FORCOAST

Earth Observation Services For Wild Fisheries, Oystergrounds Restoration And Bivalve Mariculture Along European Coasts

PROJECT DELIVERABLE REPORT

Deliverable Number: 5.3

Deliverable Title: Report Describing Common Methodology for Numerical Model Validation

Author(s): D. Pereiro, T. Dabrowski, L. Barbut, A. Caballero, L. Ferrer, G. Lacroix, S. Legrand, M. Maar, J. Murawski, S. Querin, A. Rubio, J.She, F. Campuzano, J. Staneva, L. Vandenbulcke

Work Package Number: 5

Work Package Title: Service Operationalization, Demonstration and Validation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

FORCOAST Project Information	
Project full title	Earth Observation Services For Wild Fisheries, Oystergrounds Restoration And Bivalve Mariculture Along European Coasts
Project acronym	FORCOAST
Grant agreement number	870465
Project coordinator	Ghada El Serafy, Deltares
Project start date and duration	1st November 2019, 36 months
Project website	https://forcoast.eu/

Deliverable Information	
Work package number	5
Work package title	Service Operationalisation, Demonstration & Validation
Deliverable number	D5.3
Deliverable title	Report Describing Common Methodology for Numerical Model Validation
Description	This deliverable provides standards or common methodologies to validate the FORCOAST coastal models, indicating which procedures should be followed for different ocean parameters and platforms.
Lead beneficiary	MI_IE
Lead Author(s)	D. Pereiro – Marine Institute, Ireland T. Dabrowski – Marine Institute, Ireland
Contributor(s)	L. Barbut - Royal Belgian Institute of Natural Sciences A. Caballero – AZTI F. Campuzano - Instituto superior Técnico, Universidade de Lisboa L. Ferrer - AZTI G. Lacroix - Royal Belgian Institute of Natural Sciences S. Legrand - Royal Belgian Institute of Natural Sciences M. Maar - Aarhus University, Denmark J. Murawski - Danish Meteorological Institute S. Querin - OGS Italy A. Rubio - AZTI J. She - Danish Meteorological Institute J. Staneva - HZG, Germany L. Vandenbulcke - Jailoo, Romania

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

Revision number	10
Revision Date	14/10/2022
Status (Final (F), Draft (D), Revised Draft (RV))	F
Dissemination level (Public (PU), Restricted to other program participants (PP), Restricted to a group specified by the consortium (RE), Confidential for consortium members only (CO))	PU

Document History			
Revision	**Date**	**Modification**	**Author**
0	13/10/2020	1st draft	D. Pereiro, T. Dabrowski
1	12/04/2021	Pilot contributions	Pilot Leaders
2	10/04/2021	1st Revision	D. Pereiro, T. Dabrowski
3	20/04/2021	Pilot contributions toSection 6	Pilot Leaders
4	28/04/2021	Final Revision	L. Rodriguez Galvez, G. El Serafy
5	15/04/2022	Additional Revision (see table below). Section 6 now consists of a single synthetic table of the foreseen validation efforts at each site.	D. Pereiro, T. Dabrowski
6	26/04/2022	Internal revision	A. Capet
7	28/04/2022	Final check	L. Rodriguez Galvez
8	03/10/2022	Updated Introduction section to include clarification on the model validation framework used together with flowchart	D. Pereiro, T. Dabrowski
9	11/10/2022	Review	L. Meszaros

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

10	14/10/2022	Review	Ghada El Serafy

Approvals				
	Name	**Organisation**	**Date**	**Signature (initials)**
Coordinator	Ghada El Serafy	Deltares	14/10/2022	GES
WP Leaders	Tomasz Dabrowski	Marine Institute	03/10/2022	TD

Reviewer comments and reply		
Date	**Comment**	**Reply**
Sep 2022	The deliverable should be updated to provide a clear quantified validation methodology. Additionally, the document should outline clear model selection criteria among different sites types to enhance reactivity upon users' requests	The introduction section has been updated to provide a clear validation methodology. Section 6 table provides information on model characteristics per area
Feb 2022	**General comments (page 2):** *[…] the validation methodology (D5.3) remains rather confused, very high-level, with limited information on the actual results to be expected (for instance, p22 "could also", "if and when available", or, p19 : "The only sources that are available are sea level data from tide gauges and *maybe* temperature observations", p18 "CMEMS provides [...] However, the quality of the data has not been validated").*	A synthetic table describing the foreseen validation efforts at each site has been added. This table summarizes the approach followed by each pilot.
Feb 2022	**Annex 1 (page 16):** *p22/23, maps that were shown during the review meeting are confusing. While the figures are adequately sourced as a sept 2018 report (that was made public in 2020), the text "For example we corroborated etc." could let the reader think that the author ("we") is the author of this deliverable, while the text was copied word for word from the 2018 report, which does not show clearly. The fact that these data validations were made in another 2018 project and have not been updated since should have appear clearly.*	This report is no longer referenced in the document. Instead, a table summarizes the foreseen validation efforts at each site.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

Feb 2022	*While the evaluation methodology does not raise concerns by itself, and is referring to widely accepted tools, the deliverable is not clear enough on the validation for multiple services, in multiple sites. For instance, while the Portugal site still has models to be validated, this deliverable does not mention Portugal (the word "Portugal" cannot be found in the document, which describes validation methodology for 7 sites and not 8).*	The new table describes the foreseen validation efforts in all sites, including Portugal. On the other hand, validation of services, including definitions of KPIs, will be documented in D5.5 (due August 2022), following the extended Grant Agreement schedule.
Feb 2022	*Rather than a different description for each site - descriptions are qualitatively heterogeneous- the consortium should provide the Commission with a single synthetic table of the foreseen validation efforts, for each site and service, with expected metrics for each validation (site/service/ metrics/duration of validation/hindcast-forecast/ measures/assured KPIs)*	This approach has been followed in this revision. A single synthetic table is provided now and describes the foreseen validation efforts at each site. Validation of services and associated KPIs, on the other hand, will be presented in D5.5 (due August 2022), following the Grant Agreement.
Feb 2022	*Some of the validation task will require the archiving of predictions, the timing and scope this archiving effort should be clarified, and given the project delays, the archiving of predictions should be started as soon as possible.*	The archiving of hindcasts and forecasts has been approached differently across the different partners. In some cases, considering that the same forcing has been used in the computation of both hindcasts and forecasts, then the performance of the model should be similar, and thus it is redundant to present both types of validation. This is for instance the rationale followed by Pilot 7 (Romania) to justify not presenting a forecast validation. In other cases, the services of interest in a region may rely entirely on either hindcasts or forecasts. For instance, SM-F2 Front Detection only depends on forecasts, thus making it unnecessary to provide an assessment of the performance of the model in hindcast mode for that application.
Feb 2022	*Given the project delays, even partial validation data for sites considered "validated" should be communicated as soon as possible (weeks).*	A draft of D5.4 is available, presenting model validation at each site.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

