SEAFOOD TOMORROW

Nutritious, safe and sustainable seafood for consumers of tomorrow

Grant agreement no: 773400

Deliverable 1.4

Protocol to manage risks of harmful algal blooms and microbiological impacts of sewage discharges on shellfish production areas

Date of deliverable: 31/10/2019

Actual Submission date: 31/10/2019

Start date of the project: 01/11/2017 **Duration: 36 months**

Organisation name of lead contractor: Centre for Environment, Fisheries and Aquaculture Science (Cefas)

Revision: V1

Project co-funded by the European Commission within the H2020 Programme	
Dissemination Level	
PU Public	X
PP Restricted to other programme participants (including the Commission Services)	
RE Restricted to a group specified by the consortium (including the Commission Services)	
CO Confidential, only for members of the consortium (including the Commission Services)	

Table of Contents

Grant agreement: 773400

1. List of acronyms

CTD – Conductivity, Temperature, Depth

CSO – Combined Sewer Overflow

DSP - Diarrhetic shellfish poisoning

EFSA – European Food Safety Authority

FIB – Faecal indicator bacteria

FFT – Flow to Full Treatment

HABs – Harmful Algal Blooms

IRTA – Institute of Agrifood Research and Technology

ISSC – Interstate Shellfish Sanitation Conference

NoV – Norovirus

NSSP – National Shellfish Sanitation Program (USA)

Ppb – parts per billion

SPA – shellfish production area

STW – sewage treatment works

USFDA – United States Food & Drug Administration

UV – Ultraviolet

2. Glossary of terms

Combined Sewer Overflow – discharge from a combined sewer system where rainfall runoff into sewers combines with sewage flow from homes and businesses that prevents damage to the sewerage system when the volume of flow exceeds the capacity

Geographic Information Systems – a computer-based system for analysing, displaying, interpreting, and storing spatially referenced data.

Microtidal – coastal areas in which the tidal range is less than 2 m.

Macrotidal – coastal areas in which the tidal range is greater than 4 m

Grant agreement: 773400

Neap tides - – when the difference between high and low tides is the least, during the first and third quarters of the moon

Norovirus – small, round structured RNA viruses often implicated in foodborne gastroenteritis outbreaks

Sanitary survey – an inventory of the sources of faecal contamination to a harvesting area together with an assessment of their impact on the microbiological status of the bivalve molluscs grown there.

Sewage – wastewater and excrement carried in sewers

Sewage treatment works - facility for the treatment of sewage wastes from domestic and commercial premises.

Spring tides - when the difference between high and low tides is greatest, at or just after the full and new moons

Tidal range – the difference in vertical height between high and low tide

3. Summary

This project aimed to develop an approach to determining Buffer zones around sewage sources to shellfisheries. Dye tracing studies were undertaken at two study sites, the Upper Fal Estuary in the United Kingdom and Alfacs Bay in Spain, in order to establish and quantify impact over a tidal cycle. Hydrodynamic modelling was utilised to extrapolate this impact over time in order to predict dispersion and dilution under different conditions.

Although current shellfish hygiene regulations are effective in protecting public health from bacterial pathogens in shellfish, outbreaks of illness associated with shellfish consumption still occur due to norovirus and other pathogenic viruses that may be carried in partially-treated or untreated sewage discharges and persist longer in the environment and in the shellfish. Buffer zones can be used to identify areas around sewage sources to be excluded from a shellfish production area due to the risk of contamination with human pathogens, particularly viruses such as norovirus.

This protocol is proposed to help regulators and industry identify the information required to determine Buffer zones and how they can be used to identify locations where new production could be better sited away from hazards.

Grant agreement: 773400

4. Section 1: Introduction to the protocol

4.1. Purposes of buffer zones

Bivalve shellfish (oysters, mussels, clams) are nutritious, delicious food. However, as filter-feeding organisms they pose a unique food-safety risk: when grown in coastal areas, shellfish are exposed to contaminants originating from both human and agricultural wastes in catchment areas adjacent to the waters in which they are grown as well as naturally occurring pathogens and toxins present in their environment.

Human wastes, in particular, are carried to the growing area in sewage effluents, rivers and streams. Illnesses associated with shellfish consumption originate mostly from microbiological pathogens (viruses and bacteria) and from toxin-producing phytoplankton concentrated by shellfish during the filter-feeding process (Rippey 1994) (Lees 2000).

Current official hygiene controls are effective at managing the risk from sewage-derived pathogenic bacterial contamination through the monitoring of faecal indicator bacteria (FIB) such as *Escherichia coli*. However, pathogenic viruses such as norovirus are retained for longer than FIB in at least some species of shellfish and therefore FIB results can underrepresent the norovirus load present. Epidemiological evidence indicates that the largest proportion of cases of foodborne illness is caused by infection with human norovirus (NoV) (Campos and Lees 2014). Environmental factors found to be associated with elevated levels of NoV in shellfish production areas were:

- proximity of shellfish beds to sewage discharges
- insufficient treatment of sewage
- rainfall events
- high river flows
- low salinity
- low water temperatures

Regulators and commercial producers of shellfish therefore need to ensure that these potential health risks are satisfactorily addressed, ideally through appropriate choice of production area and/or adequate post-harvest processing, so that the consumers can enjoy a safe final product.

To better safeguard public health and assist regulatory agencies and members of the industry in identifying the most appropriate locations for shellfish production, a “buffer” or exclusion zone can be established. Buffer zones are areas around point source inputs of human wastewater (usually municipal sewage treatment works; STW) where harvesting of shellfish for human consumption is not permitted. These zones are a preventive public health measure aimed at protection against contamination of shellfish with enteric pathogens.

The use of buffer zones is not currently required under the legislation that determines the hygiene controls on the commercial production of shellfish in the European Union. However, Annex II, Chapter II: C.1 of Regulation (EC) No 854/2004 states that “where the results of sampling show that the health standards for molluscs are exceeded, or that there may be otherwise a risk to human health, the competent authority must close the production area concerned, preventing the harvesting of live bivalve

Grant agreement: 773400

molluscs". The EFSA Panel on Biological Hazards and the European Reference Laboratories for monitoring bacteriological and viral contamination in bivalve molluscs have recommended the implementation of Buffer zones in shellfish production areas (SPAs) (EFSA 2012).

The United States of America (USA) National Shellfish Sanitation Programme (NSSP) contains specific guidance for delineation of buffer zones (Interstate Shellfish Sanitation Conference 2014). Following an agreement made between the Directorate General for Health and Consumer Affairs of the EU Commission and the US Food and Drug Administration during trade negotiations, national authorities of EU Member States that wish to export shellfish to the USA must establish buffer zones around point sources of wastewater discharges impacting SPAs (Anon, Community guide to the principles of good practice for the microbiological classification and monitoring of bivavle mollusc production and relaying areas with regard to Regulation 854/2004 2017).

The original proposal for this work package suggested that the buffer zone concept could be applied to help control the risk of both HABs and norovirus contamination of bivalves.

No causal link has been found between nitrogen to phosphorous ratios and an increase HABs (Davidson, et al. 2012) and particle tracking studies undertaken on a bloom of *Dinophysis acuminata* in the Bay of Biscay suggested that the modellable physical processes alone could account for the dispersion of this bloom (Velo-Suarez, et al. 2010). Therefore, the nature and source of HABs have not been clearly linked to sewage discharges and there is insufficient evidence on which to base this type of approach to HABs controls in bivalves.

However, improvement in the prediction of HABs would be helpful to both industry and regulators. There has been progress in the development of short-term model predictions for both *E. coli* and biotoxins in shellfish in St. Austell Bay and the Fal Estuary (Schmidt, et al. 2018), which showed some promise but have yet to be tested over a longer time range.

A neural network model was attempted for St Austell Bay for this project by Dr. Joaquin Trinanes on behalf of Cefas on a no-cost basis. Satellite remote sensing data were used, however there was a paucity of in-situ data for the area and what was available was found to be insufficient to train the model.

IRTA completed a neural network modelling study for Alfacs Bay, which was well studied and had a rich data set to support the modelling. The outcome of their studies is presented in Appendix 1. The performance of the classical models investigated was not sufficient to support use in forecasting toxic blooms in Alfacs Bay. The neural network model showed greater promise, particularly in forecasting events above the phytoplankton threshold alert concentrations for the area. However, these threshold values vary by member state and in some cases region and the UK threshold values are substantially lower than those used in Alfacs Bay, therefore there is still a need to investigate whether the model developed for Alfacs could be used to predict blooms in other regions.

4.2. Objectives of the protocol

This protocol details the development of a protocol for the implementation of Buffer zones to manage risks associated with norovirus in SPAs.

The protocol provides a means for determining Buffer zones, using predictive models, and effectively communicating their outputs in a way that facilitates effective risk management.

Guidance on predictive modelling for NoV is based on the use of hydrographic and pollution source-tracing studies to determine the time of travel and dispersion and dilution of sewage discharges in SPAs. Previous work in this area set a good foundation for understanding the use of the methods recommended in this protocol and their applicability to prediction of dilution downstream of sewage sources (Campos, Goblick, et al. 2017)

The inclusion of predictive modelling for HABs buffer zones, based on neural-network modelling, was investigated using data from St. Austell Bay in Cornwall. This area has been subject to recurring closures for detection of DSP toxins above maximum permitted levels and was considered a suitable location for study.

5. Approach to buffer zones

5.1. Identify areas suitable for application of buffer zones as part of a monitoring programme

Not all shellfish production areas are directly impacted by point sources of human sewage. The application of buffer zones applies only to areas where human sewage has been identified as impacting the shellfishery.

5.1.1. Role of sanitary survey

The sanitary survey forms the basis for assessment of risks at primary production for any bivalve fishery. In the EU, the conduct and content of sanitary surveys is guided by the European Union Reference Laboratory Guide to Good Practice: Technical Application (GPG) (Anon, Microbiological monitoring of bivalve mollusc harvesting areas - guide to good practice: technical application Issue 7 2018)

Regulation (EC) No 854/2004 states that if the competent authority decides in principle to classify a production or relaying area, it must:

(a) make an inventory of the sources of pollution of human or animal origin likely to contribute contamination to the area,
(b) consider how this vary with seasonal changes in both human and animal sources,
(c) determine how these contaminants are likely to move around the area in accordance with the bathymetry and tidal flows,
(d) establish a sampling programme of bivalve molluscs that ensures the results of monitoring are as representative as possible for the area
(European Communities 2004).

The Guide to Good Practice (GPG) provides more detailed guidance on the types of data to be collected and considered, including data on both continuous and intermittent sewage discharges and how these are likely to impact the area being assessed (Anon 2018).

Grant agreement: 773400

If a sanitary survey has not yet been conducted on the site of interest, it will be necessary to complete one prior to identifying buffer zones.

5.1.2. Protocol – using the sanitary survey

Step 1. The sanitary survey should be reviewed to identify whether human sewage may impact the shellfishery, either through effluent discharge directly to sea or via watercourses that then enter the waters adjacent to or within the shellfishery.

a. If there is no impact from human sewage discharges to the shellfishery, then Buffer zones will not be required.

b. If impact from human sewage discharges is identified, go to step 2.

2. The sanitary survey should provide data relevant to input into the model, including:
   a. Location, volume, and treatment level of continuous sewage discharges
   b. Location, volume and treatment level of intermittent sewage discharges
   c. Location and volume of any trade effluent discharges with sanitary content
   d. Information on tidal cycle, bathymetry and any hydrodynamic modelling studies undertaken for the area.
   e. Meteorological data relevant to the shellfishery
   f. Shellfish monitoring results

3. Depending on the date of the sanitary survey, updated data should be sought to ensure most recently available data are used. These will form the basis of the hydrographic study.

For modelling of norovirus impacts specifically, the treatment level applied to sewage effluent is important. Tertiary treatment, and particularly UV disinfection, significantly decreases detectable norovirus loadings. Treatment works applying lower levels of treatment would be more likely to pass viable norovirus into the effluent stream. Of primary concern is the operation of Combined Sewer Overflows, which allow partially-or untreated effluent to bypass the sewage works and discharge to the environment. Climate change may exacerbate these overflows in areas affected by more frequent and heavy rainfall than the sewage systems were originally designed to handle.

5.2. Step 2: Dilution/dispersion studies and modelling

The sanitary survey alone will not provide sufficient information with which to develop buffer zones.

Two study areas were identified for the project in order to provide different case studies for development of the concept: the Upper Fal Estuary in the southwest of England and Alfacs Bay on the Mediterranean coast of Spain.

5.2.1. The Upper Fal Estuary

The Fal Estuary (Cornwall) is a flooder river valley with extensive drying areas along its margins and tributaries. The mean tidal range at Truro, on the upper estuary, is 3.5 m on spring tides and 2.4 m on neap tide, and the greatest risk of microbiological contamination comes from upstream catchment

Grant agreement: 773400

sources around Truro and Tresillian. Truro STW discharges UV-treated effluent and has a population equivalent of 25,000. The main focus of the buffer zone study is on the mussel production area at King Harry Reach (Fig. 1).

Fig. 1. The Upper Fal Estuary sampling locations

The sanitary surveys for this production area indicated that the main point sources of faecal contamination of human origin impacting the King Harry mussel farm are sewage discharges from Truro (Newham) STW and storm overflows in and around Truro (Cefas, 2010; 2012).

The hydrographic characteristics of the upper Fal Estuary have been described in Cefas sanitary survey reports (2010, 2012). Briefly, the estuary is a flooded river valley with extensive drying areas along its margins and tributaries. The mean tidal range at Truro is 3.5m on springs and 2.4m on neaps. The greatest risk of microbiological contamination is from upstream catchment sources at Truro and Tresillian.

In the study area, mean *E. coli* values generally decrease from Malpas to Turnaware Bar, although they are sometimes higher at Ruan Pontoon than at Malpas (Cefas, 2010). *E. coli* contamination also increases from spring to autumn and following rainfall events due to the operation of storm overflows, land runoff, increased river loadings and resuspension of contaminated sediments (Cefas, 2010).

Samples were taken from these between March 2018 and October 2018 and tested for *E. coli*, norovirus genogroup I (NoV GI) and norovirus genogroup II (NoV GII). Results of the *E. coli* testing by site over the period are shown in Fig.2.

