. Limnology and Oceanography, 33, 796-822.

Hosmer, D.W. and Lemeshow, S. (2000) Applied Logistic Regression. 2nd Ed. New York (NY): John Wiley and Sons.

Jahnke, J. (1989) The light and temperature dependence of growth rate and elemental composition of *Phaeocystis* globosa scherffel and P. Pouchetii (HAR.) Lagerh. in batch cultures. Netherlands Journal of Sea Research, 23, 15-21.

Kudela, R.M., Stumpf, R.P. and Petrov, P. (2017) Acquisition and analysis of remote sensing imagery of harmful algal blooms. In: D.M. Anderson, S.F.E. Boerlage, M.B. Dixon, eds., Harmful Algal Blooms (HABs) and Desalination: A Guide to Impacts, Monitoring and Management, Manual and Guides 78, Intergovernmental Oceanographic Commission of UNESCO 2017, Paris, 119:132.

Kurekin, A., Miller, P. and Van der Woerd, H. (2014) Satellite discrimination of Karenia mikimotoi and *Phaeocystis* harmful algal blooms in European coastal waters: Merged classification of ocean colour data. Harmful Algae, 31, 163-176.

Labarta, U. and Fernández-Reiriz, M.J. (2019) The Galician mussel industry: Innovation and changes in the last forty years. Ocean & Coastal Management, 167, 208-218.

Lancelot, C., Spitz, Y., Gypens, N., Ruddick, K., Becquevort, S., Rousseau, V., Lacroix, G. and Billen, G. (2005) Modelling diatom and *Phaeocystis* blooms and nutrient cycles in the Southern Bight of the North Sea: the MIRO model. Marine Ecology Progress Series, 289, 63-78.

Lancelot, C., Gypens, N., Billen, G., Garnier, J., Roubeix, V. (2007) Testing an integrated river–ocean mathematical tool for linking marine eutrophication to land use: The *Phaeocystis*-dominated Belgian coastal zone (Southern North Sea) over the past 50 years. Journal of Marine Systems, 64, 216-228.

Lary, D.J., Alavi, A. H., Gandomi, A.H. and Walker, A.L. (2016) Machine learning in geosciences and remote sensing. Geoscience Frontiers, 7(1), 3-10.

van Leeuwen, S., Tett, P., Mills, D. and van der Molen, J. (2015) Stratified and nonstratified areas in the North Sea: Long-term variability and biological and policy implications. Journal of Geophysical Research: Oceans, 120, 4670-4686.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Lubac, B., Loisel, H., Guiselin, N., Astoreca, R., Felipe Artigas, L. and Mériaux, X. (2008) Hyperspectral and multispectral ocean color inversions to detect*Phaeocystis* globosablooms in coastal waters. Journal of Geophysical Research, 113.

Ma, S., Tao, Z., Yang, X., Yu, Y., Zhou, X., Ma, W. and Li, Z. (2014) Estimation of Marine Primary Productivity From Satellite-Derived Phytoplankton Absorption Data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7, 3084-3092.

Mercer, J. (1909) Functions of positive and negative type, and their connection with the theory of intergral equations. Philosophical Transactions of the Royal Society of London Series A – Containing Papers of a Mathematical or Physical Character, 209, 415-446.

Miller, P., Shutler, J., Moore, G. and Groom, S. (2006) SeaWiFS discrimination of harmful algal bloom evolution. International Journal of Remote Sensing, 27, 2287-2301.

Moore, T.S., Campbell, J.W. and Dowell, M.D. (2009) A class-based approach to characterizing and mapping the uncertainty of the MODIS ocean chlorophyll product. Remote Sensing of Environment, 113, 2424-2430.

Neil, C., Spyrakos, E., Hunter, P. and Tyler (2019) A global approach for chlorophyll-a retrieval across optically complex inland waters based on optical water types. Remote Sensing of Environment, 229, 159-178.

Nogueira, E., Pérez, F. and Ríos A.F. (1997) Seasonal Patterns and Long-term Trends in an Estuarine Upwelling Ecosystem (Ría de Vigo, NW Spain). Estuarine, Coastal and Shelf Science, 44(3), 285-300.

Palma, S., Mouriño, H., Silva, A., Barao, M. and Moita, M. T. (2010) Can *Pseudo-nitzschia* blooms be modeled by coastal upwelling in Lisbon Bay? Harmful Algae 9(3), 294-303.

