HiSea Deliverable 3.5

Data processing algorithms

Work Package Number: 3

Work Package Title: Service Specifications and User Requirements

HISea Project Information	
Project full title	High Resolution Copernicus-Based Information Services at Sea for Ports and Aquaculture
Project acronym	HiSea
Grant agreement number	821934
Project coordinator	Dr. Ghada El Serafy
Project start date and duration	1st January, 2019, 30 months
Project website	https://hiseaproject.com/

Deliverable Information	
Work package number	3
Work package title	Service Specifications and User Requirements
Deliverable number	3.5
Deliverable title	Data processing algorithms
Description	Description of the general approach and specific methods used in the HiSea project concerning data processing.
Lead beneficiary	Ascora
Lead Author(s)	Daniel Wegmann (Ascora)
Contributor(s)	Danny Pape (Ascora), Anna Spinosa (Deltares), Philippe Bryère (ARGANS-F), Pedro Galvão (Hidromod)
Revision number	0.4
Revision Date	20.03.2020

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

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))	P

Document History			
Revision	**Date**	**Modification**	**Author**
0.1	25.02.2020	Initial draft	Daniel Wegmann
0.2	28.02.2020	Added chapter 3	Danny Pape
0.3	19.03.2020	Merging partner inputs and revising previous entries	Daniel Wegmann
0.4	20.03.2020	Added executive summary and rewrote sections	Daniel Wegmann
0.4	24.03.2020	Review	Sandra Gaytan
0.5	24.03.2020	Review	Adélio Silva

Approvals				
	Name	**Organisation**	**Date**	**Signature (initials)**
Coordinator	Ghada El Serafy	Deltares	31-03-2020	GES
WP Leaders	Danny Pape	Ascora	31-03-2020	DP

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

Executive Summary

The data sources presently being used by the HiSea project includes remote sensing data sources (such as Copernicus), in-situ data and modelling data that is used to produce user focused information.

This Deliverable intends to provide an overview of the various techniques employed by the HiSea partners to gather, process and store data. It provides a general description of the overall methodology as well as in-depth descriptions of algorithms and models that are being utilized to be able to properly deal with it. This includes the need to deals with the large amount of data daily produced and to make usable by non-expert users data that, in some cases, may start to seem quite complex and, even more important, to transform these data sets into useful information.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

Content

1 HiSea data and services 6

2 Data Sources 7

3 Algorithms 9

3.1 Pre-Processing 9

3.1.1 Outliers 10

3.1.2 Box-and-Whisker plot 10

3.1.3 Tietjen-Moore test 11

3.1.4 Generalized Extreme Studentized Deviate (GESD) test 11

3.1.5 Smoothing 12

3.1.6 Data Association 12

3.1.7 Spatial Alignment 13

3.2 High resolution models 16

3.3 Post-Processing 20

3.3.1 δ-MAPS method 21

3.3.2 Density, cluster and neighbourhood influence 21

3.3.3 Serial correlation 22

3.3.4 Ensemble forecasting 22

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

1 HiSea data and services

HiSea is daily handling large data files (from Earth Observation, local measurements and models) which poses some relevant challenges in relation with different processes such as storage, format harmonization, quality control and data managing. Being the final goal to offer the users a service delivering easy to understand information, even if derived from data sets that may assume complex forms (in terms of dimension and/or parameters), the use of proper models and algorithms capable to properly work the data to extract the required information represents a critical component.

The algorithms that are the main focus of this document may be useful already during the data pre-processing phase (data quality control, data fusion, etc.) but they will be necessary all along the whole service chain to perform data analytics, to compute key performance indicators, to assimilate data into models or to add intelligence to the system allowing to cross different data sources to extract relevant information or to foresee critical situations.

An algorithm may be something so simple such as to apply some kind of data formatting procedure to harmonize the data sets to make them interoperable or something quite complex such as a neural network learning with the accumulated daily experience.

Within HiSea different algorithms are being used along the service chain. They start to be used at the data pre-processing level to handle the data acquired from different data providers (this usually means different formats, different time and space resolutions and even different levels of accuracy) and merge them to provide a harmonized view of the processes to be addressed. This, beyond the formatting harmonization process, also means to perform another type of analysis to identify the potential existence of data gaps, data outliers or trends that may indicate a loss of accuracy along time.

Once all data sets are considered ready to be used another set of algorithms may be used to perform different types of statistical analysis or to extract some relevant indicators for the user activity. Similar statistical analysis algorithms are also used to keep a continuous check on the high resolution models performance by performing models results validation analysis against the available field data.