Grant agreement: 773400

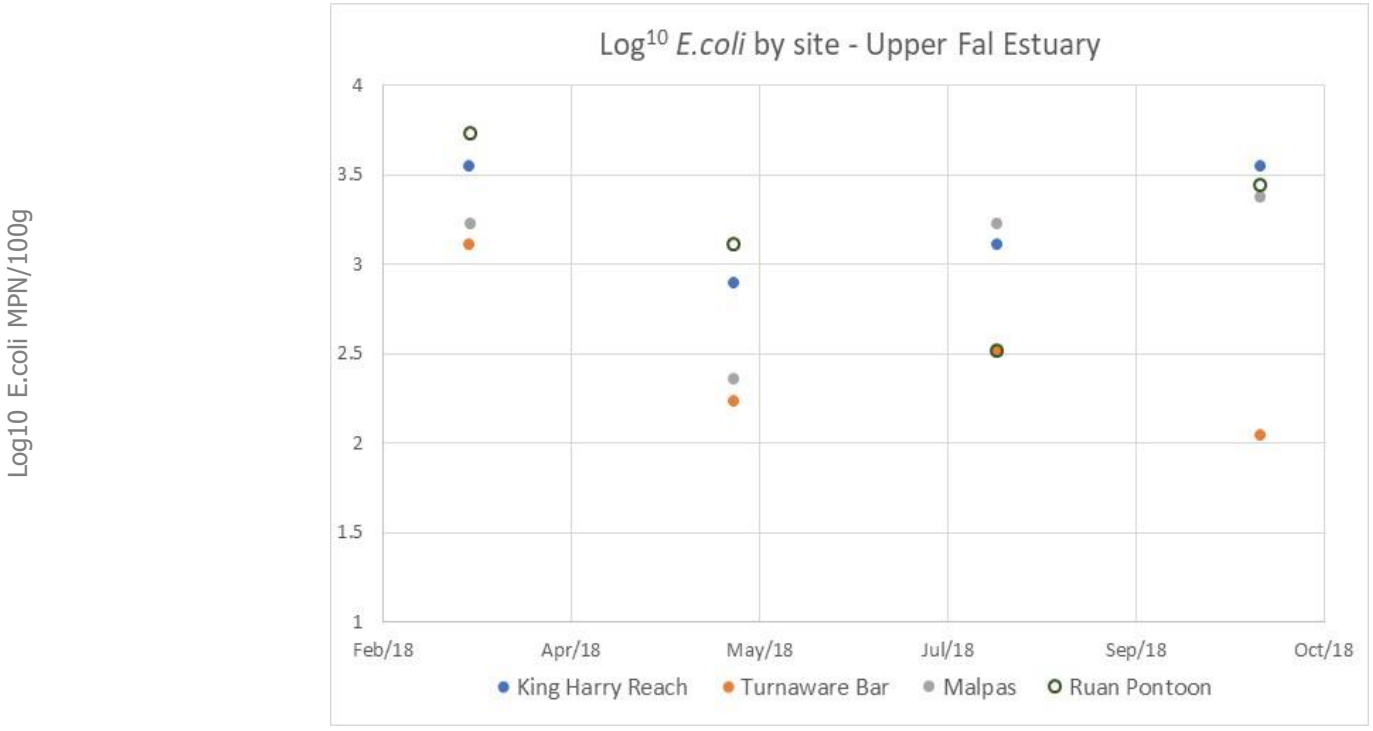
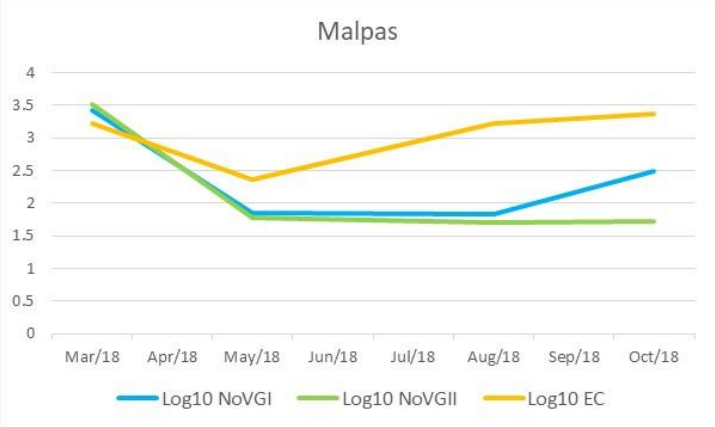


Fig. 2. *E. coli* results for the Upper Fal Estuary sampling locations

Results of samples taken Turnaware Bar were consistently the lowest, and this sample point is the furthest downstream from the Truro STW discharge. There was little consistency to be found in results from the other three sampling points; however, the sample size was limited. Fig. 3 compares results of all testing by site for the four sample locations.

E. coli results remained largely well above those for Nov GI and GII over the sampling period, though NoV results had begun trending upward at most sites by October. Samples taken from King Harry Reach, however, had not shown any signs of increased NoV content, even though there was an uptick in the GI result for Turnaware Bar downstream.



Grant agreement: 773400

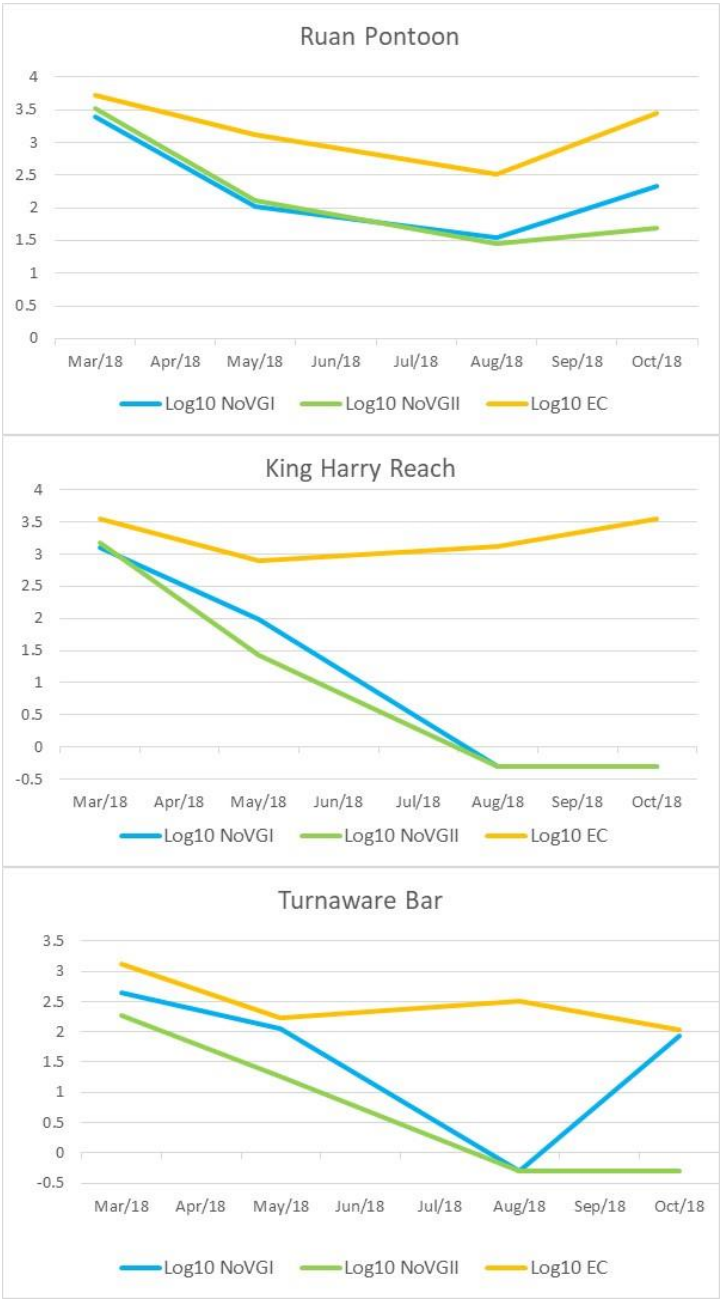


Fig. 3. Testing results for the Upper Fal Estuary by site

Due to objections from the local shellfish industry regarding the potential of dye taint to shellfish stock during the harvesting season, it was not possible to undertake a dye tracing study in the Upper Fal Estuary during the oyster harvesting season without significant reputational risk. Consideration was given to use of phage tracers, but it was not possible for Cefas to produce these in the volumes that would have been required at the time.

The decision was therefore taken to undertake hydrodynamic modelling of the upper Fal using existing data from other sources in order to determine potential dilution and dispersion of faecal indicators and norovirus to support development of the buffer zone protocol.

Grant agreement: 773400

5.2.2. Alfacs Bay

A more comprehensive study was conducted at Alfacs Bay, Catalonia, Spain, which was the second site chosen for this project. This site was selected in partnership with the Institute for Agrifood Research and Technology (IRTA) in Spain, who are partners in SeafoodTomorrow. Alfacs Bay has been extensively studied and IRTA were interested to confirm suspected sources of norovirus and how they impacted across the bay. Three sampling sites were selected to represent different pollution impacts in the bay; two were points already in use under the official control monitoring programme and a third was selected to represent reported contamination from the Poble Nou Delta. These sites are shown in Fig.4.

Shellfish samples were collected by IRTA from these sites on 15 sampling occasions during the year and the glands removed and sent to Cefas for norovirus analysis. Additional samples were collected at the same sites analysed at IRTA laboratory for *E. coli* testing. This sampling was carried out during the period March 2018–February 2019 to provide information on the prevalence of NoV contamination in the bay.

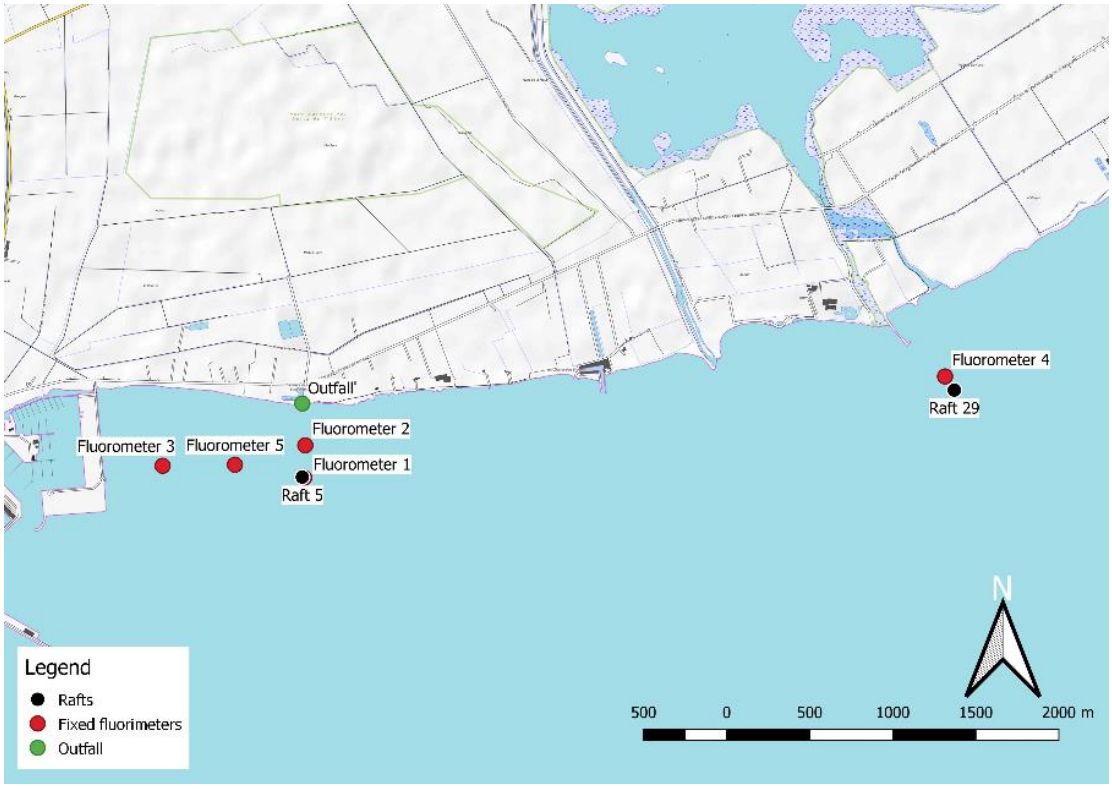


Fig. 4. Locations of sampling stations at Alfacs Bay.

The results of testing from all sites showed consistently higher *E. coli* contamination at Poblenou (Fig. 5).

Grant agreement: 773400

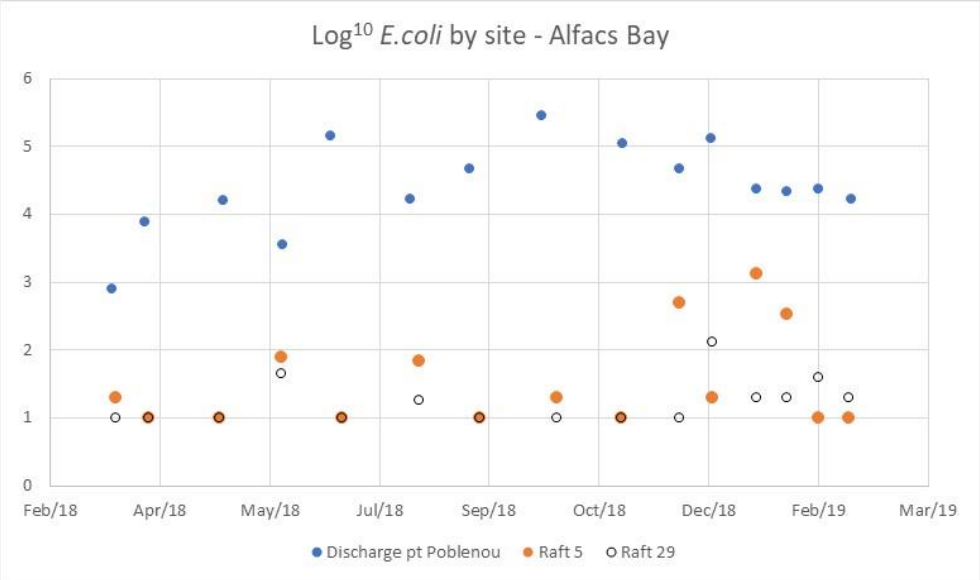
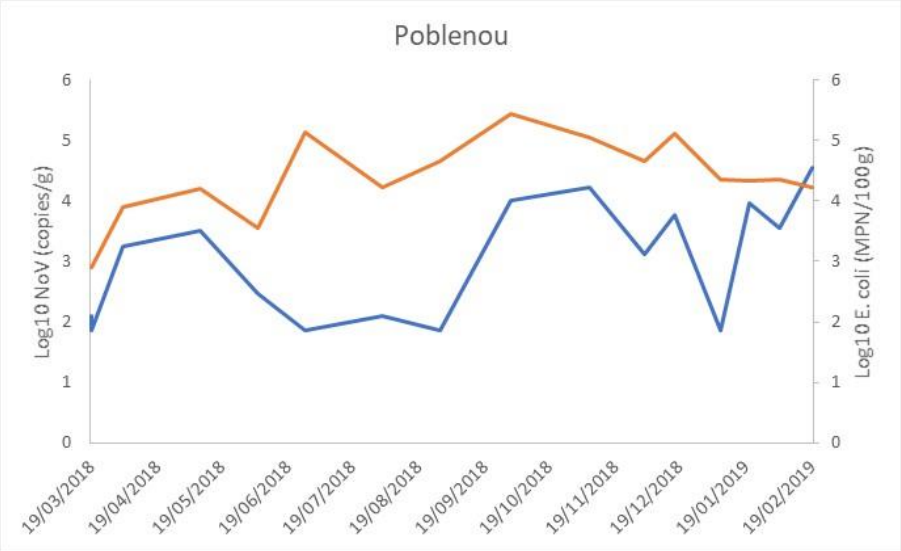


Fig. 5. Results of *E.coli* monitoring by site at Alfacs Bay

A look at the trends by site identified some interesting differences between them. Overall, contamination levels appeared to decrease in order of Poblenou>Raft5>Raft29. At both Raft 5 and Raft 29, there were clear peaks in norovirus results for both genogroups around December and January (fig. 6).



Grant agreement: 773400

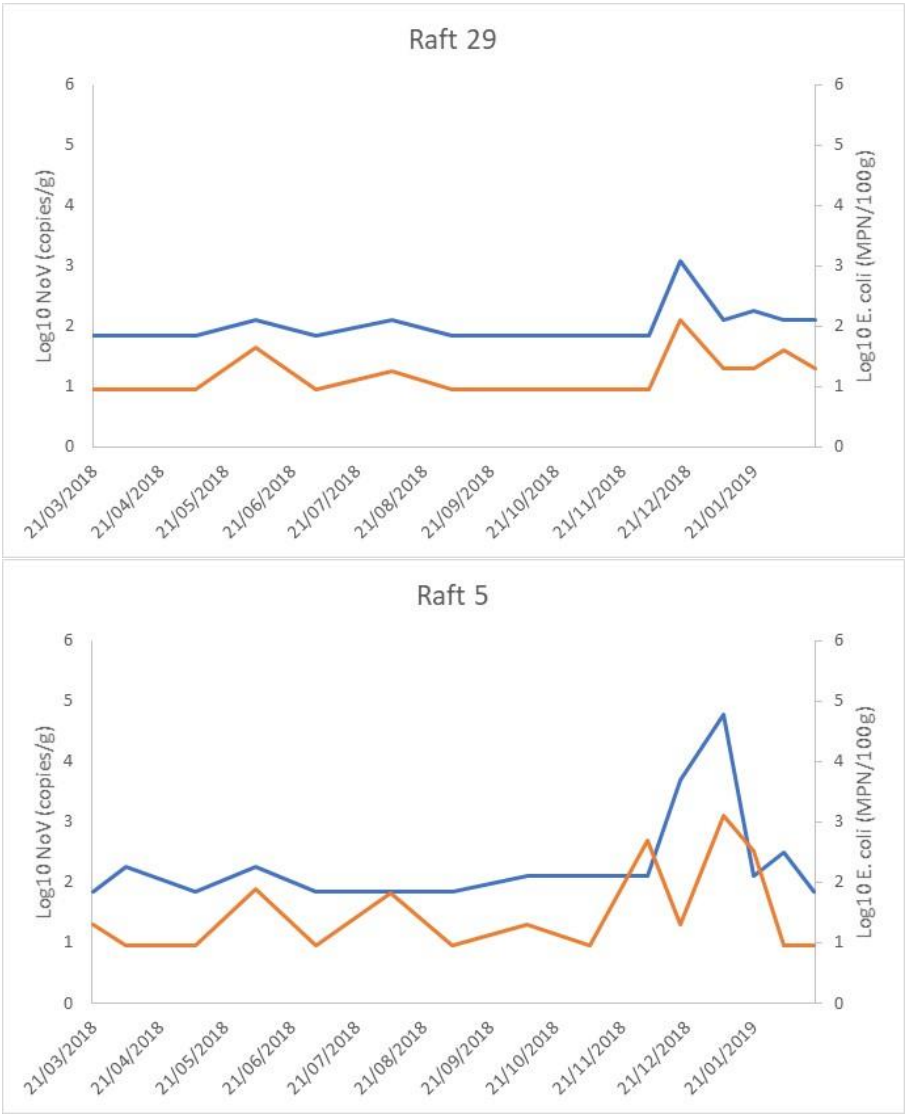


Fig. 6. Sample results by site at Alfacs Bay; where gold line represents *E. coli* and blue line represents total norovirus results.