PROPRIETARY RIGHTS STATEMENT

This document contains information, which is proprietary to the FORCOAST consortium. Neither this document, or the information contained within may be duplicated, used or communicated except with the prior written permission of the FORCOAST coordinator.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

Executive Summary

This deliverable provides standards or common methodologies to validate the FORCOAST coastal models, indicating which procedures should be followed for different ocean platforms. Different quality assessment metrics are proposed in section 2: Mean Error (ME), Root Mean Square Error (RMSE), correlation coefficient (CORR), Adjusted Relative Mean Absolute Error (ARMAE) and Receiver Operating Characteristic (ROC) curve. Then, methods to be followed for hindcast (section 3), forecast (section 4) and process-oriented validation (section 5) are specified as well. A table summarizing the foreseen validation efforts at each site is provided. Assessment of coastal model performance following the standards described in this deliverable will result in the productionof a final coordinated pilot model evaluation report (D5.4).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

Table of Contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

Table of Figures

Table of Tables

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

1. Introduction

The objective of this deliverable is to provide standard procedures for the validation of the FORCOAST coastal models. First, different metrics will be proposed to evaluate the goodness of fit between model predictions and observations (section 2). Next, section 3 will show how to apply the proposed metrics for hindcast validation and with several types of ocean datasets from different platforms, such as time series (e.g. from tide gauges and moored buoys), vertical profiles (e.g. from CTD), gridded datasets (e.g. remote sensing from satellites and High Frequency Radar data) and local sampling (e.g. Niskin sampling for analysis of biogeochemical properties of the seawater). Then, section 4 deals with forecast validation. Section 5 offers an approach to process-oriented validation. The goal is to ensure that the same procedures are followed between different partners.

The flowchart below shows the **FORCOAST model validation framework**. Validation starts with the selection of observational data, either *in-situ* or remote sensing. Observational datasets are then used for (a) hindcast validation, (b) forecast validation and (c) process-oriented validation. These different approaches to validation (whose definitions are explained in the sections below) are applied to both (a) hydrodynamic and biogeochemical coastal models, and (b) FORCOAST services. Process-oriented validation focuses on processes which are of particular interest to end users (e.g. storm surges), and thus is more related to the final services. The comparison between the observational datasets and the model and service outputs is displayed as both (a) figures for visual inspection, and (b) Estimated Accuracy Numbers (EANs, see Section 2). The process ends with a proper product evaluation, determining if the model/service is suitable for its purpose or if it has to be improved. The application of this model validation framework to each pilot site is summarized in Section 6.

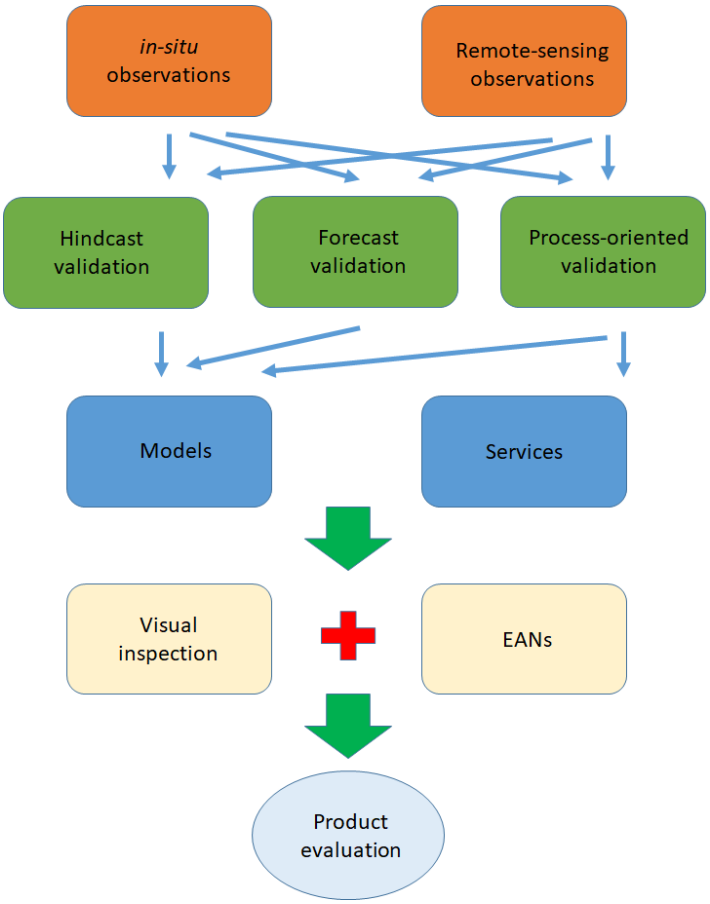

Figure 1. Validation framework

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

2. Metrics for quality assessment

In the following expressions, *x* represents the observed values, whereas *y* represents the predicted values. *N* is the number of data points.

2.1. Mean Error (ME)

The Mean Error (ME) is the mean of the differences between observations *x* and predictions *y*:

$$ME = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)$$

2.2. Root Mean Square Error (RMSE)

The Root Mean Square Error (RMSE) is the square root of the mean squared error between observations *x* and predictions *y*:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2}$$

2.3. Correlation coefficient (CORR)

The correlation coefficient is the covariance divided by the product of the standard deviations:

$$CORR = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{x_i - \overline{x}}{\sigma_x}\right)\left(\frac{y_i - \overline{y}}{\sigma_y}\right)$$

where $\bar{x}$ and $\bar{y}$ represent the mean and σ_x and σ_y represent the standard deviations of observations and predictions.