Finally, in the service chain last phase, in the presence of all data sets and model results, another set of algorithms are used to extract focused information from the large amount of data gathered along the whole process. This information may be the result of combining different data sets (for instance analyse a number of parameters to derive an indicator that may trigger an alert) or just the result of the analysis of a single data set (for instance transform a continuous time series in a sequence of coloured indicators).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

2 Data Sources

As referred, in Hisea several different data sources (Earth Observation (EO), local sensors, modelling, etc.) are used to derive the information that is distributed via HiSea Platform. The data sets produced from these different data sources usually require some type of treatment (transformation/adaptation) to make them usable by the common user either because they are too complex or too large. In HiSea context these kinds of procedures are being applied to EO data, global and regional modelling data from (Copernicus, NOAA, etc.) and to some data sets being acquired by local sensors in "near" real time. In this chapter some of these procedures are detailed.

Concerning EO data presently the main provider is Copernicus. This programme is based on a composition of dedicated satellites, contributing mission, services and in-situ data and is the largest space data provider in the world. There are 5 dedicated satellite families, which are:

- Sentinel-1 for a high-quality resolution
- Sentinel-2 for optical vegetation
- Sentinel-3 for optical/thermal information of oceans
- Sentinel-5P for atmosphere measurements word-wide
- Sentinel-6 for sea level elevations

Whereas Sentinel-4 (air quality in Europe) and Sentinel-5 (atmosphere measurements world-wide) are measuring instruments on a EUMETSAT Satellite. The data presently acquired by these satellites, together with data produced by global and regional models are made available through different operational services:

- Copernicus Atmosphere Monitoring Service (CMAS) provides continuous data and information on atmospheric composition.
- Copernicus Marine Environment Monitoring Service (CMEMS) provides regular and systematic reference information on the physical and biogeochemical state, variability and dynamics of the ocean and marine ecosystems.
- Copernicus Land Monitoring Service (CLMS) provides geographical information about land cover and its changes, land use, vegetation state, etc.
- Copernicus Climate Change Service (C3S) provides authoritative information about the past, present and future climate.
- Copernicus Security Service (CSS) improves the surveillance of borders and maritime and the support to EU external actions.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

- Copernicus Emergency Management Service (CEMS) consists of mapping and an early warning component to manage natural disasters, man-made emergencies, and humanitarian crises.

HiSea is making the most use of Copernicus Marine Environment Monitoring Service both through the direct use of some EO data sets (SST, Cl-a, Turbidity, etc.) and through the use of global or regional modelling results.

Another part of HiSea data is coming from local sources either via data locally acquired or via high resolution models. In all cases, it is usually required to make some kind of data working to fit it to the users or high-resolution models requirements. These data sets have been the object of a detailed description in *D6.3 Service Specification and User Requirements* and mostly refers to:

- Meterology
- Waves
- Currents
- Water level
- Water turbidity
- Water temperature
- Salinity
- Oxygen level
- pH
- Chlorophyll-a

Some of these parameters are measured over an area with a low time frequency (case of satellite data or drones missions), over an area with a high time frequency (case of radars), locally with a high time frequency (case of sensors) or locally with a low time frequency (case of local sampling).

Once the data gathering process is completed different algorithms are required to fit the data to the areas of interest, transform signals in an easy to use parameter (ex. transform a frequency acquired by a satellite in turbidity), adjust formats, check for consistency, etc..

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

3 Algorithms

3.1 Pre-Processing

As previously referred the use of algorithms starts with the acquisition of data to the platform. Many data sources are being imported from different providers that require to be properly worked to be possible to make them available to be used in the platform.

The procedures taking place in this phase involves proper actions to, i) assure that all data sets are kept in NetCDF –CF format; ii) assure a first automatic screen to detect eventual issues that require some action (data gaps, outliers, etc.); iii) assure that the data that will be required to be used by the high resolution models has the proper space and time resolutions.

The CF metadata conventions[1] are designed to promote the processing and sharing of files created with the NetCDF API. The conventions define metadata that provides a definitive description of what the data in each variable represents and the spatial and temporal properties of the data. This enables users of data from different sources to decide which quantities are comparable and facilitates building applications with powerful extraction, regridding, and display capabilities.

The standard is fully documented by a PDF manual accessible from a link from the CF metadata homepage[2]. Note that CF is a developing standard and access via the homepage, rather than through a direct URL, is recommended to ensure that the latest version is obtained. The current version of this document was prepared using version 1.6 of the conventions dated 5 December 2011. The approach adopted in HiSea follows the SeaDataNet profile based on CF-1.6:

- Profile (x, y, t fixed; z variable). The specification given is for the storage of a single profile such as a CTD cast or bottle profile. However, the design is such that very little change is required to facilitate the storage of multiple profiles in a single netCDF file.
- TimeSeries (x, y, z fixed; t variable). The specification given is for the storage of a single time series, such as a current meter record. However, the design is such that very little change is required to facilitate the storage of multiple time series in a single netCDF file.