At Raft 29, there was a clear match between peaks in *E. coli* and total Norovirus (GI + GII). Results at the other two monitoring stations were both higher and more variable. The results make it impossible to suggest that Poblenou is the definitive source of norovirus contamination, although it does appear that the site is generally subject to higher levels of faecal contamination than the other two sites. Both Raft 5 (nearest to Sant Carles de la Rapita STW) and Poblenou were subject to greater contamination from human faecal sources during the study period. At Poblenou, higher *E. coli* results than norovirus results could suggest faecal contamination there is arising from both human and non-human sources. Results overall suggest that both Poblenou and the STW at Sant Carles de la Rapita contribute to norovirus contamination in the bay.

During preparations for the field studies, we were informed by IRTA that the planned dye injection site near Poblenou had been blocked and was no longer flowing. At least some flow is needed to carry the dye to the bay, so a decision was taken, in consultation with IRTA, to move the location of dye injection to the San Carles de la Rapita STW, which is approximately 4km west of the original site. This would

Grant agreement: 773400

inform the hydrographic modelling but would unfortunately not directly relate to the suspect source discharge at Poblenou (see Fig 4).

The methods used in the dye tracing study are detailed in Campos *et al*. (2017). The dye tracer used in this study was Rhodamine WT because of its benign character in the aquatic environment, ease of use, low adsorptive tendency, strong fluorescence and chemical stability. This tracer is certified by the UK Marine Management Organisation for use in water quality studies (MMO 2014).

Dye standards were prepared by serial dilution (100,000 ppm to 0.1 ppb) to calibrate the WET Labs fluorometers used to monitor dye fluorescence in estuary.

Deionised water was added to the dye to create a 1:2 dye dilution mixture to facilitate the pumping of dye. The dye mixture was injected from a large plastic holding bin into the final effluent plant at a constant rate over 12.4h using a Masterflex variable speed peristaltic pump (Cole-Palmer) and Masterflex Tygon tubing.

Two approaches were used to measure dye concentrations in the estuary as follows:

- One fluorometer with internal battery and memory was attached to each of the four shellfish cages at the locations identified in Fig 4. These fluorometers were set up to start recording background fluorescence data one day before the dye injection and left in the water to monitor dye concentrations for a period of 7 days.

- On the day of the dye injection, one fluorometer was towed from a boat to identify the edges of the dye plume as it travelled down the estuary and thus measured dye fluorescence over a larger area of the estuary.

The dye measurements obtained at the fixed cage locations was used to determine the build-up and the steady-state dilutions (i.e. when the rate of effluent entering the bay equals that at which the effluent is flushed by the tides) of sewage contamination in the estuary. The dye measurements obtained with the tracking fluorometer were used to map areas impacted by the sewage plume.

Water temperature and salinity data were recorded at the shellfish cage locations during the tracing studies using mini loggers attached to the cages. A tracking CTD was also used in conjunction with the towed fluorometer on the day of the dye injection.

Planned locations for the fixed fluorometers were relocated accordingly to take account of the change to dye injection location. The locations of the sampled sites at Alfacs Bay are shown in Fig. 4.

The dye tracing exercise was originally scheduled to take place at the end of February 2018 with participation of experts from the United States Food and Drug Administration (USFDA), who developed the methodology planned for use. Unfortunately, availability of the USFDA expert was impacted by the shutdown of the US government at the end of February, and the field work was rescheduled for March 27-31, 2019. In the end, USFDA experts were not able to participate and the Cefas team undertook the study with assistance from IRTA. On the scheduled day of dye release, wind conditions constrained the dye plume along the shoreline after release (Fig.7). Whilst these conditions were safe for survey, they did not represent the steady-state conditions the project had hoped to target.

Grant agreement: 773400

Fig. 7. Dye tracing at Alfacs Bay showing dye plume at shore

All data from the dye study have been used to inform the hydrodynamic model and have been made available via the project data repository.

Step 3: Hydrodynamic modelling

5.2.3. Introduction of Upper Fal Estuary

This section describes the hydrodynamic modelling of the Upper Fal Estuary in South West England, as part of the H2020 Seafood Tomorrow project. The models aim to simulate the release of effluent from the combined sewer overflow (CSO) via the storm tank outfall, under high flow conditions, at Newham STW. In addition to the storm overflow contributing to the microbiological load in the estuary which affect downstream shellfish growing areas, the waters of the Estuary are impacted by excessive agricultural runoff and nutrient enrichment from the CSO leading to periodic eutrophic conditions and algal blooms. The model simulates creates CSO releases and predicts the temporal and spatial distribution, and the decay of *E. coli* concentration in the water column as a function of the physical parameters applied by the model. The effluent releases are made in a set of scenarios which reproduce a range of tidal conditions in the Upper Fal Estuary, through the Rivers, Truro, Tressilian and Fal.

5.2.4. Modelling system

The TELEMAC hydrodynamic modelling system is a modular, hydro-informatics system currently used by up to 4,000 individual scientists and engineers worldwide. The modelling code is published, maintained and released by Electricity de France and the system managed by scientists from a consortium of Scientific Institutes across Europe (Open TELEMAC-MASCARET,- 2019 - Available at: http://www.opentelemac.org/).

The strength of the TELEMAC system is based upon its adoption of a finite element numerical scheme (Hervouet 2007) and the ability to create model domains using an unstructured grid, enhancing and relaxing spatial resolution to aid computational efficiency. The system is modular giving a wide range of

Grant agreement: 773400

capabilities from a simple simulation of depth averaged (2D) tidal flow without baroclinic functions, to a fully resolved three-dimensional flow including temperature, salinity and fully developed stratification throughout an estuarine environment.

For this study a 2D model was constructed initially and used to scope the parameters for a full baroclinic model in which the effluent releases were made.

5.2.5. Model domain

The Upper Fal Estuary is a Ria, a drowned river valley and its formation and geographical setting are fully described in the sanitary survey (Cefas 2012). The estuary is classified as macro-tidal with a mean spring tidal range of 4.6 m measured in Carrick Roads, at the estuary mouth.

The model covers the area of the Truro, Tressilian and Fal river channels, Ruan and Callenick Creeks, Carrick Roads and Falmouth Bay. The model includes a sea boundary which extends in an arc from approximately Wilkins Point in the west, to St Anthony's head on the east side of the Bay opening.

The mesh of the model domain has a resolution of 500m at the open sea boundary reducing to 20 m along the shoreline of the upper estuary, the rivers and through the Bivalve production area. This gives the mesh 15,376 nodes.

Two sources of bathymetry data were acquired; for the offshore region and outer Fal estuary data from the DEFRA UKSeaMap (McBreen, et al. 2011) providing point data at a resolution of 1 arc-second (~30m) and for the upstream regions and river channels from the UK Environment Agency Lidar 2m resolution DTM (https://environment.data.gov.uk/hydrology/landing, 2019). The datums of these two datasets are different, UKSeaMap is related to Chart Datum and EA Lidar data related to OSGB Ordnance Datum Newlyn.

Hydrodynamic forcing to drive the model is provided at the sea boundary where 11 tidal constituents are applied (M2, S2, N2, K2, K1, O1, P1, Q1, M4, MS4, MN4) from the OSU TPXO European shelf 1/30o regional model. The boundary has values for sea surface elevation and tidal velocity, interpolated along the telemac boundary from the regional model. Sea surface elevations used for the duration of the model run are illustrated in Fig. 8.

Grant agreement: 773400

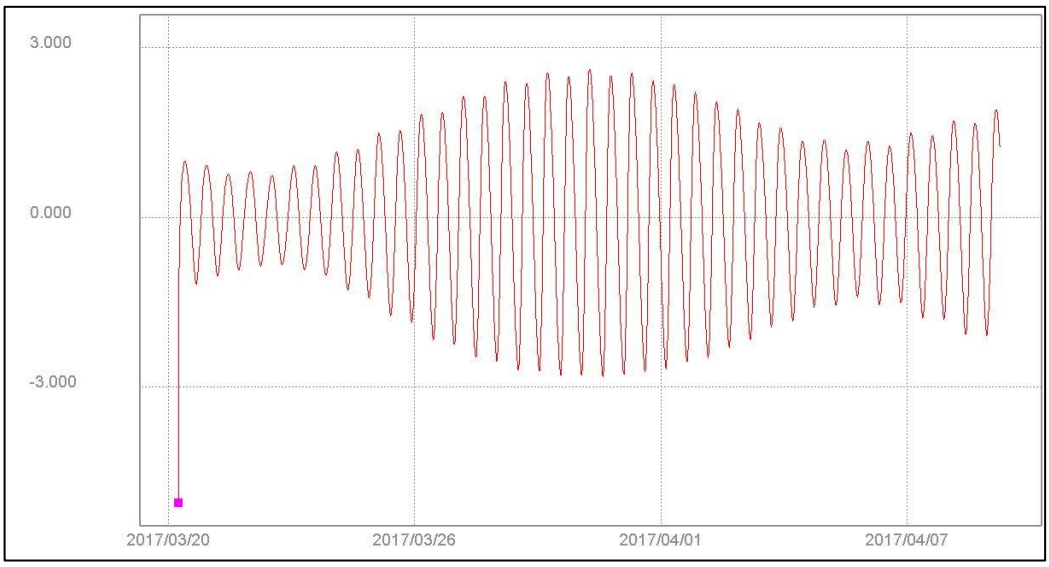


Fig.8 - Sea Surface elevation mid channel in Carrick Roads over the duration of the full model run - Lon 5.043 deg W Lat 50.189 deg N. The datum of sea level is adjusted to suit the value of Mean Sea Level specified by the TPXO model data base values, used to drive the offshore boundary sea level curve.

The vertical datum of these three datasets (UKSeamap, EA Lidar and TPXO) are different and require alignment. UK SeaMap is above Chart Datum (CD), the Lidar data related to Ordnance Datum Newlyn (ODN) and TPXO to mean sea level (MSL). The conversion from ODN to CD for Newlyn of +3.05m was used and a further offset of -3.032m applied to relate the model directly to MSL and the TPXO forcing. Bathymetry of the study area is shown in Figures 9 and 10.

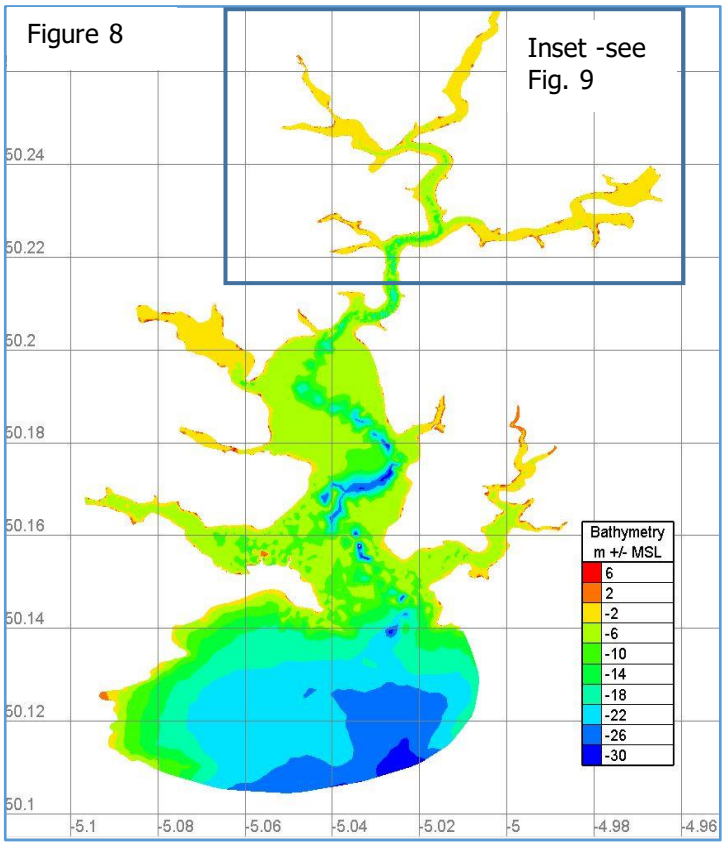
Freshwater river inputs are applied at the head of Truro, Tressilian and Fal rivers with data for average flow in April 2017 from the EA Hydrology portal. A value for Calenick Creek was also added from information supplied by the Environment Agency. The rivers are applied as point sources in the model as opposed to full discharge boundaries. The benefit of this method is that a simpler setup is achieved although riverine discharges remain constant in time throughout the model runs as time varying data was not readily available.

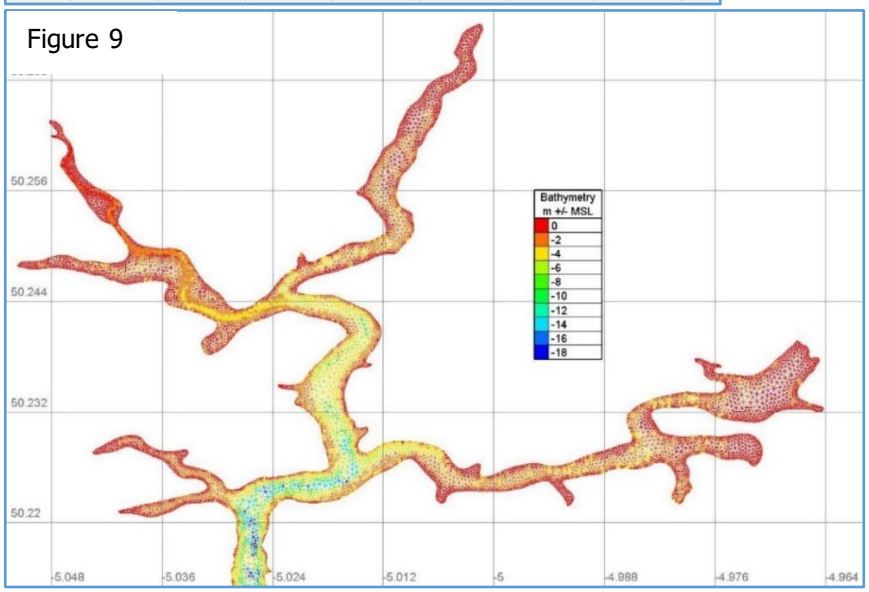
Table 1 - Table of freshwater and effluent sources applied in the model

Discharge Location	Discharge rate
River Truro	0.377 m^3s^{-1}
River Tressilian	1.2 m^3s^{-1}
River Fal	2.025 m^3s^{-1}
Newham STW effluent	0.127 m^3s^{-1}
Calenick Creek	0.3 m^3s^{-1}

The discharge at the Newham STW is taken from FFT values provided by South West Water of an average of 11,000 m3 day-1 which equates to 0.127 cumecs of total organic load.

Grant agreement: 773400





Figures 9 &10- Bathymetry and model mesh – Fig. 4 is the detail of Upper Fal Estuary; the resolution of the mesh in this region reduces to ~ 20m at the shoreline. A channel mesh is introduced from Truro town to the confluence with the River Tressilian to maintain continuity of flow in the model over the shallow wetting and drying areas upstream.

5.2.6. Parameterisation of the model

All TELEMAC3D runs use an internal timestep of 1 second with main variable outputs saved at 30 minute intervals. Turbulent energy and turbulent dissipation are modelled using a k-e approach in both horizontal and vertical planes. Bottom and lateral boundary friction is applied using the laws of

Grant agreement: 773400

Nikuradse with a roughness length selected from Soulsby (Soulsby 1997) 1997 of 0.7mm for estuarine muds.