2.4. Adjusted Relative Mean Absolute Error (ARMAE)

The Adjusted Relative Mean Absolute Error (ARMAE, Sutherland et al., 2004) is:

$$\mathrm{ARMAE} = \frac{\langle |Y - X| - \mathrm{OE} \rangle}{\langle |X| \rangle}$$

where the angular brackets represent an average. Here *X* are observations and *Y* are predictions. In other words, the ARMAE is the mean absolute error divided by the mean absolute value of the observations. An observation error OE is subtracted to account for measurement errors, which are related to the instrument, measurement and conversion principles (Sutherland et al., 2004). For

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

example, an observation error of 1 cm s^{-1} was assumed in ADCP currents in Dabrowski et al., 2016. Depending on the ARMAE number, it is possible to assess the performance of the model as "excellent", "good", "reasonable", "poor" or "bad", according to Table 1, where an additional category is considered when there is not enough data available (N/D).

Performance	Range of ARMAE
Excellent	< 0.2
Good	0.2 – 0.4
Reasonable	0.4 – 0.7
Poor	0.7 – 1.0
Bad	> 1.0
N/D	-

Table 1. Evaluation of model's performance according to the ARMAE value

2.5. Receiver Operating Characteristic (ROC) curve

The receiver operating characteristic (ROC) curve is a graphical plot evaluating the predictive power of a binary classification system as its discrimination threshold is varied. The method is described in Brown and Davis (2006) and Fawcett (2006). The base is a yes/no decision, based on the comparison of two independent information sets (observations and model results in the FORCOAST case) with respect to a threshold value. The decision process is illustrated by Fig. 1; there are four possible outcomes for each trial, either correctly positive (CP), correctly negative (CN), incorrectly positive (IP) and incorrectly negative (IN).

This approach can be used to make an analysis of similarity of how well the model fits the data (Allen et al., 2007, 2008). In a perfect model all the points in a scatter diagram of model results vs. data lie on the x=y line (Fig. 1). Setting a threshold criteria (t) dividing the data into two sets and then comparing it with the model results using the same threshold (Fig. 1) allows to assess the similarity between model results and data at that threshold, effectively assessing the model's ability to discriminate that threshold. The perfect model will only give CP and CN outcomes. The more scatter there is in the model–data relationship, the more IP and IN conditions will occur and the worse the model performance will be. By varying the threshold across the full range of observations, a non-parametric measure of the model's ability to simulate a given variable is obtained. The decision process can be further assessed by calculating the Correct Negative Fraction (CNF) and the Correct Positive Fraction (CPF) as follows:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

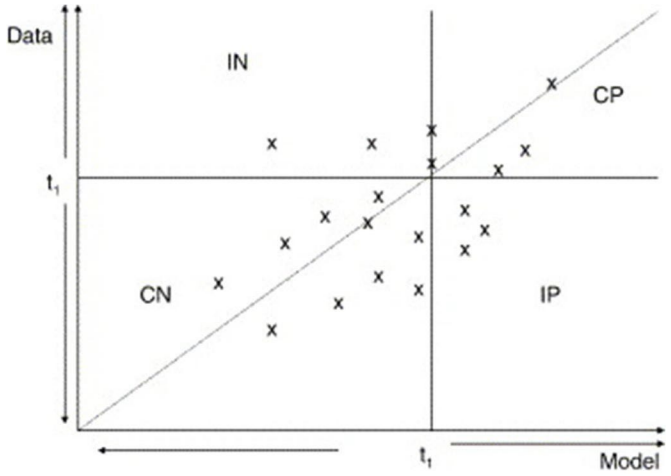

Figure 2. Schematic diagram of the discrimination analysis from Allen et al. (2007)

Correct Negative Fraction:

$$CNF = \frac{CN}{CN + IP}$$

False Positive Rate:

$$FNF = 1 - CNF$$

Correct Positive Fraction or True Positive Rate:

$$CPF = \frac{CP}{CP + IN}$$

True and false positive rates allow building a plot for the different thresholds. Thereby the model accuracy can be evaluated by estimating the Area Under the Curve as illustrated in Figure 2.

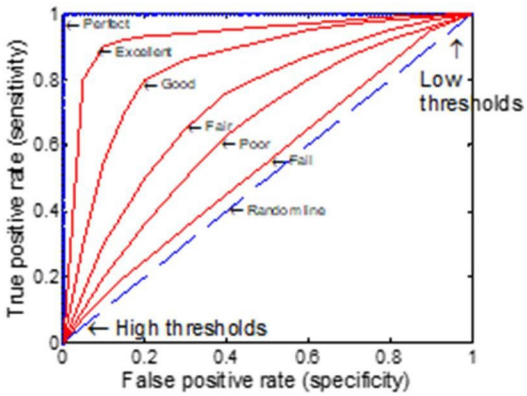

Figure 3. Receiver Operating Characteristic (ROC) plots of model performance

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

Performance	AUC Roc curve
Excellent	0.9 - 1
Good	0.8 – 0.9
Reasonable	0.7 – 0.8
Poor	0.6 – 0.7
Bad	0.5-0.6

Table 2. Evaluation of model's performance according to the Roc, by measuring the area under the curve (AUC), [http://gim.unmc.edu/dxtests/roc3.htm]

Positive and negative probabilities of correct decision. In case where a decision has to be taken based on thresholds (e.g. install oyster spat collectors or harvest seaweeds when the water temperatureis above or below a given threshold) it can be useful to know the probability of a correct decision. The probability that a positive decision (value > threshold) or that a negative decision (value < threshold) is correct can be computed with the Positive Predicted Value (PPV) and the Negative Predicted Value (NPV) respectively, as follow:

$$PPV = \frac{CP}{CP + IP} \qquad NPV = \frac{CN}{CN + IN}$$

NPV and PPV values are between [0,1], where high values indicate that a decision can be trusted, whereas low values suggest the decision should be regarded with suspicion.