[1] https://cf-trac.llnl.gov/trac
[2] https://cf-trac.llnl.gov/trac

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

- Trajectory (x, y, z, t, all variable). The specification given is for storage of a single trajectory, but this may be easily modified to store several trajectories in a single file.

3.1.1 Outliers

An outlier is defined as an observation extremely far from the expected value. Outliers may have importance and can explain the events that happen. However, computed quantities such as average and least squares regression can be badly affected by such values. The presence of outliers can also change the fit estimates and predictions for predictive models. For this reason, the identification and removal of outliers are of high importance to get a representative data set.

Often, the decision on how to treat the outliers is left to the best judgement of the analyst, and threshold is applied to clean the dataset. Test and plot can also be used to identify the outliers, among which the Box-and-Whisker plot, the Tietjen-Moore test and the Generalized Extreme Studentized Deviate (GESD) test.

3.1.2 Box-and-Whisker plot

The Box-and-Whisker plot is a tool with a specific objective to visualize outliers. The Box-and-Whisker plot displays the lower, median and upper quantiles of the data set at the 'box', and the minimum and maximum excluding the outliers at the 'whiskers'. The outliers are shown as separate plotted points. The Box-and-Whisker plot is represented in Figure 1, and as an example of its application, Figure 2 shows the use of the Box-and-Whisker box in the representation of shoreline regression along the Dutch coastline.

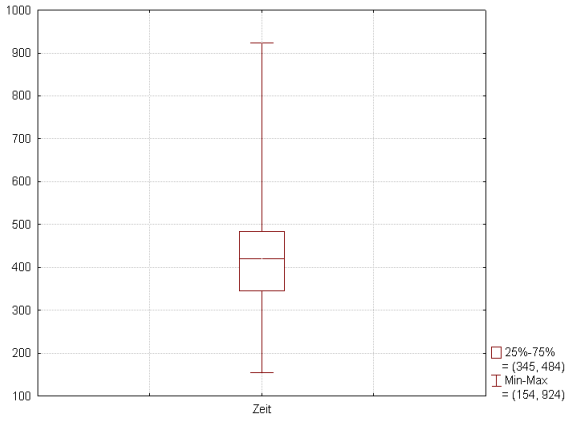

Figure 1: Boxplot with whiskers from minimum to maximum. Source: Wikipedia

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

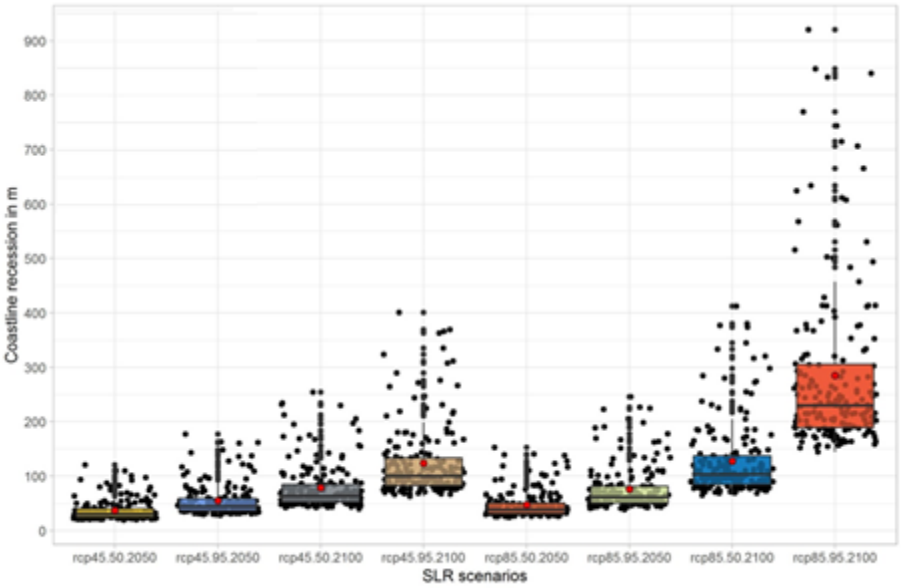

Figure 2: Representation of the recession rates along a transect for the Dutch coastal area. The X-axis shows the eight scenarios. Y-axis shows the total coastline recession in meters for the given year compared to 2018. The horizontal location of the scattered black dots is based on random variation to handle overplotting as a result of the discreteness of the dataset. Red dots are the average values that correspond with the values in the table below the graph