The TELEMAC3D model is coupled with a water quality processes module which includes the atmospheric exchange. The effects of wind speed and direction, atmospheric temperature, humidity and pressure, nebulosity, precipitation and evaporation are included via an input file sourced each hour from the ECMWF ERA-5 model data for the duration of the model runs.

Salinity and temperature at the offshore boundary and applied as initial conditions across the domain are derived from the FOAM AMM7 ocean model NORTHWESTSHELF_ANALYSIS_FORECAST Phy 004 013, published by the Copernicus Marine Environment Portal from and downloaded for March and April 2017.

The TELEMAC3D model uses 4 layers in the vertical, equally spaced though the water column however the estuary is well mixed in the upper reaches with many areas drying completely for several hours each side of low water and stratification is limited.

5.2.7. Spin up and prep

The four-week period of 20/03/2017 to 20/04/2017 was determined for the hydrodynamic model runs. A TELEMAC2D setup was run initially to scope tidal conditions (surface elevation, tidal range and velocities) in the upper and lower estuary and compared with existing data from the studies carried out in CES Ltd, H. R. L. (CES, Ltd 1989). These weeks in March and April coincide with periods of approximately equal periods of light through the day and night which is used in the parameterised version of the decay rate for tracers (T90) applied to *E. coli* in TELEMAC2D.

5.2.8. Scenarios

E.coli

The *E. coli* released into the flow at the Newham STW is simulated by a series of high-volume releases from the CSO over 12 and 24 hours. Low water neap tide releases are designated with suffix EC and Spring tide releases with suffix EB. High water neap releases suffix FC and springs with FB.

For *E. coli*, an implicitly varying T90 value was applied based on the research from Huang et al (Huang, Falconer and Lin 2017). Values of T90=8.56 h for day-time and 30.64 h for night-time were applied. For each scenario a background value of 100 cfu/100ml was applied to the *E. coli* tracer at the discharge representing the value emitted from a discharge with UV filter operational. For the discharge period the concentration was increased to 2.5x106 cfu/100ml representing the combined storm overflow.

Norovirus

The model was amended to simulate the release of 7.5x104 copies/100ml of a passive tracer representing Norovirus G1+G2 from the Newham STW. For this emission the tracer was treated as completely passive with no decay or other behaviour programmed in the model. The material is diffused and advected through the water column by the hydrodynamic functions of the model only.

Grant agreement: 773400

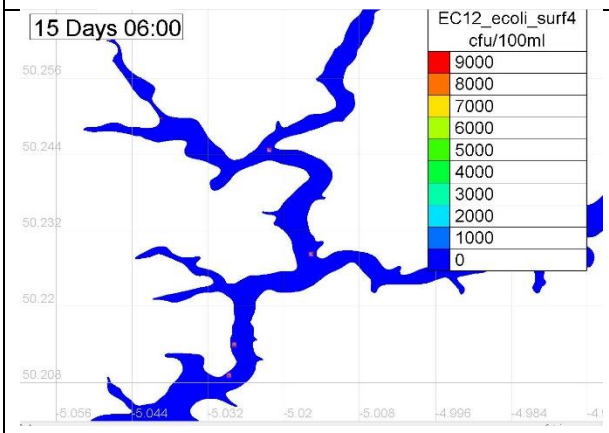
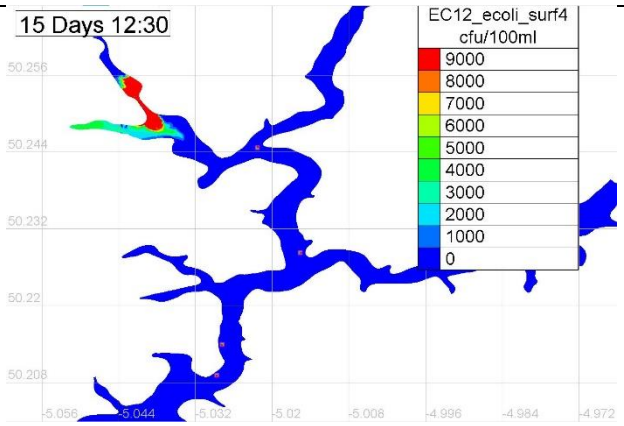
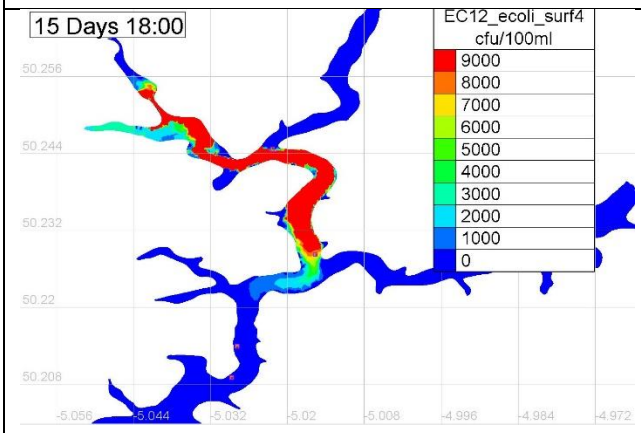
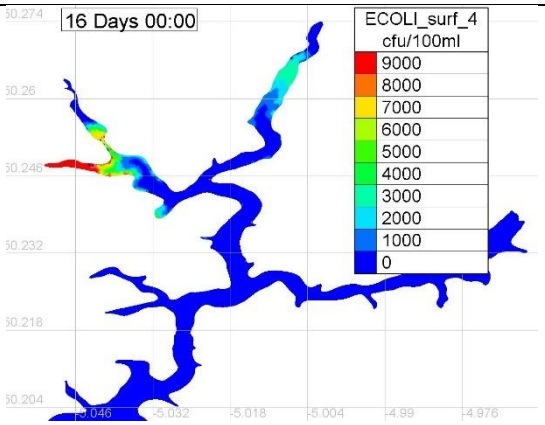
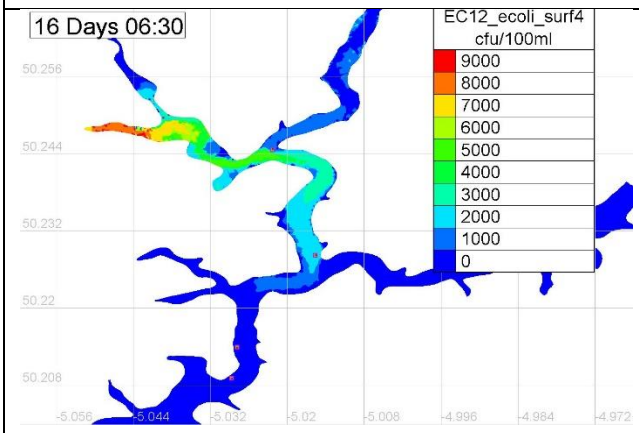
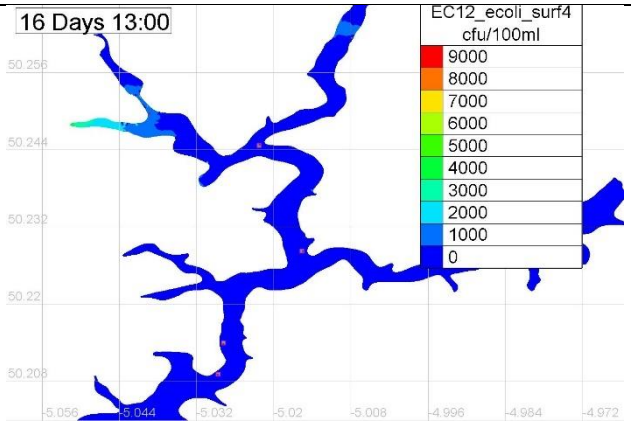
Natural dissipation in the tidal flow by advection and turbulence degraded the material during a 7-day period the decay value for the virus in water is not currently understood. By ignoring a decay rate, the values transmitted to the production area are considered to be conservative in terms of the exposure at the sampling locations.

Table 2 - Summary of Scenarios run and spill durations

Tide Type at Release	Tracer	Duration of CSO spill	Max Concentration	Results File Reference
LW Neap	*E. coli*	12 hour	$2.5x10^6$ CFU/100 ml	EC_r3d_12hr.slf
LW Neap	*E. coli*	24 hour	$2.5x10^6$ CFU/100 ml	EC_r3d_24hr.slf
LW Spring	*E. coli*	12 hour	$2.5x10^6$ CFU/100 ml	EB_r3d_12hr.slf
LW Spring	*E. coli*	24 hour	$2.5x10^6$ CFU/100 ml	EB_r3d_24hr.slf
HW Neap	*E. coli*	12 hour	$2.5x10^6$ CFU/100 ml	FC_r3d_12hr.slf
HW Neap	*E. coli*	24 hour	$2.5x10^6$ CFU/100 ml	FC_r3d_24hr.slf
HW Spring	*E. coli*	12 hour	$2.5x10^6$ CFU/100 ml	FB_r3d_12hr.slf
HW Spring	*E. coli*	24 hour	$2.5x10^6$ CFU/100 ml	FB_r3d_24hr.slf
HW Spring	Norovirus	12 hour	$7.5x10^4$ copies/ml	GB_r3d_12hr.slf
HW Spring	Norovirus	24 hour	$7.5x10^4$ copies/ml	GB_r3d_24hr.slf
HW Neap	Norovirus	12 hour	$7.5x10^4$ copies/ml	GC_r3d_12hr.slf
HW Neap	Norovirus	24 hour	$7.5x10^4$ copies/ml	GC_r3d_24hr.slf
LW Spring	Norovirus	12 hour	$7.5x10^4$ copies/ml	HB_r3d_12hr.slf
LW Spring	Norovirus	24 hour	$7.5x10^4$ copies/ml	HB_r3d_24hr.slf
LW Neap	Norovirus	12 hour	$7.5x10^4$ copies/ml	HC_r3d_12hr.slf
LW Neap	Norovirus	24 hour	$7.5x10^4$ copies/ml	HC_r3d_24hr.slf

The point values of the *E. coli* and norovirus tracers are exported as time series at each sampling point at the surface and bed (layers 4 and 1) of the 4 layer TELEMAC3D model. The analysis and comparison of these time series examines the duration of time (and percentage of time) the exposure at each sampling point exceeds a set of threshold values.

Grant agreement: 773400

Neap – 12 hr Release from Newham STW on 05-04-2017 (Day 15) at 06:00 (T+15:06:00)	
Low water	High Water
	
	
	

Grant agreement: 773400

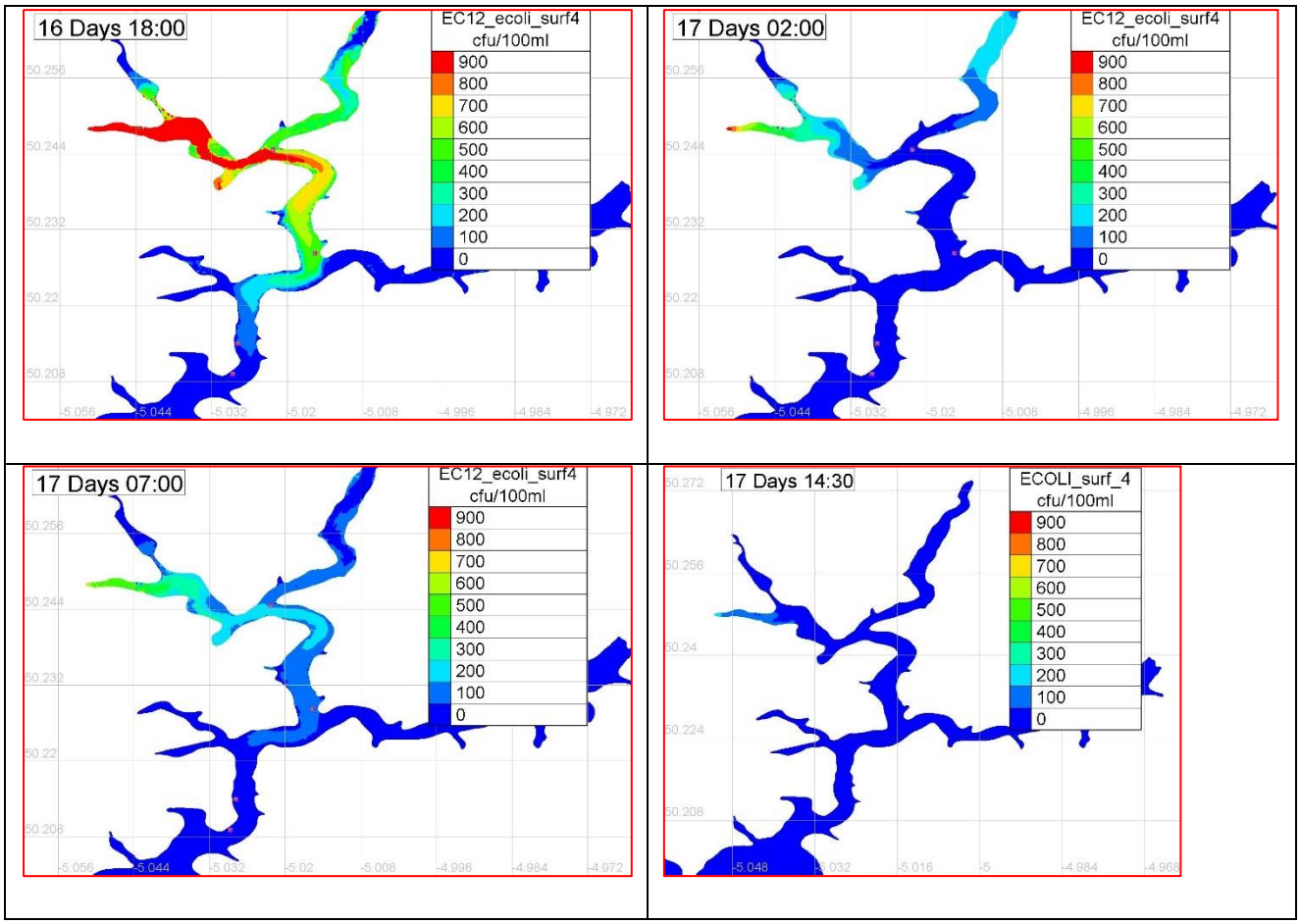


Fig 11 Image snapshots of *E. coli* concentration in the surface layer from LW1 (start of a 12 hour release) to HW 10 at subsequent alternating maximum high and low water during and after release. Note the re-scaling of the legend in the last 4 images, reflecting decayed values of *E. coli* in the water column.

The model was amended to simulate the release of tracer representing Norovirus from the Newham STW. For this the tracer was treated as completely passive with no decay or other behaviour programmed. The material is diffused and advected through the water column by the hydrodynamic functions of the model only.

5.2.9. Sampling station analysis

Scenarios used for the assessment of norovirus dispersion are identified in Table 4 and box plots of predicted norovirus loadings at both the surface and the seabed at the different study stations are presented in Fig. 11.

Grant agreement: 773400

Table 3 – Scenarios used in norovirus modelling

Scenario	Tracer	Type Phase	Release time	Release Duration	Release quantity
GB12	NoV G1+G2	Spring	High Water	12 hrs	75000 copies/ml
GB24	NoV G1+G2	Spring	High Water	24 hrs	75000 copies/ml
GC12	NoV G1+G2	Neap	High Water	12hrs	75000 copies/ml
GC24	NoV G1+G2	Neap	High Water	24 hrs	75000 copies/ml
HB12	NoV G1+G2	Spring	Low Water	12 hrs	75000 copies/ml
HB24	NoV G1+G2	Spring	Low Water	24 hrs	75000 copies/ml
HC12	NoV G1+G2	Neap	Low Water	12 hrs	75000 copies/ml
HC24	NoV G1+G2	Neap	Low Water	24 hrs	75000 copies/ml

Very little difference was seen between modelled effects at the surface and seabed at the four identified sample points.