3. Classical hindcast validation

One of the main objectives of hindcast validation should be to demonstrate that the new FORCOAST service model provides an improvement over the parent model in which it is nested. In this sense, the same procedures (e.g. observational datasets, quality assessment metrics) should be followed to validate the parent model and the new FORCOAST service model. In particular, when the new system is taking boundary conditions from a product delivered by CMEMS (Copernicus Marine Environment Monitoring Service), the hindcast validation should apply the same methods as in the latest published QUID (Quality Information Document).

Below, methods to be followed for different ocean platforms and devices are specified. The examples provided below concern mainly physical parameters, but the same approach should be undertaken when validating biogeochemical variables, e.g. in-situ dissolved oxygen, chlorophyll-a or gridded chlorophyll datasets from satellite ocean color data, etc.

3.1. Time series

A time series is any sequence of data points ordered in time. The metrics in section 2 can be determined if $x = \{x_1, x_2, \ldots, x_N\}$ is taken as the observed time series and $y = \{y_1, y_2, \ldots, y_N\}$ as the predicted time series. Ocean sensors producing time series are:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

Tidal Constituent		Gauge		NEA_ROMS	
		Amp	Phase	Amp	Phase
Aranmore	M2	1.22	161	1.29	156
	S2	0.44	194	0.46	189
	N2	0.25	139	0.26	135
	K1	0.12	154	0.14	142
	O1	0.08	359	0.08	5
	Q1	0.03	299	0.03	292
Ballycotton	M2	1.42	148	1.26	146
	S2	0.44	193	0.41	190
	N2	0.26	128	0.23	128
	K1	0.02	178	0.02	178
	O1	0.03	35	0.02	55
	Q1	0.01	340	0.01	348
Ballyglass	M2	1.16	158	1.21	152
	S2	0.41	191	0.43	185
	N2	0.23	136	0.25	131
	K1	0.14	122	0.16	116
	O1	0.09	337	0.08	342
	Q1	0.03	277	0.03	268
Castletownbere	M2	1.12	131	1.09	124
	S2	0.36	161	0.36	155
	N2	0.23	108	0.22	101
	K1	0.04	50	0.05	49
	O1	0.01	279	0.02	277

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

	Q1	<0.01	185	<0.01	184
Dunmore East	M2	1.38	150	1.20	150
	S2	0.45	199	0.43	198
	N2	0.25	133	0.22	136
	K1	0.04	178	0.04	177
	O1	0.04	29	0.03	49
	Q1	0.01	344	0.01	341
Galway	M2	1.57	141	1.64	137
	S2	0.55	172	0.58	170
	N2	0.32	119	0.34	117
	K1	0.09	76	0.10	79
	O1	0.06	311	0.05	311
	Q1	0.02	261	0.02	243
Howth	M2	1.44	325	1.39	318
	S2	0.41	357	0.41	349
	N2	0.28	297	0.28	289
	K1	0.10	198	0.11	187
	O1	0.08	37	0.07	54
	Q1	0.03	343	0.03	340
Malin Head	M2	1.12	178	1.19	172
	S2	0.42	206	0.44	201
	N2	0.23	156	0.26	151
	K1	0.09	167	0.11	154
	O1	0.07	7	0.07	18
	Q1	0.03	310	0.03	306

Table 3. Example of tidal harmonic analysis validation. The amplitudes in meters and phases in degrees for six of the principal tidal constituents calculated for the measured and modelled data (Nagy et al., 2020).

1. Tidal gauges, measuring the time-varying sea level. In this case, in addition to the regular time-series validation described above, a tidal harmonic analysis should be carried out. The objective of such harmonic analysis is to obtain a decomposition of the tidal signal into several harmonic constituents characterized by a given amplitude, frequency and phase.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

Harmonic analysis of both the observed (tidal gauge) and predicted (model) sea level time series should yield similar results. An example is shown (Table 3).

2. Sensors attached to moorings and buoys.

3. Acoustic Doppler Current Profilers (ADCPs) produce several time series of current velocity at given depths. Each time series can be treated separately and decomposed into a *u*-time series and a *v*-time series.

4. Drifters and gliders produce along-trajectory time series of various ocean parameters and the same general procedure can be applied. In the case of gliders, it is also desirable to provide transect plots side by side to compare observed versus predicted results.