Peperzak, L., Colijn, F., Gieskes, W. and Peeters, J. (1998) Development of the diatom-*Phaeocystis* spring bloom in the Dutch coastal zone of the North Sea: the silicon depletion versus the daily irradiance threshold hypothesis. Journal of Plankton Research, 20, 517-537.

Peperzak, L. (2002) The wax and wane of *Phaeocystis* globosa blooms. 19-4-2002. PhD thesis, University of Groningen (the Netherlands).

Peperzak, L. and Poelman, M. (2008) Mass mussel mortality in The Netherlands after a bloom of *Phaeocystis globosa* (prymnesiophyceae). Journal of Sea Research, 60, 220-222.

Peperzak, L., van der Woerd, H. and Timmermans, K. (2015) Disparities between in situ and optically derived carbon biomass and growth rates of the prymnesiophyte *Phaeocystis globosa*. Biogeosciences, 12, 1659-1670.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Pitcher, G.C., Figueiras, F.G., Hickey, B.M. and Moita, M.T. (2010) The physical oceanography of upwelling systems and the development of harmful algal blooms. Progress in Oceanography, 85, 5-32.

Riegman, R. and Van Boekel, W. (1996) The ecophysiology of *Phaeocystis globosa*: A review. Journal of Sea Research, 35, 235-242.

Rodríguez, G.R., Villasante, S. and García-Negro, M.C. (2011) Are red tides affecting economically the commercialization of the Galician (NW Spain) mussel farming? Marine Policy, 35, 252-257.

Sarle, W.S. (1994) Neural networks and statistical models. In: Nineteenth Annual SAS Users Group International Conference, Cary, NC, USA, pp. 1538–1550.

Schoemann, V., Becquevort, S., Stefels, J., Rousseau, V. and Lancelot, C. (2005) *Phaeocystis* blooms in the global ocean and their controlling mechanisms: a review. Journal of Sea Research 53, 43-66.

Spyrakos, E., González Vilas, L., Palenzuela, J.M.T. and Barton, E.D. (2011) Remote sensing chlorophyll a of optically complex waters (rias Baixas, NW Spain): application of a regionally specific chlorophyll *a* algorithm for MERIS full resolution data during an upwelling cycle, Remote Sensing of Environment, 115, 2471–2485.

Steinmetz, F., Deschamps, P.-Y. and Ramon, D. (2011) Atmospheric correction on the presence of sun glint: application to MERIS. Optics Express, 19, 9783-9800.

Sun, Y., Wang, Y., Lei, J., Qian, C., Zhu, X., Akbar, S., Huang, Y. and Yang, Z. (2018) Temperature-dependent elimination efficiency on *Phaeocystis globosa* by different initial population sizes of rotifer *Brachionus plicatilis*. Chemosphere, 203, 474-481.

Teixeira, I., Crespo, B., Nielsen, T. and Figueiras, F. (2012) Role of microzooplankton during a *Phaeocystis* sp. bloom in the Oosterschelde (SW Netherlands). Journal of Marine Systems, 94, 97-106.

Terseleer, N., Gypens, N. and Lancelot, C. (2013) Factors controlling the production of domoic acid by Pseudo-nitzschia (Bacillariophyceae): A model study. Harmful Algae, 24, 45-53.

Utermöhl, H. (1958) Zur vervollkommnung der quantitativen phytoplankton-methodik. Mitteilung Internationale Vereinigung Theoretische unde Amgewandte Limnologie, 9, 1-38.

de Vries, I., Duin, R., Peeters, J., Los, F., Bokhorst, M. and Laane, R. (1998) Patterns and trends in nutrients and phytoplankton in Dutch coastal waters: comparison of time-series analysis, ecological model simulation, and mesocosm experiments. ICES Journal of Marine Science, 55, 620-634.

van der Woerd, H., Blauw, A., Peperzak, L., Pasterkamp, R. and Peters, S. (2011) Analysis of the spatial evolution of the 2003 algal bloom in the Voordelta (North Sea). Journal of Sea Research, 65, 195-204.

Zapata, M., Rodríguez, F. and Garrido, J.L. (2000) Separation of chlorophylls and carotenoids from marine phytoplankton: A new HPLC method using a reversed phase C8 column and pyridine-containing mobile phases. Marine Ecology Progress Series, 195, 29-45.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348