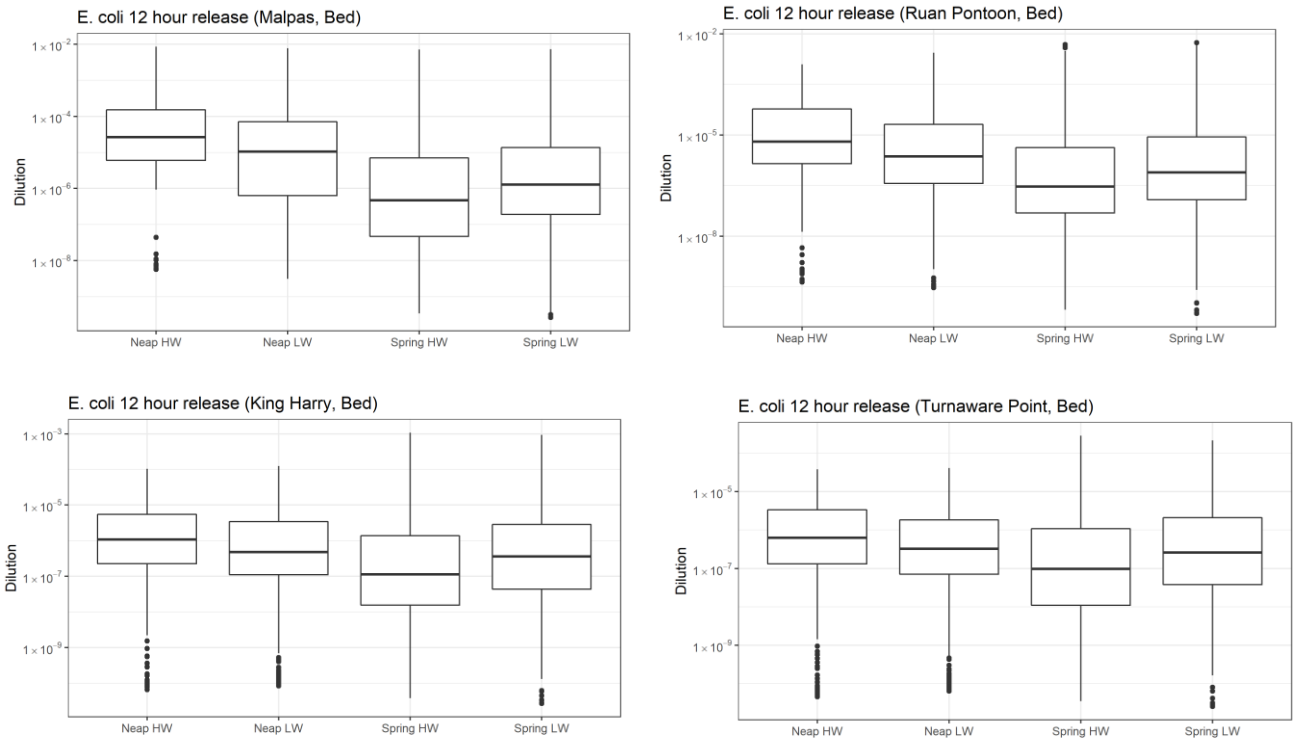


Fig 12 Dilution at all stations for 12 hour CSO release

At all stations, there was an increase in the dilution observed when releases occurred during HW springs. The differences were more pronounced the higher up the estuary the station, so that at Turnaware Point (most downstream) there was very little variation in dilution between the different tidal states at release. This is consistent with the changes in depth and area available for dilution lower down the estuary. At Malpas, the minimum modelled dilution achieved was $1x10^{-2}$ across all release states, though the median dilution for releases on spring tides achieved the target dilution of $1x10^{-5}$.

Grant agreement: 773400

Table 4 – Modelled geomean concentrations at each station for 12 and 24 hour CSO releases

	Malpas Surface	Ruan Surface	King Harry Surface	Turnaware Surface	Malpas Bed	Ruan Bed	King Harry Bed	Turnaware Bed
E. coli Releases 2.5 x 10^6 CFU/100ml - 4.75 day post release								
Spring Tide - Low Water Release								
EB12	3.66	2.64	0.93	0.71	3.86	2.64	0.93	0.70
EB24	12.76	9.72	3.38	2.53	13.17	9.73	3.37	2.53
Spring Tide - High Water Release								
FB12	3.48	2.46	0.84	0.62	3.57	2.44	0.83	0.62
FB24	10.50	7.44	2.46	1.79	10.74	7.45	2.45	1.79
Norovirus Releases 7.5 x10^4 copies/ml - 7 day post release								
Spring Tide - High Water Release								
GB12	4.65	3.37	1.57	1.20	4.78	3.53	1.58	1.20
GB24	10.00	7.19	3.13	2.33	10.29	7.57	3.14	2.33
Spring Tide - Low Water Release								
HB12	6.35	3.50	1.47	1.16	6.47	3.48	1.47	1.16
HB24	13.13	7.17	2.78	2.16	13.37	7.12	2.79	2.16

	Malpas Surface	Ruan Surface	King Harry Surface	Turnaware Surface	Malpas Bed	Ruan Bed	King Harry Bed	Turnaware Bed
E. coli Releases 2.5 x 10^6 CFU/100ml - 4.75 day post release								
Neap Tide - Low Water Release								
EC12	18.59	6.11	0.90	0.51	18.95	6.24	0.93	0.52
EC24	54.59	16.46	2.05	1.10	55.45	16.77	2.11	1.12
Neap Tide - High Water Release								
FC12	59.11	15.07	1.43	0.72	60.90	15.50	1.48	0.73
FC24	158.76	35.95	2.86	1.39	164.41	37.37	2.97	1.42
Norovirus Releases 7.5 x10^4 copies/ml - 7 day post release								
Neap Tide - High Water Release								
GC12	19.36	12.46	6.61	5.41	19.30	12.58	6.65	5.42
GC24	42.90	18.97	3.93	2.42	43.63	19.33	4.05	2.45
Neap Tide - Low Water Release								
HC12	16.36	5.55	1.47	0.99	17.01	5.61	1.49	0.99
HC24	31.30	10.11	2.49	1.64	32.58	10.24	2.53	1.65

For *E. coli* dispersion is significantly higher under a release during a spring tide, with very little difference between a low water or high water release. For the neap tide releases, much more material remains in the zone as advection over the tidal cycles is significantly is reduced.

Norovirus dispersion is much reduced and therefore longer periods were considered in the analysis (7 day post release against 4.75day for *E. coli*). No decay was applied to the norovirus dispersion, so a 7 day cutoff was applied to the model. The time for recovery to background values was assessed for *E.coli*.

Grant agreement: 773400

In each scenario, concentrations remained marginally higher at the bed as transport is slightly reduced due to friction and the buoyancy of the slightly fresher water at the surface. However, these effects are small as the water column in the upper estuary remains well mixed, particularly in areas where tidal flats exist at low water and wetting and drying occurs.

To get a geomean calculation of norovirus between the release point and the time of first value at the stations, a value of $1x10^{-6}$ was applied as using a geomean of zero results in computation error.

5.2.10. Alfacs Bay modelling

The same methodology was used to develop a hydrodynamic model for Alfacs Bay. At the time of reporting, this model had not yet yielded successful output. Modelling the influx of freshwater in the bay has repeatedly caused the program to fail and as yet no suitable fix has been identified.

If the assumption of no freshwater input to the bay is made, long residence times for any sewage input would be expected due to the shallow depth and minimal tidal flushing in the bay. There is potential for impact from norovirus, when present, across most if not all of the bay. The spread of this would be highly dependent on wind direction and strength, as in the absence of freshwater this is the main driver of water circulation in the bay.

Further attempts are being made to resolve this model and the resulting data and output will be made available via the data repository for the project.

5.3. Developing and implementing Buffer zones

Proposed dilution guidance for prohibited zones associated with wastewater discharges (Interstate Shellfish Sanitation Conference 2014) to the USFDA identified that under a worst-case scenario of a raw sewage release to an estuary, a dilution of 1:100000 be achieved. Where dilution is greater than this, the STW would be considered to not impact the shellfishery.

Therefore, using the model developed for the Upper Fal Estuary for *E. coli*, the impact at the four monitoring points would be as follows under the scenarios identified in 5.3.3.

Raw sewage from a 12-hour CSO discharge at low water on a neap tide (worst case scenario) would impact at the upper two monitoring points, Malpas and Ruan Pontoon, within 12 hours. Elevated levels of microbial contamination would persist over 4 tidal cycles at these two sites before returning to background levels.

Table 4 shows the predicted elevation of *E. coli* levels at all four sampling points at the modelled end points. Impact at the shellfisheries will be dependent on the state of tide at which the discharge occurs and also how much of the tidal cycle the shellfish at each location remain submerged and filtering.

Different scenarios could be modelled based on the actual duration of any spill to identify at what point contamination levels in the water would be expected to return to background levels. As long as telemetry was in place to allow the managers of the STW to alert industry and regulators to a spill condition within 6 hours, the fishery could be closed before any impact reached the shellfish.

Grant agreement: 773400

Criteria for reopening are a matter not resolved by this study. These depend largely on the ability of the shellfish to depurate contaminants accumulated during the spill.

6. Resources required

6.1. Data resources

This type of approach is highly data-dependent and will require access to a number of sources of data relevant to model input and development. It is important to have a good understanding of the physical characteristics of the fishery, which should be set out in the sanitary survey.

In addition to this, it will be necessary to have access to Geographic Information Systems (GIS) software, statistical analysis and representation software (eg. R, SPSS, Minitab), TELEMAC modelling capability, as well as the expertise required to utilise these tools. Hydrodynamic models are, by their nature, specific to the bathymetry, terrain, and hydrography of the area for which they are built. Large scale models covering large sea areas do not provide the spatial resolution required to describe the near-shore processes found around most shellfisheries and therefore many areas would benefit from specific models.

6.2. Other resources

If there is insufficient information available regarding the movement of contaminants in an area, or if it is not clear whether effluent from a sewage source would impact a shellfishery, a source-tracing study may be undertaken as outlined is the example in Section 2. Rhodamine dye was used as a tracer is this study, but other types of tracers may also be suitable to this type of study. Tracer studies are by nature resource intensive exercises and require a suitably trained team with access to work boats, fluorometers, CDT probes, and the associated software to operate them and record data.

Once a model has been developed for a particular water body, this should be validated against monitoring data to ensure it remains representative over a range of real-life conditions. This will require sampling equipment and laboratories capable of carrying out accredited testing for relevant bacteria and viruses in support statutory shellfish monitoring.

7. Reporting and communicating information

7.1. Reporting CSO releases

A key part of using the buffer zone concept to manage fisheries subject to contamination from operation of CSOs is reporting of these to the relevant stakeholders in good time for them to take action. In the case of the Fal Estuary, this would ideally mean communicating information on a spill to the Local Enforcement Authority and relevant shellfish harvesters within 6 hours of initiation of a spill. This would allow for harvesting at the most impacted sites in the upper estuary to be ceased in good time to prevent placement of contaminated shellfish on the market.

Grant agreement: 773400

7.2. Reporting progress on buffer zone development

The data collected as a part of this project and the modelling output are all being made available via the project data portal. It is felt that the modelling protocol would benefit greatly from further development and validation, which we will seek to progress in future. Once a validated protocol is completed, this will be shared directly with stakeholders and regulators.

8. Conclusions

The model developed for the Upper Fal Estuary identified that in the event of a combined sewage overflow, microbial contamination from that source would be expected to impact a large part of the upper estuary, with the extent of impact depending on the duration of the spill and tidal state at the time. Greatest concentration of contaminants tends to occur at neap tides, when there is less flushing of the upper estuary. This tended to run counter to assumptions made at the beginning of the project, when a higher impact was expected to coincide with spring tides and stronger tidal flows.

Alfacs Bay was found to pose complications for the hydrodynamic model, which are still being rectified. Onshore winds at the time of the dye release entrained the dye near to shore, preventing it from reaching the shellfishery. The monitoring that had been undertaken for the study showed significant faecal impacts at the Poblenou monitoring station, whilst the other two stations were substantially cleaner. The bay receives heavily seasonal human impacts, with a strong summer tourist season and seasonal freshwater inflows tied to seasonal draining of the rice fields along the head of the bay.

There is currently insufficient evidence to suggest that sewage effluents affect HABs in a readily predictable way and no evidence that these arise in a manner that would support implementation of a buffer zone approach. An attempt at developing a neural network model for HABs in St. Austell by was unsuccessful and this was attributed to a paucity of observational data with which to train and then validate the model. This was an important outcome, as we had considered that the data available for St. Austell Bay would be fairly easily support this. It has certainly highlighted that modern technologies in remote sensing and modelling require dense observational datasets. The success of IRTA in developing a similar model for Alfacs Bay, where there were rich datasets on which to train the model, confirmed that this was possible. It will be useful for this model to be tested more widely in areas around Europe where phytoplankton alert thresholds are significantly different than those used for Alfacs. For example, the threshold alert concentration for *Pseudo-nitzschia* spp. vary by orders of magnitude between Alfacs and areas in the UK (2.000.000 cells L^{-1} at Alfacs Bay, 150.000 cells L^{-1} in England & Wales and 50.000 cells L^{-1} in Scotland). This is a key area for further development and validation of the model.

There is further potential for extending the knowledge gained in this study through the conduct of field studies during an actual spill event to validate that the hydrodynamic model for the Fal Estuary is able to adequately predict the water quality impact and an investigation into how long it takes shellfish to depurate contaminants taken up during the event, particularly with regard to norovirus.

Grant agreement: 773400

9. References

Anon. 2018. *Microbiological monitoring of bivalve mollusc harvesting areas - guide to good practice: technical application Issue 7.* December . Accessed 10 21, 2019. https://eurlcefas.org/media/14117/20181231gpg_issue-7-final.pdf.

Anon. 2017. "Community guide to the principles of good practice for the microbiological classification and monitoring of bivavle mollusc production and relaying areas with regard to Regulation 854/2004." *European Commission.* 03 January. Accessed October 21/10/2019, 2019. https://ec.europa.eu/food/sites/food/files/safety/docs/biosafety_fh_guidance_community_guide_bivalve_mollusc_monitoring_en.pdf.

Campos, Carlos JA, and David N Lees. 2014. "Environmental transmission of human noroviruses in shellfish waters." *Applied and Environmental Microbiology* 80 (12): 3552-3561.

Campos, Carlos JA, Gregory Goblick, Ron Lee, Ken Wittamore, and David N Lees. 2017. "Determining the zone of impact of norovirus contamination in shellfish production areas through microbiological monitoring and hydrographic analysis." *Water Research* 124: 556-565.

Cefas. 2012. "Sanitary Survey - Upper Fal Estuary, Cornwall." Accessed 10 21, 2019. https://www.cefas.co.uk/media/41315/final-fal-upper-estuary-sanitary-survey-report-2010.pdf.

CES, Ltd. 1989. *Enviromental survey and mathematical modelling of the Truro River system.* HRL.

Davidson, Keith, Richard J Gowen, Paul Tett, Eileen Bresnan, Paul J Harrison, April McKinney, Stephen Milligan, David K Mills, Joe Silke, and Anne-Marie Crooks. 2012. "Harmful algal bloom: How strong is the evidence that nutrient ratios and forms influence their occurrence?" *Estuarine, Coastal and Shelf Science* 115: 399-413.

EFSA. 2012. "Scientific opinion on norovirus (NoV) in oysters: methods, limits and control options." *EFSA Journal* (European Food Safety Authority) 10 (1): 2500.

European Communities. 2004. "Regulation (EC) No 854/2004 of the European Parliament and of the Council of 29 April 2004 laying down specific rules for the organisation of official controls of products of animal origin intended for human consumption." *Official Journal of the European Union* L226, 25.06.04: 83-127.

Hervouet, Jean-Michel. 2007. *Hydrodynamics of free surface flows.* Chichester: John Wiley and Sons Ltd.

Huang, Guoxian, Roger A Falconer, and Binliang Lin. 2017. "Integrated hydro-bacterial modelling for predicting bathing water quality." *Estuarine, Coastal and Shelf Science* 188: 145-155.

Interstate Shellfish Sanitation Conference. 2014. "Dilution Guidance for Prohibited Zones Associated with Wastewater Discharges." *Proposal No. 13-118.* Columbia, SC, 05 05. http://www.issc.org/Data/Sites/1/media/2013%20summary%20of%20actions/13-118%2006-10.pdf.

Grant agreement: 773400

Lees, DN. 2000. "Viruses and bivalve shellfish." *International Journal of Food Microbiology* 81-116.

McBreen, F, N Askew, A Cameron, D Connor, H Lillis, and A Carter. 2011. *UK SeaMap 2010: Predictive mapping of seabed habitats in UK waters.* No. 446, Joint Nature Conservation Committee, Petersborough: JNCC. https://www.researchgate.net/publication/290604692_UK_SeaMap_2010_Predictive_mapping_of_seabed_habitats_in_UK_waters.