3.1.3 Tietjen-Moore test

The Tietjen-Moore test (1972) is a generalization of the Grubbs test used to detect more than one outlier in a univariate data set that follows an approximately normal distribution. The R package 'EnvStats' is used to perform the test. The outlier test hypothesis is given as:

- Null hypothesis: Ho = There are no outliers in the data set
- Alternative hypothesis: H1 = There are exactly k number of outliers in the data set

3.1.4 Generalized Extreme Studentized Deviate (GESD) test

Similar to the Tietjen-Moore test, the GESD test, also known as the Rosner's test [1] is used to search for outliers in univariate data that is normally distributed. An advantage of this method over the former is that it requires only an upper bound for the suspected number of outliers be specified. Therefore, the alternative hypothesis for this test is; H1 = There are up to r outliers in the data set.

The R package 'climtrends' and 'EnvStats' are used to perform these tests. Tests are implemented also on in-situ measurements use for validation purpose.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

3.1.5 Smoothing

Data sets (in situ measurements, satellite retrieved variables, modelled data) contain noise or random variation (outliers). Noisy data can be filtered or reduced its influence through various data smoothing techniques. Moving average is the simplest form of data smoothing. However, at the edge of the window data are lost (Janert, 2010). The window defines how much 'memory' is contained in the moving average. Larger window width causes the value of the moving average calculated to change slowly, which means it is more influenced by the past values, and less influence from short term fluctuations. (Marshall, 2012). The dimension of the window can be computationally limited.

Another method of smoothing is the locally estimated scatterplot smoothing (LOESS) by Cleveland and Devlin (1988), a very flexible non-parametric smoother. A low degree polynomial is fitted to each subset of the data. An advantage of this method is that is only the smoothing parameter value and degree of polynomial has to be provided (NIST, 2013; Jacoby, 2000). An application of LOESS and moving average window is shown in Figure 3, where total inorganic matter values are smoothed.

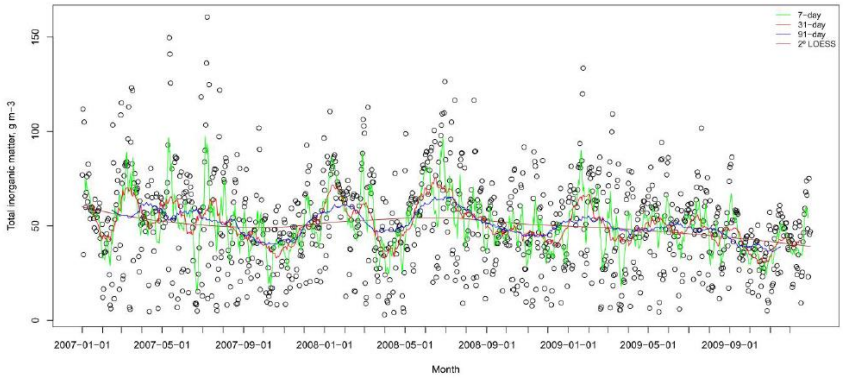

Figure 3: Example of Moving average window

3.1.6 Data Association

Data association is needed to determine if the set of measurements correspond to each target and therefore if the set of observations or measurements is generated by the same target. Data association can be computed using Nearest Neighbors (NN) and K-mean technique. As described before, NN is a simple algorithm able to recognize a pattern in a cluttered environment. A conceptual framework of NN is shown in Figure 5. K-means is a variation of the NN algorithms that divide the dataset values into clusters and compute the best centroid of the cluster.

More complex algorithms can also be applied to the input dataset to perform data association, among which the Probabilistic Data Association (PDA) and Joint Probabilistic Data Association (JPDA). PDA

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

associate probability to each hypothesis from a valid measure input, generally in situ measured. JPDA is similar to PDA, with the difference that the association probabilities are computed using all of the observations and all of the targets. Thus, in contrast to PDA, JPDA considers various hypotheses together and combines them (Castanedo, 2013).