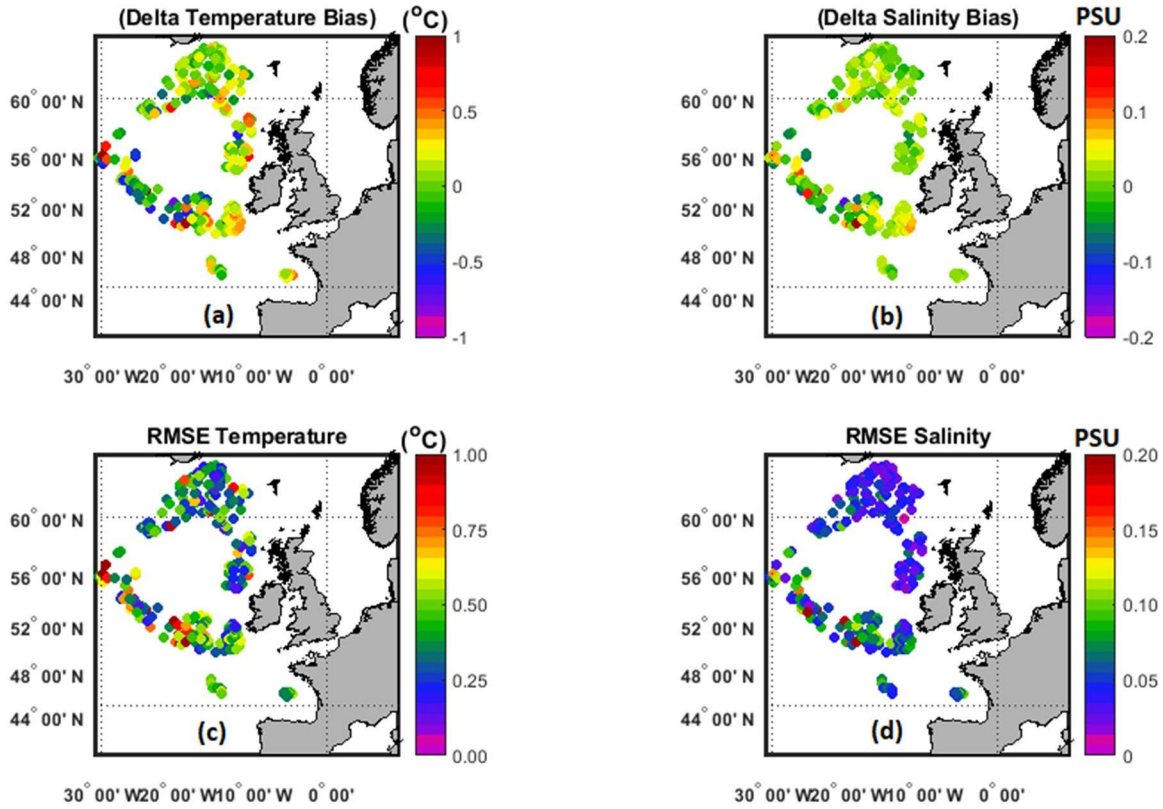

Figure 4. Quality assessment metrics (RMSE, bias) determined separately for 304 Argo float profiles locations and conveniently represented in a colored scatter plot for model evaluation (Nagy et al., 2020)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

3.2. Vertical profiles

Observations from Argo buoys, CTDs and rosette samplers are presented as vertical profiles, where multiple measurements are taken along the water column. Metrics from section 2 will be determined taking $x = \{x_1, x_2, \ldots, x_N\}$ as measurements, $y = \{y_1, y_2, \ldots, y_N\}$ as predictions and N as the number of data points along the water column. In this way, a set of metrics will be obtained for each profile and for each measured parameter (e.g. salinity, temperature, nutrients, etc.). Results can be shown in a colored scatter plot (Fig. 3). Also, a T-S diagram presenting observations versus predictions should be provided (Fig. 4) as well as vertical profile plots (Fig. 5).

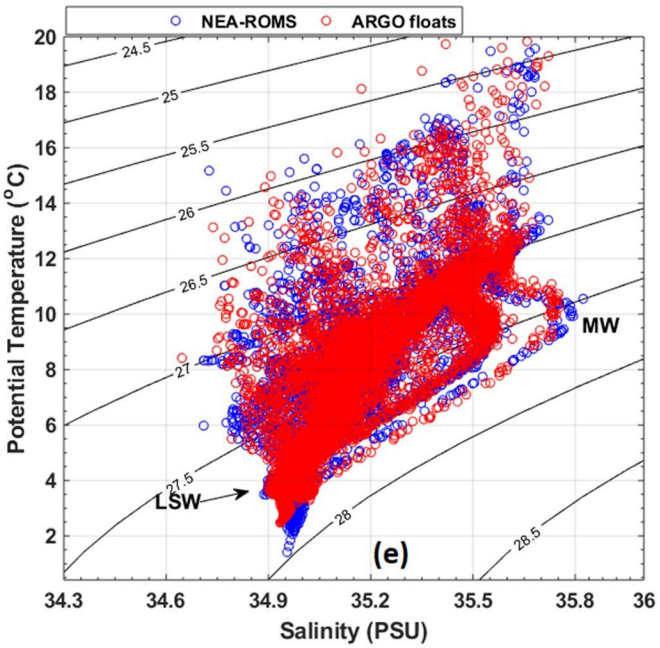

Figure 5. T-S diagram comparing NEA-ROMS model and Argo float observations (Nagy et al., 2020).

3.3. Gridded datasets

Remote sensing data, such as satellite altimetry or HF-Radar data, are often delivered as gridded datasets, where data from multiple points on the ocean surface are provided as a single product. For every grid node inside the model domain, it is possible to extract the corresponding time series and thus give it the same treatment as in section 3.1. As a result, a spatial distribution of the metrics in section 2 can be obtained and conveniently plotted for evaluation. In addition, the time series of the spatially-averaged data and quality assessment metrics can be obtained and conveniently plotted for evaluation (Fig. 6).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