Rippey, Scott R. 1994. "Infectious diseases associated with molluscan shellfish consumption." *Clinical Microbiology Reviews* 419-425.

Schmidt, Wiebke, Hayley L Evers-King, Carlos JA Campos, Darren B Jones, Peter I Miller, Keith Davidson, and Jamie D Shutler. 2018. "A generic approach for the development of short-term predictions of Escherichia coli and biotoxins in shellfish." *Aquaculture Environment Interactions* 10: 173-185. doi:https://doi.org/10.3354/aei00265.

Soulsby, RL. 1997. *Dynamics of Marine Sands: A Manual for Practical Applications.* London: Thomas Telford.

Velo-Suarez, L, B Reguera, S Gonzalez-Gil, M Lunven, P Lazure, E Nezan, and P Gentien. 2010. "Application of a 3D Lagrangian model to explain the decline of a Dinophysis acuminata bloom in the Bay of Biscay." *Journal of Marine Systems* 83: 242-252.

Visciano, Pierina, Maria Schirone, Miriam Berti, Anna Milandri, Rosanna Tofalo, and Giovanna Suzzi. 2016. "Marine biotoxins: Occurrence, toxicity, regulatory limits and reference methods." *Frontiers in Microbiology* 7 (1051): 1051.

10. Appendices

1. Forecasting modelling for Harmful Algal Blooms in Alfacs Bay

1. Experimental design

1.1. Forecasting modelling for Harmful Algal Blooms -Alfacs Bay

Harmful Algal Blooms (HABs): Building on previous studies, data on hydrodynamics, meteorology, nutrient pollution and ecological aspects of algal blooms were used to develop forecasting capabilities for the occurrence and impacts of HABs in SPAs. The accuracy and representativeness of existing forecasting models was improved through training of neural nets for a larger number of model variables to ensure that predictions represent the proliferation of HABs at appropriate temporal and spatial scales. Forecasting models were developed for a sentinel site, Alfacs Bay in Spain. For which there are robust historical datasets on bloom dynamics, toxicity patterns and water movements. The species and genera of algae that were considered in the models were *Pseudo-nitzschia* spp. implicated in amnesic shellfish

Grant agreement: 773400

poisoning, *Dinophysis* spp. implicated in diarrheic shellfish poisoning and *Alexandrium* spp. implicated in paralytic shellfish poisoning. The results of the modelling studies may be used to for a risk management protocol for HABs that will consider the implications of implementing a dynamic buffer zone approach in SPAs across Europe. This protocol will be prepared in a simple format that can be easily and quickly understood by shellfish farmers. The risk management plans will be developed and validated in collaboration with our industry partners (FEPROMODEL, Spain). This will assist more targeted and flexible approaches to risk management and will ensure that the plan can be used by other members of the industry thus improving the economic sustainability of shellfish farming businesses and consumer confidence.

1.1.1. Data compilation

To develop forecasting models a large data set is required. The Monitoring Programme for the Quality of Waters in Shellfish Harvesting Areas in the Catalan coastline carried out and managed by IRTA is collecting data since 1990. Weekly sampling was performed in five stations in Alfacs Bay (Figure 1).

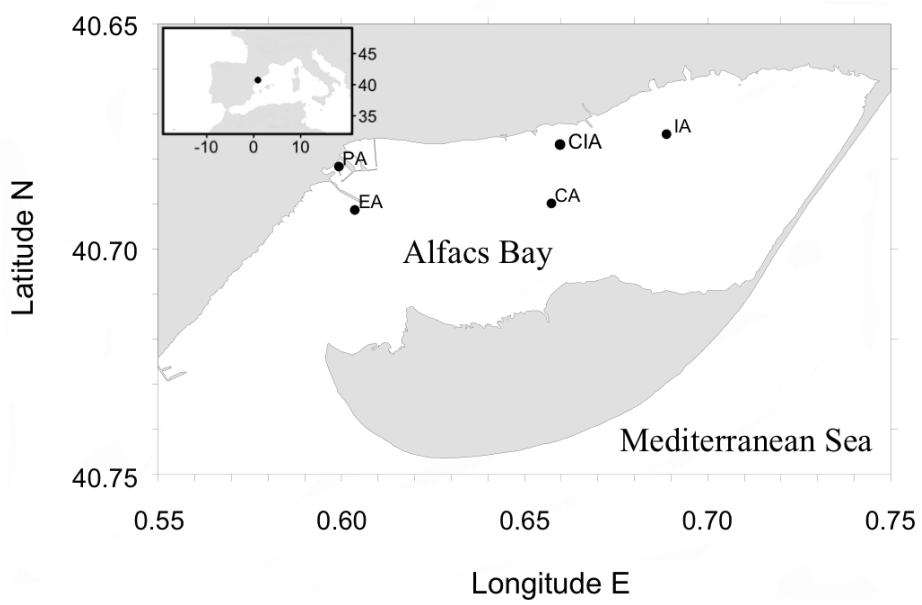


Figure 1. Alfacs Bay and the stations sampled in the Monitoring Programme.

In each station, physical and biological parameters are measured. Water temperature, salinity, oxygen and turbidity are measured by means of a portable probe (WTW). Phytoplankton species composition was determined from formalin preserved samples until 2007 and lugol's preserved samples from 2008 until present, using the Utermöhl technique (1958). Chlorophyll-a is measured using the fluorometric technique. This data is complemented with parameters from other sources. Meteorological data from an automated meteorological station of the Catalan Meteorological Service (METEOCAT, webpage: http://www.meteo.cat/servmet/index.html), placed on the north coast of Alfacs Bay. Daily air temperature (ºC), precipitation (mm), wind direction (◦) and speed (ms^{-1}), barometric pressure (hPa) and solar irradiation ($W m^{-2}$) were gathered. The daily river flow rate ($m^3 s^{-1}$) from Ebre River at Tortosa (North-East Alfacs Bay) was obtained from the Confederación Hidrográfica del Ebro (CHE, webpage:

Grant agreement: 773400

http://195.55.247.237/saihebro/), the governmental organisation that regulates the use of Ebro River water and its infrastructures.

Table 1. Summary of the main data sources for this task

	Starting year	Frequency	Source
Physical	1990	Weekly	IRTA
Biological	1990	Weekly	IRTA
Meteorological	1994	Daily	Meteocat
River Flow	1990	Daily	CHE

These datasets were explored to detect possible errors and missing data. This step was paramount because the models obtained depend on the veracity of the data used and therefore time is needed to accomplish this objective. Different techniques were applied. Visual data of the plotted profiles allowed a fast detection of errors in the water column and summary statistics and multivariate analysis, as the Principal Component Analysis, detected possible outliers in the data set. Once the error was detected, the original data set and the source of data (field sheets) was examined to solve a possible error of data transcription. Once the errors were solved and the possible missing values were complemented, the derived variables were recalculated as the water salinity and density. Following, integrated or summary variables of the water column for each parameter were calculated: maximum, minimum, mean, maximum gradients of the variable. For the daily variables, seven day mean, maximum and minimum statistics from the sampling day were calculated.

Once each data type was ready, all data sets were gathered in a unique data set with one row per sampling day and station, including metadata information (station, sampling day, month and year, station, coordinates), physical, meteorological, river flow and biological data. The structure of the final data set presented 622 columns and 16 313 rows, resulting in a total of more than 10^7 cells. This data set was then used to perform the training of the models to forecast Harmful Algal Blooms in Alfacs Bay.

In Guallar et al. (2016), the strategy to develop forecasting models was to use all the available data from the five sampling stations in Alfacs Bay. In this case, we focused the effort in a unique station to have the information from a single point, in view to use this approximation to other cases where only information from a single station is available. Therefore, the first step was to select the station to model. From the five available stations in Alfacs Bay, we decided to focus on the central station. The main reason was because it presented the longer and more consistent time series where all the samples were taken as an integrated sample using a hose, as it is recommended at present.

With this criterion, we subset the data to present only the integrated samples from the central station and we performed a data set for each taxon objective. In Alfacs Bay, we focused on *Alexandrium minutum*, *Pseudo-nitzschia* spp., *Dinophysis* spp. (resulting from adding all the counts corresponding to *Dinophysis* species), and also, we differentiated between *Dinophysis caudata* and *Dinophysis sacculus*. Therefore, five data sets were obtained and used to perform the forecasting models.

1.1.2. Architecture of the models developed

For the purposes of this study, feed-forward neural network with sigmoid (or logistic) activation function was used. This is the most popular and widely-used network paradigm in many applications. The procedure used for training the artificial neural network (ANN) was the back-propagation with momentum term and flat spot elimination algorithm (Fahlman, 1988; Rumelhart et al., 1985; Rumelhart et al., 1986). The procedure of back-propagation repeatedly adjusts the weights of the connections in the network in order to minimise the difference between the actual output vector of the net and the desired output vector (Rumelhart et al., 1986). This procedure is controlled by the learning rate (η) and the maximum ignored error (dmax). This method is usually considered slow and time consuming because it takes long time to converge. Several ways were developed to speed up the training (Almeida et al., 1990; Fahlman, 1988; Vogl et al., 1988). Adding the momentum term (μ) increase the convergence rate and stabilise the learning procedure (Hagiwara,1992; Phansalkar and Sastry, 1994; Rumelhart et al., 1985). Flat spot elimination (c) is a parameter to avoid the blockage of the neurons during the training (Fahlman, 1988). In this study, the following neural network parameters were used: learning rate, η = 0.1; momentum term, μ = 0.5; flat spot elimination, c = 0.0005; and maximum ignored error, dmax= 0. These values are based on bibliography, SNNS manual (Stuttgart Neural Network Simulator; Zell et al., 1998) and trial and error procedure.

The neural network architecture is composed of three single layers. The input layer of the model depends on the number of the predictor variables selected. The number of the neurons in the input and hidden layer was determined by trial and error procedure. The final input and hidden configuration chosen in each model were determined by two criteria: the best performance of the models and their cross-validation to avoid overfitting (Tetko et al., 1995). The model type determined the output configuration. One output neuron was necessary for abundance models and two output neurons were necessary for absence-presence or threshold models (Figure 2). Standardization of the data was performed to avoid computational problems, to meet algorithm requirement and to facilitate network learning (Lapedes and Farber, 1988; Sharda and Patil,1992; Srinivasan et al., 1994; Zhang et al., 1998). Data until 2015 is used to perform the models and were randomly separated in two groups: Training set (80% samples) and test set (20% samples). Training and test sets were used during the training of the model. The first was used to adjust the weights and the latter to visualise the overtraining of the model developed. Once the model was trained, the contiguous three years of data (2015–2018) were used to evaluate the forecasting power of the models.

Grant agreement: 773400

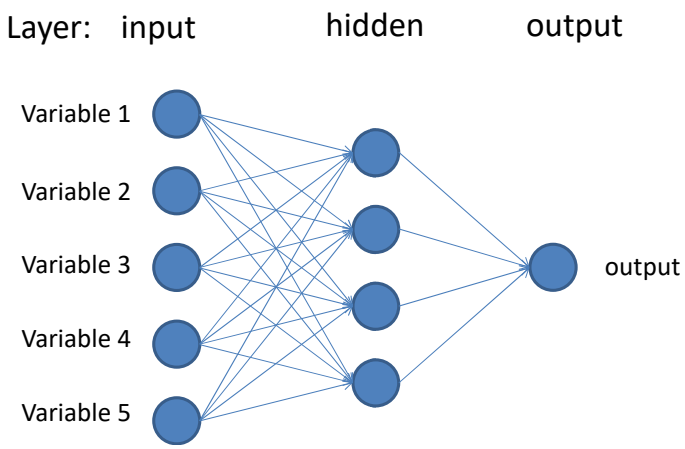


Figure 2. Example of neural network architecture with an input layer with five variables (input neurons), a hidden layer with four hidden neurons and an output layer with one neuron.

1.1.3. Model evaluation

The performance of the models was assessed through different tests. During the training of the model, in each iteration step, the internal test used was the Sum of Square Error (SSE). Once the model training was finished, the model was evaluated. In absence-presence and threshold alert models two methods were used, the Miss-Classification Error (MCE, %) and the Area Under the Curve (AUC) of a Receiver-Operating Characteristics (ROC) plot. MCE quantifies the bad classified cases (Hand, 1997). AUC tests whether the predicted results differed significantly from a random prediction (Elith et al., 2006; Fielding and Bell, 1997). Following Swets' scale (Swets, 1988), predictions are considered random when AUC is under 0.5, poor when it is in the range 0.5–0.7, and useful in the range 0.7–0.9. AUC values greater than 0.9 are considered good to excellent (1 = perfect). In abundance models, the coefficient of determination (R^2) was used. It shows the proportion of the variance explained by the model.

These evaluation tests were performed for each of the data sets (training, test and validation). To choose the best combination of the three results, we developed a simple algorithm to obtain the best model according to present the best evaluation values and avoiding too much difference between evaluation values in each data set. By this way we chose a model with low dispersion of performance between data sets, and hence, the model was the less overtrained. This algorithm was calculated with the mean of the evaluation results of the three data sets and then, the standard deviation was subtracted (in the coefficient of determination and in the AUC) or added (in the MCE) to penalise the possible dispersion of the values.

1.1.4. Model strategy

The strategy to develop models is the same for each taxon objective. To explore the data, first we performed simple models as linear models (LM), generalised linear models (GLM), or more complex models used to forecast HABs, as the zero-inflated negative binomial (ZINB) model (Cusack et al, 2015). The results obtained with these models are used as a reference to compare the models developed here. To obtain these models, all available cases of the corresponding species are used.

Grant agreement: 773400

Then, the training of the neural networks is addressed. Many options are available to perform the neural network model and therefore, a strategy to obtain the best forecasting model was developed. Four different strategies were combined to obtain the best possible model. We called it **variable selection**, **temporal**, **typology** and **neural network structure** strategies. They were applied simultaneously for each of the taxon addressed in this project.

The **variable selection strategy** refers to choose the best combination of available variables to perform the forecasting model. In neural networks, there is no automatic algorithm to obtain the best combination of variables as for example in linear models. Here, we try different models adding or suppressing variables from the model and comparing the results obtained according to the following criteria:

A. The criterium to add a variable into a model is to choose the one with the highest correlation with the data case to forecast.
B. Avoid those variables that also had high correlation with one of the variables already taking part of the neural network model.

In the **temporal strategy**, three different models were developed regarding the lag of the variables chosen from the sample to predict. In all the models, the five preceding weeks of the abundance of the taxon objective were always used due to their proved prediction capability (Guallar et al. 2016). Then, the environmental variables with the maximum correlation with the sample to predict were chosen according to the lag from this sample:

1) Lag with maximum correlation: The variables and lags used were those with the maximum correlation with the taxon abundance of the sample objective. The lags used start from one week before.
2) One lag week with the maximum correlation: Only variables with lag -1 week were considered, except the variables regarding taxon abundance, as explained before.
3) Two lag weeks with the maximum correlation: Only variables with lag -2 week were considered, except the variables regarding taxon abundance, where lag weeks between -2 and -4 were allowed to develop the forecasting model.

The typology strategy presented three different ways to address the development of forecasting depending of the output of the prediction given by the model:

a) Concentration model with all range of abundances: The objective was to obtain a prediction of the taxon concentration.
b) Presence – absence model combined with concentration model with only taxon presence cases: two consecutive models were performed where the first predict the presence or absence of the species, and the second predict the taxon concentration only from data cases where the taxon is present. The principle behind this strategy is the idea that the drivers that determine the presence or absence of the species is different from the drivers that increase or decrease the cell

Grant agreement: 773400

concentration in the ecosystem (Cusack et al, 2015). This strategy is already used in Alfacs Bay (Guallar et al, 2016).

c) Threshold alert model: In this classification model the final result was to determine if the taxon concentration was higher or lower than its threshold alert concentration, which was the final objective of this task. It was the same procedure as to perform a presence / absence model but fixing the reference value as the threshold alert concentration. The threshold level will depend on the taxon (Table 2).

Table 2. Species threshold alert values in Alfacs Bay.

Taxon	Threshold alert concentration (Cells L^{-1})
Alexandrium minutum	1.000
Pseudo-nitzschia spp.	2.000.000
Dinophysis spp.	500
Dinophysis caudata	500
Dinophysis sacculus	500

As a result, for each taxon, five different types of neural network models were developed. The variables used to develop the models were selected based on prior knowledge of the underlying process involved (bibliography and experts) and with autocorrelation and cross-correlation analyses based on the generalised linear model using the logistic function. This is the same function used as the activation function in the neural networks developed. This analysis showed the better-lagged period to use for the model development. In addition, the final variables selected had to be registered during enough time to avoid eliminating too many data cases. A final data set without gaps and with all the available cases to develop the model was obtained. Each row corresponded to a data case and columns corresponded to the abundance of the taxon target and the chosen lagged variables used to develop the models.