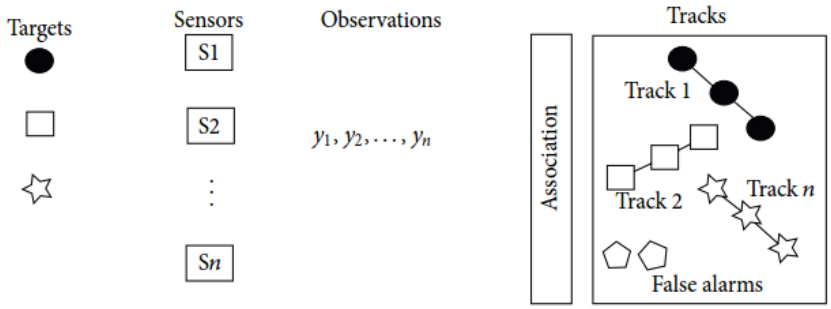

Figure 4: Conceptual overview of the data association process from multiple sensors and targets. Source: Castanedo, 2013

3.1.7 Spatial Alignment

Data collected from several sources can be available in different coordinate system, therefore a conversion of the data to a common coordinate system can be required. This process involves geoprocessing and geo-referencing of the locations. The latest can be performed using gdal package in Python or QGIS, as well as in Matlab and R. Visualization of the grids involved in re-projection can also be performed, as shown in Figure 5.

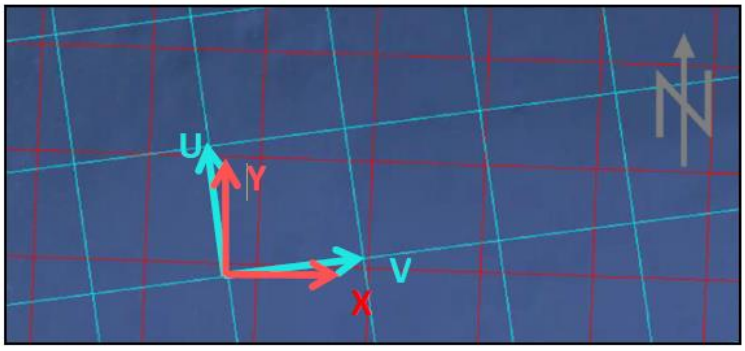

Figure 6: Visualization of the grids Involved in re-projection. The orientation of the plain presents the direction of the original wind velocity data, the light-blue grid the u/v orientation and the red grid the x/y orientation.

Resampling and Interpolation Methods

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

The spatial resolution of data from multiple sources can highly differ. For this reason, the resampling method can be applied. Common resampling methods for raster file data are NN, bilinear and cubic transformation. Data could also be resampled to a common pre-defined grid using interpolation techniques to find the values in the nodes or the center of the grid. Symmetrical and

Separable interpolation kernels are used to reduce computational complexity, Figure 7 shows common kernel functions used for image interpolation

Name	$h(x)$
Nearest Neighbour	1 if $0 \leq \lvert x\rvert < \frac{1}{2}$; otherwise 0.
Linear	$1 - \lvert x\rvert$ if $0 \leq \lvert x\rvert < 1$; otherwise 0.
Quadratic	$\begin{cases} -2\lvert x\rvert^2 + \frac{1}{4} & \text{if } 0 \leq \lvert x\rvert < \frac{1}{2}, \\ \lvert x\rvert^2 - \frac{5}{2}\lvert x\rvert + \frac{3}{8} & \text{if } \frac{1}{2} \leq \lvert x\rvert < \frac{3}{2}, \\ 0 & \text{otherwise.} \end{cases}$
Cubic ($N = 4$)	$\begin{cases} (a+2)\lvert x\rvert^3 - (a+3)\lvert x\rvert^2 + 1 & \text{if } 0 \leq \lvert x\rvert < 1, \\ a\lvert x\rvert^3 - 5a\lvert x\rvert^2 + 8a\lvert x\rvert - 4a & \text{if } 1 \leq \lvert x\rvert < 2, \\ 0 & \text{otherwise,} \end{cases}$
	where a can take the values $a = -1/2, -2/3, -3/4, -1, -4/3$.

The above formulas in the table assume x and y are given in units of the sampling interval.

Figure 7: Kernel function for Image Interpolation. Source: Mitchell 2012

Temporal Alignment

The alignment of data to a common time axis is crucial in an application that involves multi-sensor data fusion application for creating a common representation format. Originally developed for speech recognition, Dynamic Time Warping is a common technique to perform temporal alignment. The distance between two time series is minimized in such a way that the corresponding sensor observations appear at the same location on a common time axis.

Radiometric Alignment

Radiometric alignment or normalization is performed using statistical analysis. Precise parametrization is desirable when a number of fusion techniques are applied. Figure 9 represents the complexity of scene-based radiometric alignment analysis per-pixel.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

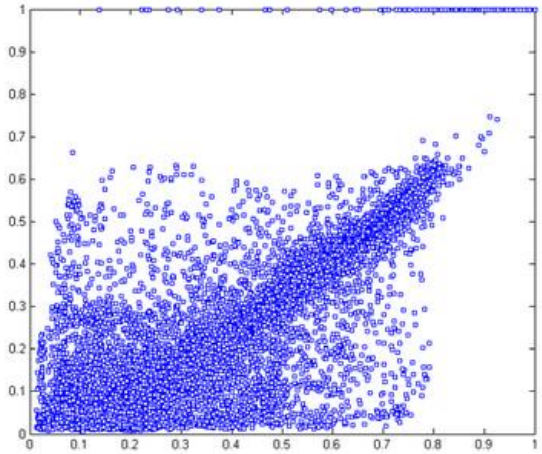

Figure 8: Per-pixel radiometric relations between two normalized snapshot of the same area at different exposures.