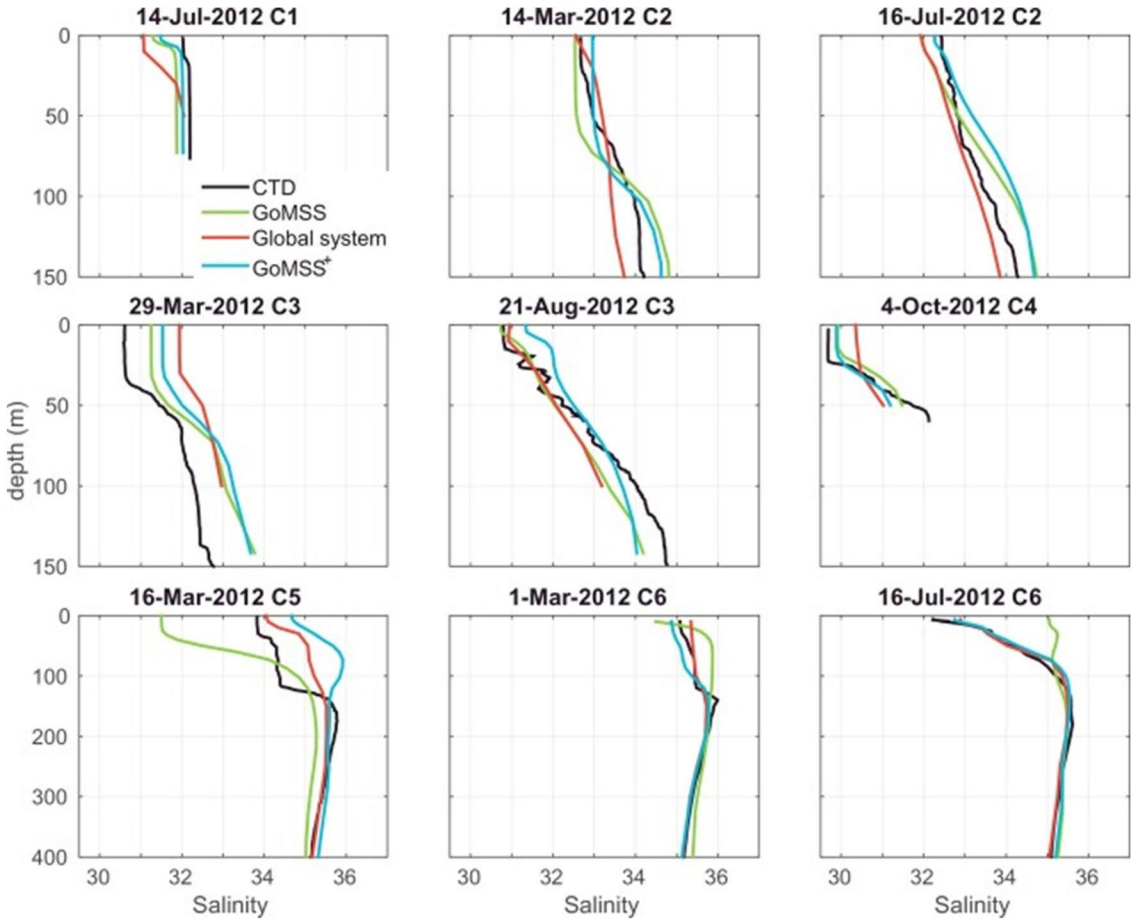

Figure 6. Vertical profiles of salinity from CTD and ocean models (Katavouta and Thompson, 2016)

3.4. Local sampling

Local sampling refers to any collection of measurements taken at different times and locations. For example, water sampling for chemical analysis along a latitudinal transect inside the model domain. In this case, observations $x = \{x_1, x_2, \ldots, x_N\}$ refer to measurements at each of the N sampled locations, predictions $y = \{y_1, y_2, \ldots, y_N\}$ are taken from the model using interpolation, and metrics in section 2 are calculated accordingly. Similar colored scatter plots as in Fig. 3 can be produced.

3.5. Trajectories

In the framework of the FORCOAST project, particle-tracking models can be useful to predict the paths followed by sewage discharges or larvae. Predictions can be compared with the trajectories followed by ocean drifters and a suitable validation methodology is the one described by Liu and Weisberg (2011), which is based on the definition of a Normalized Cumulative Lagrangian Separation.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

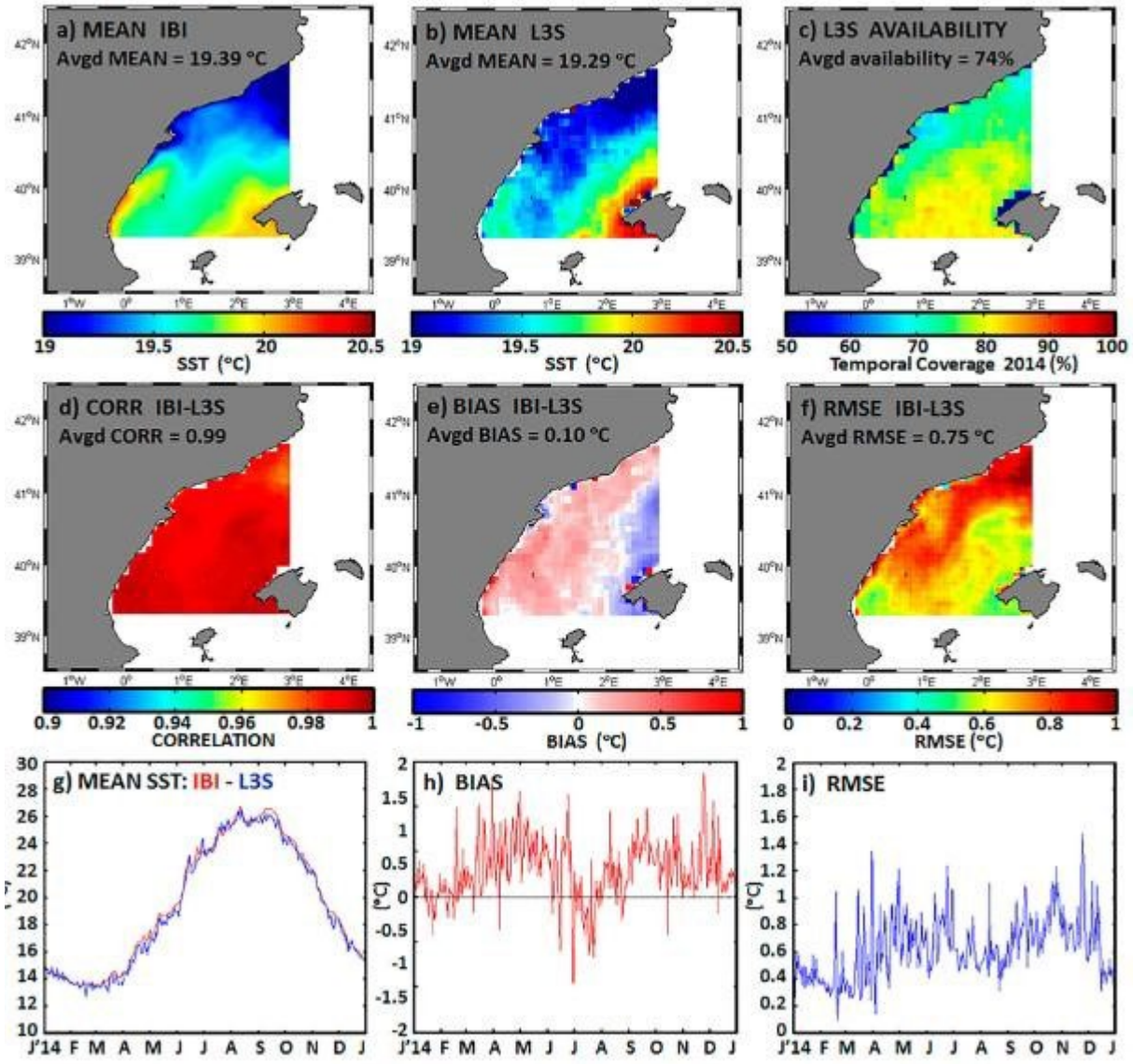

Figure 7. Example of validation methodology for L3 satellite-derived SST observations vs. the IBI-MFC model, where subplots a-f show annual statistics for year 2014, while subplots g-i show spatially-averaged data and quality assessment metrics (Lorente et al., 2016)

4. Forecast validation

Since forecast validation can become a massive task, fluent communication between scientists and stakeholders is of paramount importance to decide which parameters (e.g. salinity, dissolved oxygen concentration) and which locations are of greatest interest to end users. Then, **forecast validation should focus on the agreed parameters and locations.**

In general, the same approaches as in section 3 can be applied to forecast validation. However, it is expected that the model's performance will decrease during the forecast length, with higher accuracy at the beginning of the forecast and lower-quality predictions towards the end of the forecast. Therefore, it is possible to divide the forecast into three equal-length subsets and calculate the metrics

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

in section 2 to each of them separately. This procedure would provide end users with an estimation of the forecast quality and reliability at three different stages: beginning, middle and end of the forecast period.