The neural network structure strategy addresses the different architecture and training procedures available to develop the models. Therefore, they are assayed to find the best model. They are classified as:

I. Hidden neurons: The neural network models presents a hidden or internal layer between the input and the output neurons. The configuration of the internal layer could be as simple as one neuron to a high complexity with many hidden layers with many internal neurons each. In our case, we trained the models with only one hidden layer presenting from 4 to 10 hidden neurons. Therefore, 7 neural networks were developed for each combination of the other strategies.

Grant agreement: 773400

II. Number of iterations: To train a neural network, the parameters of the neurons were updated continuously to better fit the training data. However, the more iterations done, the higher the risk to overtrain the model. That means the model better forecast the training set but predict worst the test or validation data. Therefore, this point tries to detect the overtraining of the model. To achieve it, a minimum of two models are trained: one with 50 iterations and another with 300 iterations. Sometimes, the number of iterations reach 5000.

III. Initial neuron parameters: In the first iteration of the training procedure, the neural network parameters are assigned randomly. They are called activation parameters. The final parameters, and therefore, the final model may depend on this activation parameters. Therefore, for each combination of the strategies, 20 models were developed with different activation parameters.

IV. Training and test data sets: As commented, to develop the neural networks, the data for each taxon is split in three data sets: The validation set corresponds to the data from 2015 until present. The rest (from 1990 until 2014) is divided randomly between training (80% of data) and test (20% of data) data sets. This may also determine the performance of the model and therefore, for each combination of strategies, 50 different cycles of model training were performed each of which presented different groups of training and test sets.

1.1.5. Results organization

For each taxon, a total of 114 runs were performed. They were divided in 36 runs to obtain models to forecast cell concentration from all ranges of abundances, 36 runs to obtain models to forecast cell concentration from only presence cases and 42 runs to develop classification models divided between models to predict presence absence of the taxon and to predict cases higher and lower than the threshold alert abundance.

The neural network models trained were grouped in runs. Each run presents a combination of variables selected using one temporal strategy (max. corr., lag-1 or lag-2), one type of model (concentration, presence-absence, a combination of both or threshold alert) and one of the iterations values. However, each run is initialised many times. As an example, the number of models in one run may reach 7000. It results of the combination of 7 different number of hidden neurons (4 to 10), 20 different sets of activation parameters and 50 different combination of training and tests sets (7*20*50 = 7000 models / run).

For each taxon, a total of 114 runs were performed. They were divided in 36 runs to obtain models to forecast cell concentration from all ranges of abundances, 36 runs to obtain models to forecast cell concentration from only presence cases and 42 runs to develop classification models divided between models to predict presence absence of the taxon and to predict cases higher and lower than the threshold alert abundance. All the models developed were saved.

To manage this huge number of models and results, for each run three files were created: an excel, a pdf and a RDATA.

1) Excel file: This file presents the summary of each of the models developed in the corresponding run. Each row corresponds to a neural network and hence, there are as many rows as models developed in the corresponding run. Each row has the information of the different options of model configuration, an identification number of the model and the summary statistics of the

performance of the model for each of the data sets and the global performance value. That is, the coefficient of the correlation in the concentration models and the MCE error and the AUC in the presence – absence and threshold alert models.

2) Pdf file: This file presents graphical information about the models developed. Each file presents as many pages as models developed in the corresponding run. The page number is equivalent to the identification number of the model. For concentration models, three figures were depicted: The evolution of the sum of squared errors along the iterations of the model developed, the observed vs estimated values of each data set (training, test and evaluation data sets) and the error histogram fit. For the presence – absence or the alert threshold models, only the evolution of the sum of squared errors along the iterations of the model developed were depicted (Figure 3).
3) RDATA file: This is a R format data file where all the models developed in the corresponding run are saved. In addition, data sets, variables and configuration options were also saved.

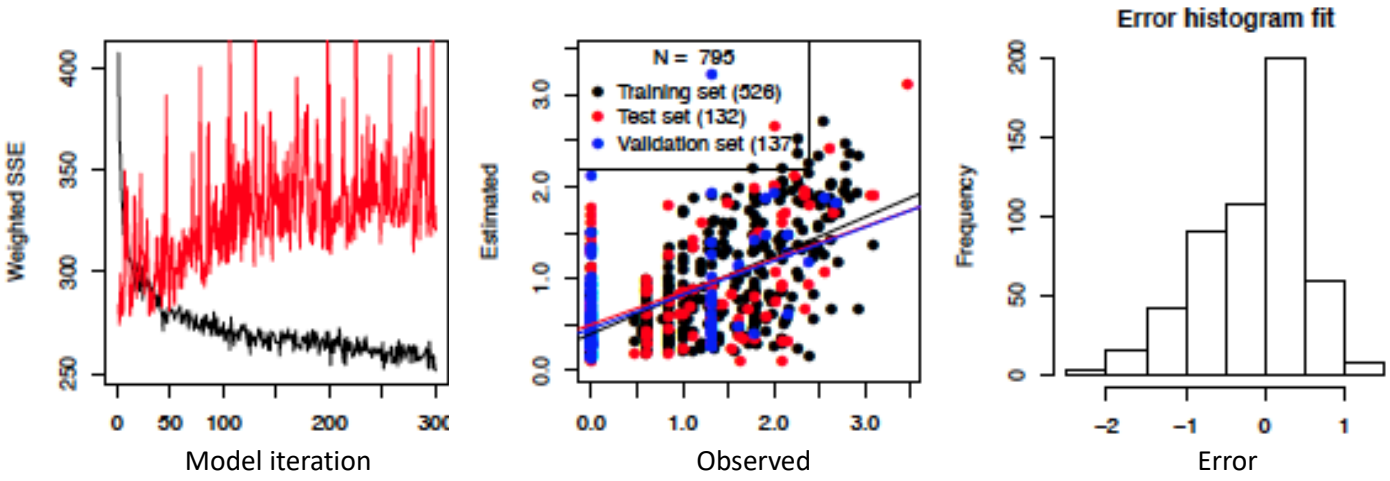


Figure 3. Example of the figures showed in the pdf file for each model. This example corresponds to the model 138 of the run 31 for the model type of forecasting *Dinophysis* spp. abundance .Left panel shows the evolution of the sum of squared errors along the iterations of the model training, middle panel is the observed vs. estimated values plot and right panel shows the error histogram fit.

Finally, an excel file is also performed where the summary of the two best models of each run is included. There is a sheet for each of the taxa studied.

1.1.6. Data analysis

All statistical computations were developed using R programming environment v.3.5.1 (RCoreTeam, 2018) with the RStudio editor v.1.1.463 (RStudio, 2018). For the neural networks training, the RSNNS package v.0.4-10 was used (Bergmeir and Benítez, 2012). This implements an R interface to the SNNS (Zell et al., 1998).

Grant agreement: 773400

2. Results and Discussion

2.1. Forecasting Modeling for Harmful Algal Blooms

2.1.1. Forecasting modeling in Alfacs Bay

In this study, the first attempts to forecast HABs in Alfacs Bay was with classical model types as linear models, generalised linear models and zero-inflated negative binomial models. The best performance obtained in this phase was a minimum coefficient of correlation (R^2) of 0.44 for a ZINB model for *Dinophysis caudata*, to a maximum of around 0.6 for a linear model for *Dinophysis sacculus* or a ZINB model for *Pseudo-nitzschia* spp. (Table 3).

Table 3. Best model results obtained for each taxon studied and performed with other model types than neural network models.

	Best model type	n	R^2
Alexandrium minutum	Generalised Linear Model	1045	0.51
Pseudo-nitzschia spp.	Zero Inflated Negative Binomial	480	0.62
Dinophysis spp.	Linear Model	468	0.60
Dinophysis sacculus	Linear Model	481	0.60
Dinophysis caudata	Zero Inflated Negative Binomial	410	0.44

Regarding neural networks, tables 4 and 5 summarise the models designed to forecast cell concentration considering all the range values of abundances; and the two consecutive models of the strategy to predict first the presence and absence of the taxon and when presence is predicted forecast concentration of the taxon considering only values higher than 0 cells L^{-1}. The performance of these models was not enough satisfactory to use them for the aquaculture industry.

Table 4. Coefficient of correlation of the best models obtained designed to forecast taxon concentration when all abundances or abundances higher than 0 cells L^{-1} are considered.

R^2	Type	Test	Training	Validation	Decision algorithm
Alexandrium minutum	All abundances	0.558	0.590	0.586	0.560
	Abundances > 0 cell L^{-1}	0.501	0.452	0.451	0.439

Grant agreement: 773400

Pseudo-nitzschia spp.	All abundances	0.740	0.671	0.553	0.560
	Abundances > 0 cell L^{-1}	0.579	0.605	0.517	0.522
Dinophysis spp.	All abundances	0.413	0.436	0.433	0.414
	Abundances > 0 cell L^{-1}	0.410	0.458	0.428	0.408
Dinophysis sacculus	All abundances	0.597	0.556	0.579	0.556
	Abundances > 0 cell L^{-1}	0.422	0.460	0.485	0.424
Dinophysis caudata	All abundances	0.506	0.503	0.001	0.045
	Abundances > 0 cell L^{-1}	0.495	0.519	NA	0.490

Table 5. Misclassification error and area under the ROC curve of the best models obtained designed to forecast taxon presence or absence.

Presence / Absence	Evaluation	Test	Training	Validation	Decision algorithm
Alexandrium minutum	MCE (%)	0.14	0.14	0.10	0.15
	AUC	0.90	0.84	0.83	0.82
Pseudo-nitzschia spp.	MCE (%)	0.05	0.06	0.01	0.07
	AUC	0.97	0.93	NA	0.93
Dinophysis spp.	MCE (%)	0.24	0.24	0.23	0.24
	AUC	0.85	0.77	0.72	0.71
Dinophysis sacculus	MCE (%)	0.14	0.16	0.16	0.17
	AUC	0.91	0.91	0.90	0.90
Dinophysis caudata	MCE (%)	0.17	0.22	0.05	0.23
	AUC	0.84	0.85	0.82	0.82

In the five taxa studied, the best models to forecast HABs were those obtained with the type of model designed to predict cases with higher or lower concentration than the alert threshold. High performance models were obtained for the three types of models regarding temporal strategy (maximum concentration without considering the lag period, models developed with variables with one lag week with the maximum

Grant agreement: 773400

correlation and models developed with variables with two lag weeks with the maximum correlation). Following the three models for each taxon are showed.

2.1.1.1. *Alexandrium minutum*

For *A. minutum*, the best model performed with the **variables and lags with the maximum correlation** was obtained in the run 40. The model number identification is 1342. Nine variables were used to develop this model, presenting 777 cases with data:

- Log-10 transformed abundances of *A. minutum* from lag -1 week to lag -5 week (5 variables).
- Surface temperature at lag -8 week.
- Seven-day mean Irradiance at lag -13 week.
- Density (Sigma-T) mean of the water column at lag -8 week.
- Difference of temperature between surface and bottom layers at lag -14 week.

The evaluation results of this model are showed in the table 6 and 7. For the MCE results, the values are 3.51%, 6.20% and 4.44% for the training, test and validation data sets, respectively. The model selection algorithm showed a result of 6.08%. For the AUC results, the values are 0.95, 0.96 and 0.95 for the training, test and validation data sets, respectively. The model selection algorithm showed a result of 0.94.

In the case of the model developed with variables with the maximum correlation and **lag -1 week**, the best model was obtained in the run 41. The model number identification is 90. Nine variables were used to develop this model, presenting 896 data cases:

- Log-10 transformed abundances of *A. minutum* from lag -1 week to lag -5 week (5 variables).
- Bottom temperature at lag -1 week.
- Log-10 transformed density (Sigma-T) mean of the water column at lag -1 week.
- Log-10 transformed 7-day mean river flow at lag-1 week.
- Difference of oxygen saturation between surface and bottom layers at lag -1 week.

The evaluation results of this model are showed in the table 5 and 6. For the MCE results, the values are 2.64%, 5.92% and 5.04% for the training, test and validation data sets, respectively. The model selection algorithm showed a result of 6.23%. For the AUC results, the values are 0.96, 0.93 and 0.95 for the training, test and validation data sets, respectively. The model selection algorithm showed a result of 0.94.

In the case of the model developed with variables with the maximum correlation and **lag -2** week, the best model was obtained in the run 42. The model number identification is 1082. Eight variables were used to develop this model, presenting 929 data cases:

- Log-10 transformed abundances of *A. minutum* from lag -2 week to lag -5 week (4 variables).
- Bottom temperature at lag -2 week.
- Density (Sigma-T) mean of the water column at lag -2 week.
- Log-10 transformed 7-day mean river flow at lag -2 week.
- Difference of oxygen saturation between surface and bottom layers at lag -2 week.

Grant agreement: 773400

The evaluation results of this model are showed in the table 5 and 6. For the MCE results, the values are 1.59%, 5.70% and 3.52% for the training, test and validation data sets, respectively. The model selection algorithm showed a result of 5.66%. For the AUC results, the values are 0.95, 0.92 and 0.89 for the training, test and validation data sets, respectively. The model selection algorithm showed a result of 0.89.

2.1.1.2. *Pseudo-nitzschia spp.*

For *Pseudo-nitzschia* spp., the best model performed with the **variables and lags with the maximum correlation** was obtained in the run 40. The model number identification is 83. Nine variables were used to develop this model, presenting 601 cases with data:

- Log-10 transformed abundances of *Pseudo-nitzschia* spp. from lag -1 week to lag -5 week (5 variables).
- Density (Sigma-T) mean of the water column at lag -11 week.
- Log-10 transformed bottom temperature at lag -14 week.
- Difference of temperature between surface and bottom layers at lag -1 week.
- Maximum gradient of temperature of the water column at lag -1 week.

The evaluation results of this model are showed in the table 5 and 6. For the MCE results, the values are 0.00%, 0.00% and 0.00% for the training, test and validation data sets, respectively. The model selection algorithm showed a result of 0.00%. For the AUC results, the value for the training data set is 1.00. No AUC value for the test and validation data sets are obtained. The reason is the absence of cases where abundance is higher than the threshold value.

Table 6. Misclassification error percentages of the models developed to forecast abundances higher or lower than the threshold alert values for the five taxa studied.

Misclassification error (%)	Type	n	Test	Training	Validation	Decision algorithm
Alexandrium minutum	Max. correlation	777	3.51	6.20	4.44	6.08
	lag -1	896	2.64	5.92	5.04	6.23
	lag -2	929	1.59	5.70	3.52	5.66
Pseudo-nitzschia spp.	Max. correlation	601	0.00	0.00	0.00	0.00
	lag -1	626	0.26	0.00	0.00	0.24
	lag -2	639	0.25	1.00	0.00	0.94
Dinophysis spp.	Max. correlation	795	0.57	0.76	0.73	0.79
	lag -1	835	0.90	0.00	0.00	0.82
	lag -2	868	2.76	3.42	0.00	3.87

Grant agreement: 773400

Dinophysis sacculus	Max. correlation	797	1.33	2.27	0.00	2.34
	lag -1	835	0.90	0.00	0.00	0.82
	lag -2	868	1.03	0.68	0.70	1.00
Dinophysis caudata	Max. correlation	815	0.00	0.00	0.00	0.00
	lag -1	905	0.00	0.00	0.00	0.00
	lag -2	940	0.00	0.00	0.00	0.00

In the case of the model developed with variables with the maximum correlation and **lag -1 week**, the best model was obtained in the run 41. The model number identification is 17. Nine variables were used to develop this model, presenting 626 data cases:

- Log-10 transformed abundances of *Pseudo-nitzschia* spp. from lag -1 week to lag -5 week (5 variables).
- Difference of temperature between surface and bottom layers at lag -1 week.
- Maximum gradient of temperature of the water column at lag -1 week.
- Salinity mean of the water column at lag -1 week.
- Log-10 transformed Chlorophyll-a at lag -1 week.