Semantic Alignment

This conversion may be required when the same object or phenomena is referred differently in each dataset which does not allow the fusion of the data. Techniques used to perform semantic alignment are the same used to perform a radiometric alignment. Co-association matrices techniques can be used to find special similarities among the input data. Semantic alignment can be performed in Matlab using the CLUSTERPACK toolbox for cluster ensemble algorithm. Representation of matching of two closed contours is shown in Figure 9.

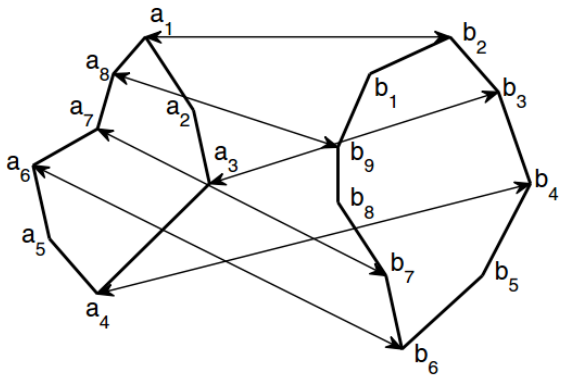

Figure 9: Matching of two closed contours. Source: Mitchell, 2012

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

3.2 High resolution models

Part of the above described data is used to keep running of high resolution models capable to provide more accurate information in the areas with more relevance for the users' activity. These models, in practice, act as complex algorithms in the sense that they pick up a number of data sets and produce higher resolution data.

Presently these models are simulating different processes such as waves, hydrodynamics and water quality. In all cases, Copernicus global or regional models data is being used to define the boundary and initial conditions and to keep a continuous validation process of these high resolution models. Models such as Delft3D and MOHID are being used to simulate hydrodynamics, D-Water Quality (DELWAQ) is being used to simulate water quality, MOHID OSS is being used to simulate the transport dispersion and fate of water pollutants (oil spills, plastics, etc.), Wave Watch III and SWAN are being used to simulate waves. A few more detailed information on the equation solved by D FLOW-FM can be found in D6.2: Report on the sustainable supply chain.

At the current stage of the project, the model is used to compute the sea water level (SWL) and the sea water temperature (SWT). SWT is one of the most important water variables affecting fish health and used as a proxy for estimating fish growth. The temperature modelled by means of D-FLOW FM is a function of atmospheric forces such as solar radiation influx, wind speed, evaporation, etc. The SWT in time series is computed for two aquaculture sites, Ortholithi and Ovrios. The vertical temperature gradient is also performed. In situ measurements are used to calibrate and validate the models, more information on the validation process can be found in D6.3: Service test and validation.

The following sections describe pre-processing operations carried out on CMEMS and EMODnet data used to set up the boundary condition of the model.

Boundary condition

The choice of boundaries depends on the phenomena to be studied and in the study area. For a coastal model, it is usually better to prescribe two different types of boundaries: water level for the alongshore open boundary and currents (Neumann/Riemann could be another possibility) for the cross shore. Each boundary is divided into a set of segments, each with a start node and an end node. In a compromise between required resolution and computational cost, the size of each segment is defined.

Copernicus Marine Environment Monitoring Service (CMEMS) is used to set up the boundary conditions. CMEMS provides free access to a range of satellite data, models, forecasts and reports. However, CMEMS data have a coarse resolution and thus downscaling techniques to extract boundary conditions needed

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

for the DELFT3D-FM are needed. Figure 8 describes the workflow required to define the boundary conditions from CMEMS data. Data are retrieved as NETCDF file. Interpolation techniques are used to downscale the data to a predefined grid. The **nearest neighbor interpolation method** is chosen because the resolution of data holding grid and the assigned at the model boundaries are almost the same. The interpolation is carried out using Python algorithm. Each boundary condition, naming water level, salinity, temperature, northward and eastward velocities, is separately processed.