It is proposed that the exact details of the forecast validation in each Pilot are agreed upon with local end users and concern only the parameters that will be used locally in a forecast mode. As pointed out above, a **general framework would follow the hindcast validation** described in section 3. For example, in the case of a 3-day forecast, such forecast can be validated for three forecast horizons:

a. Day 1 – follows the same approach as in the hindcast, except that the model output comprises forecast day 1 predictions.

b. Day 2 - follows the same approach as in the hindcast, except that the model output comprises forecast day 2 predictions.

c. Day 3 - follows the same approach as in the hindcast, except that the model output comprises forecast day 3 predictions.

5. Process-oriented validation

The objective of process-oriented validation is to assess the ability of the model to reproduce specific oceanographic processes (e.g. coastal upwelling, phytoplankton blooms, extreme salinity or temperature events, etc.) that are of particular interest to the stakeholders. Therefore, it can be considered the most important step in the ocean model validation, as it will be directly linked to the services offered in the FORCOAST platform.

Firstly, it is necessary to determine what characterizes the oceanographic process under consideration. For example, oyster farmers are concerned about sudden drops in salinity –for instance, under high river runoff conditions–, which pose a serious threat to the health of marine bivalves. It is possible to define a threshold salinity S_t such that, when salinity drops below S_t, it means that an extremely low salinity event is taking place. Here, the goal of process-oriented validation would be to evaluate how well the model predicts such events in which salinity drops below S_t. Obviously, observations in the locations of interest to the stakeholders are needed to carry out this type of validation.

Furthermore, in order to determine the appropriate thresholds used as an indicator of a given oceanographic process of interest, the expertise of both scientists and stakeholders is required. **The Receiver Operating Characteristic (ROC)** analysis described in section 2.5 is particularly suited for this form of validation and is recommended as a common approach to process-oriented validation across the Pilots.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

6. Model validation strategy per pilot

The table below presents a summary of the foreseen validation efforts at each site. For each Pilot, a process-oriented validation will ensure that the model is fit for purpose in relation to the services being develop.

Pilot	Strategy
1: Portugal	**Observations:** 1. Continuous observations from the new ExporSado Longa monitoring station, recording sea level, seawater temperature and salinity, pH, suspended sediments, dissolved oxygen and chlorophyll concentration. 2. IH tide gauges sea level series. 3. APA data collected under the Water Framework Directive program, in several areas of the estuary, including water temperature, salinity, oxygen, etc. 4. Remote-sensing sea surface temperature. **Metrics:** ME, CORR, RMSE. **Hindcast validation:** May 2018 onwards. **Forecast validation:** 48-hour forecasts from April 2022, focusing on the accuracy of water level predictions. **Process-oriented validation:** Wind and atmospheric pressure validation together with water levels to evaluate the effective working period. Salinity to evaluate the impact of river discharges in the production area.
2: Spain	**Observations:** 1. Slope Donostia Buoy, a deep-water mooring providing hourly temperature, salinity and current measurements at different levels, from 10 meters to 200 meters depth. 2. High-Frequency Radar measuring surface velocities (5 km X 5 km) on an hourly basis. 3. Daily, remote-sensing sea surface temperature (0.02° x 0.02°). **Metrics:** ME, RMSE, CORR, ARMAE. **Hindcast validation:** Not required, as the Service Module of interest in this region is the SM-F2 Front Detection and its focus is on forecasting mesoscale fronts. **Forecast validation**: 96-hour forecasts with the focus on the period from 10-Jun to 04-Sep 2021. Observational datasets (1) and (2) above to be used for forecast validation. **Process-oriented validation**: Remote-sensing SST to be used determine frontal occurrence and compare with model forecasts.
3: Bulgaria	SM-F1 Fishing Suitability Index released for this area relies on sea surface temperature and sea surface salinity measurements from the CMEMS' BLKSEA_MULTIYEAR_PHY_007_004, whose validation is already published in the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

	QUID https://catalogue.marine.copernicus.eu/documents/QUID/CMEMS-BS-QUID-007-004.pdf. Therefore, the validation will focus on the WBS wave hindcast system only. **Observations:** 1. Sentinel-3a, Sentinel-3b, Cryosat-2, SARAL/Altika, Jason-3, CFOSat, and HaiYang-2b from product WAVE_GLO_WAV_L3_SWH_NRT_OBSERVATIONS_014_001. 2. INSITU_BS_NRT_OBSERVATIONS_013_034 **Metrics:** RMSE, bias, Scatter Index (SI), Pearson correlation coefficient (CORR), and best-fit Slope (SLOPE) **Hindcast validation:** Significant Wave Height from Jun-2019 to May-2021 using both datasets above (1) remote-sensing and (2) in-situ data.
4: Belgium	**Observations:** in-situ data from multiple stations throughout the North Sea (Aberdeen, Akkaert, Barmouth, Bournemouth, Cadzand, Cuxhaven, Den Helder, Europlatform, Helgoland, Hoek van Holland, Ijmuiden, K13, Newport, Ostend, Roompot buiten, Stavanger, Vlakte van de Raan, Vlissingen, Wandelaar, Westhinder, Zeebruges). The parameters that will be considered are the sea level, seawater temperature, seawater salinity, significant wave height and zero up-crossing frequency. **Metrics**: ME, CORR, RMSE, ratio of standard deviations **Hindcast validation:** 2013-2017 for sea level, seawater temperature, seawater salinity, significant wave height and zero up-crossing frequency. **Forecast validation:** Forecasting skill of the MFC's operational models to be consulted in the website of NOOS (http://noos.eurogoos.eu/) **Process-oriented validation:** Will focus on the temperature measurements from the Westdiep buoy against the BCZ model.
5: Ireland	**Observations:** 1. Galway Port tide gauge (sea level). 2. Multi-Scale Ultra-High Resolution (MUR) Sea Surface Temperature. 3. Three ADCP moorings from spring and summer 2018. 4. Three moored CPT loggers recording temperature and salinity 5. CTD casts recording temperature and salinity throughout the bay on quarterly basis, starting on May 2021 (forecast validation). **Metrics**: ME, CORR, RMSE, ARMAE, ROC curve **Hindcast validation:** Spring and summer 2018 for the ADCP recordings, Oct 2019 – Sep 2020 for the other datasets. **Forecast validation**: From Jan 2021 onwards, assessing the 72-hour forecasts against the CPT logger temperature and salinity recordings and against quarterly CTD casts. **Process-oriented validation**: Focused on low salinity events triggering oyster mortality, using weekly salinity measurements from 2014 until present at the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