The evaluation results of this model are showed in the table 5 and 6. For the MCE results, the values are 0.26%, 0.00% and 0.00% for the training, test and validation data sets, respectively. The model selection algorithm showed a result of 0.24%. For the AUC results, the value for the training data set is 0.983. No AUC value for the test and validation data sets are obtained. The reason is the absence of cases where abundance is higher than the threshold value.

In the case of the model developed with variables with the maximum correlation and **lag -2 week**, the best model was obtained in the run 42. The model number identification is 919. Eight variables were used to develop this model, presenting 639 data cases:

- Log-10 transformed abundances of *Pseudo-nitzschia* spp. from lag -2 week to lag -5 week (4 variables).
- Difference of temperature between surface and bottom layers at lag -2 week.
- Maximum gradient of temperature of the water column at lag -2 week.
- Oxygen saturation mean of the water column at lag -2 week.
- Salinity mean of the water column at lag -2 week.

The evaluation results of this model are showed in the table 5 and 6. For the MCE results, the values are 0.25%, 1.00% and 0.00% for the training, test and validation data sets, respectively. The model selection algorithm showed a result of 0.94%. For the AUC results, the values are 0.999 and 0.907 for the training

Grant agreement: 773400

and test data sets, respectively. No AUC value for the validation data set is obtained. The reason is the absence of cases where abundance is higher than the threshold value.

Table 7. Area under the ROC curve of the models developed to forecast abundances higher or lower than the threshold alert values for the five taxa studied.

Area Under the ROC curve	Type	n	Test	Training	Validation	Decision algorithm
Alexandrium minutum	Max. correlation	777	0.946	0.960	0.946	0.943
	lag -1	896	0.956	0.934	0.950	0.935
	lag -2	929	0.952	0.921	0.895	0.894
Pseudo-nitzschia spp.	Max. correlation	601	1.000	NA	NA	1.000
	lag -1	626	0.983	NA	NA	0.983
	lag -2	639	0.999	0.907	NA	0.888
Dinophysis spp.	Max. correlation	795	0.999	1.000	NA	0.998
	lag -1	835	0.957	NA	NA	0.957
	lag -2	868	0.951	0.972	NA	0.946
Dinophysis sacculus	Max. correlation	797	0.984	0.985	NA	0.984
	lag -1	835	0.995	NA	NA	0.995
	lag -2	868	0.930	0.945	NA	0.998
Dinophysis caudata	Max. correlation	815	1.000	1.000	NA	1.000
	lag -1	905	1.000	1.000	NA	1.000
	lag -2	940	1.000	1.000	NA	1.000

2.1.1.3. *Dinophysis spp.*

For *Dinophysis* spp., the best model performed with the **variables and lags with the maximum correlation** was obtained in the run 40. The model number identification is 1397. Nine variables were used to develop this model, presenting 795 cases with data:

- Log-10 transformed abundances of *Dinophysis* spp. from lag -1 week to lag -5 week (5 variables).
- Seven-day mean Irradiance at lag -7 week.

Grant agreement: 773400

- Surface temperature at lag -3 week.
- Log-10 transformed bottom density (Sigma-T) at lag -3 week.
- Log-10 transformed maximum temperature gradient of the water column at lag -12 week.

The evaluation results of this model are showed in the table 5 and 6. For the MCE results, the values are 0.57%, 0.76% and 0.73% for the test, training and validation data sets, respectively. The model selection algorithm showed a result of 0.79%. For the AUC results, the values are 0.99 and 1.00 for the training and test data sets, respectively. No AUC value for the validation data set is obtained. The reason is the absence of cases where abundance is higher than the threshold value.

In the case of the model developed with variables with the maximum correlation and **lag -1 week**, the best model was obtained in the run 41. The model number identification is 1181. Nine variables were used to develop this model, presenting 835 data cases:

- Log-10 transformed abundances of *Dinophysis* spp. from lag -1 week to lag -5 week (5 variables).
- Mean temperature of the water column at lag -1 week.
- Bottom density (Sigma-T) of the water column at lag -1 week.
- Seven-day mean Irradiance at lag -1 week
- Log-10 transformed 7-day mean river flow at lag-1 week.

The evaluation results of this model are showed in the table 5 and 6. For the MCE results, the values are 0.90%, 0.00% and 0.00% for the test, training and validation data sets, respectively. The model selection algorithm showed a result of 0.82%. For the AUC results, the value for the training data set is 0.957. No AUC value for the test and validation data sets are obtained. The reason is the absence of cases where abundance is higher than the threshold value.

In the case of the model developed with variables with the maximum correlation and **lag -2 week**, the best model was obtained in the run 42. The model number identification is 66. Eight variables were used to develop this model, presenting 868 data cases:

- Log-10 transformed abundances of *Dinophysis* spp. from lag -2 week to lag -5 week (4 variables).
- Surface temperature at lag -2 week.
- Log-10 transformed seven-day mean Irradiance at lag -2 week
- Bottom density (Sigma-T) of the water column at lag -2 week.
- Log-10 transformed 7-day mean river flow at lag-1 week.

The evaluation results of this model are showed in the table 5 and 6. For the MCE results, the values are 2.76%, 3.42% and 0.00% for the test, training and validation data sets, respectively. The model selection algorithm showed a result of 3.87%. For the AUC results, the values are 0.951 and 0.972 for the training and test data sets, respectively. No AUC value for the validation data set is obtained. The reason is the absence of cases where abundance is higher than the threshold value.

Grant agreement: 773400

2.1.1.4. *Dinophysis sacculus*

For *Dinophysis sacculus*, the best model performed with the **variables and lags with the maximum correlation** was obtained in the run 40. The model number identification is 763. Nine variables were used to develop this model, presenting 797 cases with data:

- Log-10 transformed abundances of *Dinophysis sacculus* from lag -1 week to lag -5 week (5 variables).
- Seven-day mean Irradiance at lag -8 week.
- Mean temperature of the water column at lag -3 week.
- Log-10 transformed bottom density (Sigma-T) at lag -3 week.
- Difference of temperature between surface and bottom layers at lag -10 week.

The evaluation results of this model are showed in the table 5 and 6. For the MCE results, the values are 1.33%, 2.27% and 0.00% for the test, training and validation data sets, respectively. The model selection algorithm showed a result of 2.34%. For the AUC results, the values are 0.984 and 0.985 for the training and test data sets, respectively. No AUC value for the validation data set is obtained. The reason is the absence of cases where abundance is higher than the threshold value.

In the case of the model developed with variables with the maximum correlation and **lag -1 week**, the best model was obtained in the run 41. The model number identification is 418. Nine variables were used to develop this model, presenting 835 data cases:

- Log-10 transformed abundances of *Dinophysis sacculus* from lag -1 week to lag -5 week (5 variables).
- Bottom temperature at lag -1 week.
- Log-10 transformed density (Sigma-T) mean of the water column at lag -1 week.
- 7-day mean irradiance at lag -1 week.
- Log-10 transformed 7-day mean river flow at lag -1 week.

The evaluation results of this model are showed in the table 5 and 6. For the MCE results, the values are 0.90%, 0.00% and 0.00% for the test, training and validation data sets, respectively. The model selection algorithm showed a result of 0.82%. For the AUC results, the value for the training data set is 0.995. No AUC value for the test and validation data sets are obtained. The reason is the absence of cases where abundance is higher than the threshold value.

In the case of the model developed with variables with the maximum correlation and **lag -2 week**, the best model was obtained in the run 42. The model number identification is 834. Eight variables were used to develop this model, presenting 868 data cases:

- Log-10 transformed abundances of *Dinophysis sacculus* from lag -2 week to lag -5 week (4 variables).
- Mean temperature of the water column at lag -2 week.
- Log-10 transformed bottom density (Sigma-T) of the water column at lag -2 week.
- 7-day mean irradiance at lag -2 week.
- Log-10 transformed 7-day mean river flow at lag -2 week.

The evaluation results of this model are showed in the table 5 and 6. For the MCE results, the values are 1.03%, 0.68% and 0.70% for the test, training and validation data sets, respectively. The model selection algorithm showed a result of 1.00%. For the AUC results, the values are 0.930 and 0.945 for the training and test data sets, respectively. No AUC value for the validation data set is obtained. The reason is the absence of cases where abundance is higher than the threshold value.

2.1.1.5. *Dinophysis caudata*

For *Dinophysis caudata*, the best model performed with the **variables and lags with the maximum correlation** was obtained in the run 40. The model number identification is 878. Nine variables were used to develop this model, presenting 815 cases with data:

- Log-10 transformed abundances of *Dinophysis caudata* from lag -1 week to lag -5 week (5 variables).
- Log-10 transformed bottom temperature at lag -6 week.
- Seven-day mean Irradiance at lag -14 week.
- Log-10 transformed 7-day mean river flow at lag -3 week
- Density (Sigma-T) mean of the water column at lag -6 week.

The evaluation results of this model are showed in the table 5 and 6. For the MCE results, the values are 0.00%, 0.00% and 0.00% for the test, training and validation data sets, respectively. The model selection algorithm showed a result of 0.00%. For the AUC results, the values are 1.000 and 1.000 for the training and test data sets, respectively. No AUC value for the validation data set is obtained. The reason is the absence of cases where abundance is higher than the threshold value.

In the case of the model developed with variables with the maximum correlation and **lag -1 week**, the best model was obtained in the run 41. The model number identification is 17. Nine variables were used to develop this model, presenting 905 data cases:

- Log-10 transformed abundances of *Dinophysis caudata* from lag -1 week to lag -5 week (5 variables).
- Log-10 transformed bottom temperature at lag -1 week.
- Log-10 transformed 7-day mean river flow at lag-1 week
- Density (Sigma-T) mean of the water column at lag -1 week.
- Difference of oxygen concentration between surface and bottom layers at lag -1 week.

The evaluation results of this model are showed in the table 5 and 6. For the MCE results, the values are 0.00%, 0.00% and 0.00% for the test, training and validation data sets, respectively. The model selection algorithm showed a result of 0.00%. For the AUC results, the values are 1.000 and 1.000 for the training and test data sets, respectively. No AUC value for the validation data set is obtained. The reason is the absence of cases where abundance is higher than the threshold value.

In the case of the model developed with variables with the maximum correlation and **lag -2 week**, the best model was obtained in the run 42. The model number identification is 886. Eight variables were used to develop this model, presenting 940 data cases:

Grant agreement: 773400

- Log-10 transformed abundances of *Dinophysis caudata* from lag -2 week to lag -5 week (4 variables).
- Log-10 transformed bottom temperature at lag -2 week.
- Log-10 transformed 7-day mean river flow at lag -2 week.
- Density (Sigma-T) mean of the water column at lag -2 week.
- Difference of oxygen concentration between surface and bottom layers at lag -2 week.

The evaluation results of this model are showed in the table 5 and 6. For the MCE results, the value s are 0.00%, 0.00% and 0.00% for the test, training and validation data sets, respectively. The model selection algorithm showed a result of 0.00%. For the AUC results, the values are 1.000 and 1.000 for the training and test data sets, respectively. No AUC value for the validation data set is obtained. The reason is the absence of cases where abundance is higher than the threshold value.

2.1.2. Performance of the models

The classical models (linear model, generalised linear models and zero-inflated negative binomial models) developed do not allow to use them to forecast HABs in Alfacs Bay with confidence. The higher performance was a coefficient of determination of 0.6. In addition, the development of the models was performed with all data cases, without considering a test and validation data sets. Furthermore, in some of the models, the variables selected were not available in all data cases and, consequently, limited the number of available cases to develop the model. In some models the cases used for the model is more than the half of the available cases for the taxon. Therefore, the model developed was less generalised as if all data cases were used.

In the case of neural networks models, the best models were those designed to predict if the taxon concentration is going to be higher or lower than the alert threshold abundance. In addition, the performance of the neural network models developed allowed to use them to forecast HABs in the central station of Alfacs Bay for the five taxa studied. Almost all models present an error of classification lower than 3.5% of the cases and an AUC value higher than 0.9, highlighting the exce llent results obtained. Only *Alexandrium minutum* models presented MCE values around 5%, still a good performance. It is important to note again that models are trained in a fraction of the available cases (training data set). Therefore, the results obtained for test data set and especially for the validation data set is a proof of the confidence of the models developed.

Furthermore, the development of models with the capacity to forecast HABs with one and two weeks in advance gives a powerful tool for the aquaculture industry. Managers may prevent the possible damage to ecosystem and to decrease the possible economic losses due to the closure of aquaculture farms when taxon abundance alert threshold is surpassed.

Grant agreement: 773400

3. Conclusions

Models to forecast HABs in Alfacs Bay have been developed. Best models were neural network algorithms designed to forecast abundances higher or lower than the threshold alert values. For each taxon, two models were developed to forecast HABs with one and two weeks before it happens. This strategy gives enough time for the aquaculture industry to manage their farms to avoid or decrease the possible economic losses due to the HABs episodes.

4. References

Almeida, L.B., Wellekens, C.J., Silva, F.M., Almeida, L.B., 1990. Acceleration techniques for the backpropagation algorithm, Neural Networks. Springer Berlin Heidelberg, pp. 110-119.

Bergmeir, C., Benítez, J.M., 2012. Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. Journal of Statistical Software 46(7), 1-26.

Cusack, C., Mouriño, H., Moita, M. T., Silke, J. 2015. Modelling Pseudo-nitzschia events off southwest Ireland. Journal of Sea Research 105, 30-41.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29(2), 129-151.

Fahlman, S.E., 1988. An empirical study of learning speed in back-propagation networks. Technical Report CMU-CS-88-162. School of Computer Science, Carnegie Mellon University, PA, Pittsburgh.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence / absence models. Environmental Conservation 24(1), 38-49.

Hagiwara, M., 1992. Theoretical derivation of momentum term in back-propagation. Neural Networks 1, 682-686.

Hand, D.J., 1997. Construction and assessment of classification rules. Wiley, University of Michigan.

Guallar, C., Delgado, M., Diogene, J., & Fernandez-Tejedor, M. (2016). Artificial neural network approach to population dynamics of harmful algal blooms in Alfacs Bay (NW Mediterranean): Case studies of Karlodinium and Pseudo-nitzschia. Ecological modelling, 338, 37-50.

Lapedes, A., Farber, R., 1988. How neural nets work, In: Anderson, D.Z. (Ed.), Neural Information Processing Systems. American Institute of Physics, New York, pp. 442-456.

Phansalkar, V.V., Sastry, P.S., 1994. Analysis of the back-propagation algorithm with momentum. IEEE Transactions on Neural Networks 5(3), 505-506.

RCoreTeam, 2018. R: A language and Environment for Statistical Computing, 3.5.1 ed. R Foundation for Statistical Computing, Vienna, Austria.

RStudio, 2018. RStudio: Integrated development environment for R RStudio, Boston, MA.

Rumelhart, D.E., Hinton, G., Williams, R.J., 1985. Learning internal representations by error propagation, In: Rumelhart, D.E., McClelland , J.L. (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition. MIT Press, Cambridge, MA.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323(6088), 533-536.Utermöhl, H., 1958. Zur vervollkommnung der quantitativen phytoplankton-methodik. Mitt. int. Ver. theor. angew. Limnol. 9, 1-38.

Sharda, R., Patil, R., 1992. Connectionist approach to time series prediction: an empirical test. Journal of Intelligent Manufacturing 3(5), 317-323.

Srinivasan, D., Liew, A.C., Chang, C.S., 1994. A neural network short-term load forecaster. Electric Power Systems Research 28(3), 227-234.

Tetko, I.V., Livingstone, D.J., Luik, A.I., 1995. Neural network studies. 1. Comparison of overfitting and overtraining. Journal of Chemical Information and Computer Sciences 35(5), 826-833.

Vogl, T.P., Mangis, J.K., Rigler, A.K., Zink, W.T., Alkon, D.L., 1988. Accelerating the convergence of the back-propagation method. Biological Cybernetics 59(4-5), 257-263.

Zell, A., Mamier, G., Vogt, M., Mache, N., Hübner, R., Döring, S., Herrmann, K.-U., Soyez, T., Schmalzl, M., Sommer, T., Hatzigeorgiou, A., Posselt, D., Schreiner, T., Kett, B., Clemente, G., Wieland, J., Gatter, J., 1998. SNNS Stuttgart Neural Network Simulator User Manual, Version 4.2. IPVR, University of Stuttgart and WSI, University of Tübingen. URL http://www.ra.cs.uni-tuebingen.de/SNNS/.

Zhang, G., Eddy Patuwo, B., Y. Hu, M., 1998. Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting 14(1), 35-62.

Grant agreement: 773400