Recently, a script to automatic download Copernicus data and convert them in ready to use Delft3D-FM boundary condition has been developed and made available for future flow models set-up that requires deep-sea data, instant data acquisition and interpolation, adjusted to the desired boundary condition point location.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

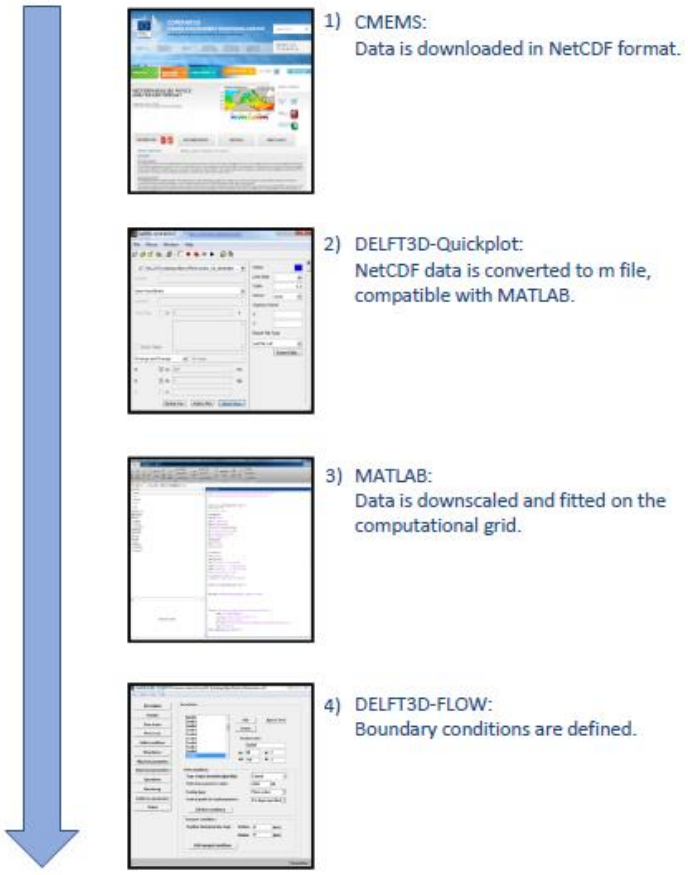

Figure 10: Boundary condition definition, from NETCDF to Delft3d-FM format

Bathymetry

Bathymetry should be defined on a parallel grid using a depth file (*.dep file format). To do so, first bathymetric data should be acquired and then interpolated to fit into the grid. Data are downloaded from the digital terrain model (DTM) developed by EMODnet. Data are visualized and if necessary cleaned-up in Qgis. By visualizing the data, it is indeed possible to recognize if there are any 'obstacles' that could cause local flow irregularities such as vortexes and artefacts which do not represent any real entity.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

Meteorological dataset

The most complicated transformation and extraction processes take place for the meteorological data set. Variables need to be extracted from the NETCDF format and converted to text file format (ASCII). Within the text file format used as model input, the variables are to be given in certain units. Also, the header required formatting and naming according to certain guidelines, including a given title for each quantity. The text file for each variable is equipped with a predefined extension (see Figure 11) by means of specific Matlab codes.

Parameter	Abbreviation	Extension	Quantity	Unit
Near Surface Air Temperature	tas	*.amt	air_temperature	Celsius
Eastward Near Surface Wind	uas	*.amu	x_wind	m/s
Northward Near-Surface Wind	vas	*.amv	y_wind	m/s
Surface Pressure	ps	*.amp	air_pressure	Pa
Near-Surface Relative Humidity	hurs	*.amr	relative_humidity	%
Total Cloud Cover	clt	*.amc	cloudiness	%

Figure 12: extension requirement for meteorological input in Delft3D-FM.

Deltares aims to further develop the model by D-Water Quality (DELWAQ). DELWAQ is a mathematical water quality model that delivers an approximate quantitative description of one or more water quality "state variables" in a given water system. In D-Water Quality models the"state variables" are "substances" that represent a pollutant, a naturally present substance or an aquatic organism. Substances may enter the modelled area through model boundaries, lateral inflows or "dry" waste loads. Unless attached to or located in the sediment bed, they move with currents and turbulence through the modelled water body (model domain). Water flow and dispersion coefficients are usually derived from a hydrodynamic model (e.g. Delft3D-FLOW). The rates of all water quality processes are formulated according to simplified or advanced kinetic rules, each requiring input parameters, often indicated as process coefficients. The rates are dependent on temperature, and sometimes other meteorological parameters such as solar radiation and wind speed. The input for these parameters into a model is indicated with meteorological forcing. Numerical solvers are applied to calculate the concentrations of substances as resulting from mass transport and water quality processes. Water quality modelling can be applied to a wide range of water quality problems, each of those requires the modelling of a specific substance or group of substances. In the contest of the HiSea project, DELWAQ will be employed to compute and predict eutrophication

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

problems by modelling algae biomass, or inorganic nutrients (NH4, NO3, PO4, Si), predict growth and mortality of algae and define a threshold to be used to set up an early warning system.