	Killeenaran pier, and applying the ROC curve analysis.
6: Denmark	**Observations:** in-situ data from multiple stations throughout the Limfjord. The parameters that will be considered are the sea level from tide gauge stations, CTD profiles from the NOVANA environmental monitoring cruises, remote-sensing SST, and biogeochemical in-situ observations including DIN, PO_4, chlorophyll concentration and dissolved oxygen concentration at different sites. **Metrics:** ME, RMSE, CORR, ARMAE **Hindcast validation:** 2015-2019 historical data **Forecast validation:** Operational period, from Mar-2021 to present, using profile observations of surface seawater temperature and remote-sensing SST. **Process-oriented validation:** 2015-2019 data, focusing on storm surges at the Lemvig tide gauge site and on diurnal warming in shallow waters at the aquafarming site KFO1.
7: Romania	**Observations**: 1. Remote-sensing data (SST and CHL) directly downloaded from CMEMS and/or available/tailored for the Pilot. 2. ADCP, in-situ temperature and salinity measurements collected by NIMRD. **Metrics**: ME, CORR, RMSE **Hindcast validation**: The focus will be on the period from 2019 until present. **Forecast validation**: Not required, as the forecast forcing is the same as in the hindcast. **Process-oriented validation**: The focus will be on ADCP currents, as the SM-A2 Land Pollution relies on model currents.
8: Italy	**Observations:** 1. Remote-sensing data (SST and CHL) directly downloaded from CMEMS and/or available/tailored for the Pilot. 2. Coastal sampling stations data (T, S, nutrients, CHL) provided by ISPRA (Italian Institute for Environmental Protection and Research) being updated for the period 2018-2021. **Metrics:** BIAS, RMSE, standard deviations, number of SST and CHL points. **Hindcast validation:** Updated reanalysis for the 2006-2017 period and further extension to the 2018-2021 period. **Forecast validation:** Daily NRT validation metrics for each (daily) first day forecast to produce incremental plots starting in March 2022. The last available daily surface maps of SST and CHL will also be plotted for visual comparison (model/satellite). **Process-oriented validation:** Focused on heat waves and oxygen depleted conditions.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.

7. Conclusions

This deliverable provides guidance on how to validate ocean models within the FORCOAST project with the aim of harmonizing procedures across different partners. In order to assess model accuracy, different quality assessment metrics have been proposed. Methods to be followed for hindcast, forecast and process-oriented validation have been specified as well. Model strategy validation plans have been presented for each site. Assessment of coastal model performance following the standards described in this deliverable will result in the production of a final coordinated pilot model evaluation report (D5.4).

8. References

Allen, J.I., Holt, J., Blackford, J. and Proctor, R., 2007. Error quantification of a high-resolution coupled hydrodynamic-ecosystem coastal-ocean model: Part 2. Chlorophyll-a, nutrients and SPM. Journal of Marine Systems, 68(3-4): 381-404.

Allen, J.I., Smyth, T.J., Siddorn, J. and Holt, J., 2008. How well can we forecast high biomass algal bloom events in a eutrophic coastal sea? Harmful Algae, 8(1): 70-76.

Brown, C.D. and Davis, H.T., 2006. Receiver operating characteristics curves and related decision measures: A tutorial. Chemometrics and Intelligent Laboratory Systems, 80(1): 24-38.

Dabrowski, T., Lyons, K., Nolan, G., Berry, A., Cusack, C., Silke, J., 2016. Harmful algal bloom forecast system for SW Ireland. Part I: Description and validation of an operational forecasting model. Harmful Algae 53:64-76. doi:10.1016/j.hal.2015.11.015

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters, 27: 861-874.

Katavouta, A., Thompson, K.R., 2016. Downscaling ocean conditions with application to the Gulf of Maine, Scotian Shelf and adjacent deep ocean. Ocean Model. 104:54-72. doi:10.1016/j.ocemod.2016.05.007

Liu, Y., Weisberg, R.H., 2011. Evaluation of trajectory modeling in different dynamic regions using normalized cumulative Lagrangian separation. J. Geophys. Res. 116, C09013. doi:10.1029/2010JC006837

Lorente, P., Piedracoba, S., Sotillo, M.G., Aznar, R., Amo-Balandron, A., Pascual, A., Soto-Navarro, J., Álvarez-Fanjul, E., 2016. Ocean model skill assessment in the NW Mediterranean using multi-sensor data. J. Oper. Oceanogr. doi:10.1080/1755876X.2016.1215224

Nagy, H., Lyons, K., Nolan, G., Cure, M., Dabrowski, T., 2020. A regional model for the North East Atlantic: model configuration and validation. J. Mar. Sci. Eng. 8, 673. doi:10.3390/jmse8090673

Sutherland, J., Walstra, D.J.R., Chesher, T.J., van Rijn, L.C., Southgate, H.N., 2004. Evaluation of coastal area modelling systems at an estuary mouth. Coast. Eng. 51:119-142. doi:10.1016/j.coastaleng.2003.12.003

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870465.