Mohid Oil Spill Simulator is a fast oil and inert spill lagrangian simulator integrating offline met-ocean forecasts from several different institutions.

The web interface allows end-user to have control over model simulations. Parameters such as date and time of the event, location and oil spill volume are provided to the users; this interactive tool integrates the best available met-ocean forecasts (waves, meteorological, hydrodynamics) from different institutions.

Metocean data are supplied by HiSea grid service, which automatically interpolated and pre-processes the data so it's available for simulations.

Simulations are provided to end-users in a matter of seconds, and thus, can be very useful in emergencies. The backtracking modeling feature and the possibility of importing spill locations from remote servers with observed data (flight surveillance or remote sensing) allow the potential application to the evaluation of possible contamination sources.

The numerical model used to simulate spill fate & behavior in this application is the lagrangian component of the MOHID water modeling system, including the oil spill module (www.mohid.com). The MOHID oil spill module solves the following processes: beaching, evaporation, dispersion, entrainment, sedimentation, dissolution, emulsification and dispersion.

3.3 Post-Processing

Post-Processing methods are used to evaluate the data gathered by the data sources as well as as the outputs of the various predictive models and algorithms that are employed to extrapolate predictive data.

In the post-processing phase, data analysis helps to evaluate how well the data satisfies the assumptions of specific statistical analysis, and often give preliminary indications of trends and set the stage for further trend analysis. Visual inspection of spatial data is recommended before proceeding with more detailed analysis techniques. Spatial pattern analysis can be used to describe the distribution of the value of certain parameter across the area of interest.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

3.3.1 δ-MAPS method

Given a wide variety of spatio-temporal data from single or multiple sources, traditional methods such as the principal component analysis (PCA) or independent component analysis (ICA) can be deployed to identify relationships between variables. Among the available prediction tools that can be used to identify spatially contiguous and possibly overlapping components is the δ-MAPS method. δ-MAPS is an inference method that identifies spatial components and their connection. δ-MAPS can aid in the modelling of the non-linear relationships between the variables and the provision of probabilistic relationships among random variables.

δ-MAPS differs from the commonly used PCA and ICA methods because it does not require the number of domains as an input parameter, the resulting domains are spatially contiguous and potentially overlapping, and the inferred connections between domains can be lagged and positively or negatively weighted. Furthermore, δ-MAPS can be applied towards quantifying differences across datasets and models, evaluating model performances, and investigating model biases and their propagation across different fields of the climate system.

Methodology: The input data is generated from a spatial field sampled on a grid. By means of Pearson's cross-correlation, the similarity between the activities of two contiguous cells is measured. Given homogeneity threshold δ, which can simply be a user-specified parameter for the minimum required average cross-correlation within a domain, a number of N domains are identified. Domains may be spatially overlapping, merged or expanded. Once the domains are defined, a functional network to identify relationships between the domain is constructed.

3.3.2 Density, cluster and neighbourhood influence

Cluster analysis plays a crucial role in building spatial models and data distribution over a studied area. To assess if the sampling locations are well distributed over the study area (or clustered), density analysis can be carried out. Density in the spatial analysis context is defined as the number of points per unit area and a common measure of Complete Spatial Randomness (CSR) for spatial point process. Commonly used methods to check for CSR are the quadrat count and nearest neighbour analysis. Quadrat count is a method to analyze the spatial arrangement of a point. It examines the frequency of points occurring in the quadrats cells that are superimposed on a study area. The limitation of this method is that it is highly dependent on the size of the quadrat.

Nearest Neighbor finds the nearest neighbour of each point. The Average Nearest Neighbour examines the CSR by measuring the average distance between each point. The distribution of the data is considered

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

clustered if the distance is less than the average for a hypothetical random distribution, otherwise, it is dispersed.

These methods are dependent on the computational window, which is an arbitrary area. The inferences made are only applicable to the area within the computational area and are inappropriate to be generalized to other locations.

3.3.3 Serial correlation

Serial correlation or autocorrelation is defined as the correlation between points separated by different time lags and indicatives of how past and future data are related in a time series. It measures the tendency of the time series values to 'remember' its preceding values. It is one of the most widely used methods and most useful descriptive tool in time series analysis to detect non-randomness (in other terms: dependence or non-persistence) in data. An example of the use of correlation analysis is shown in Figure 13.

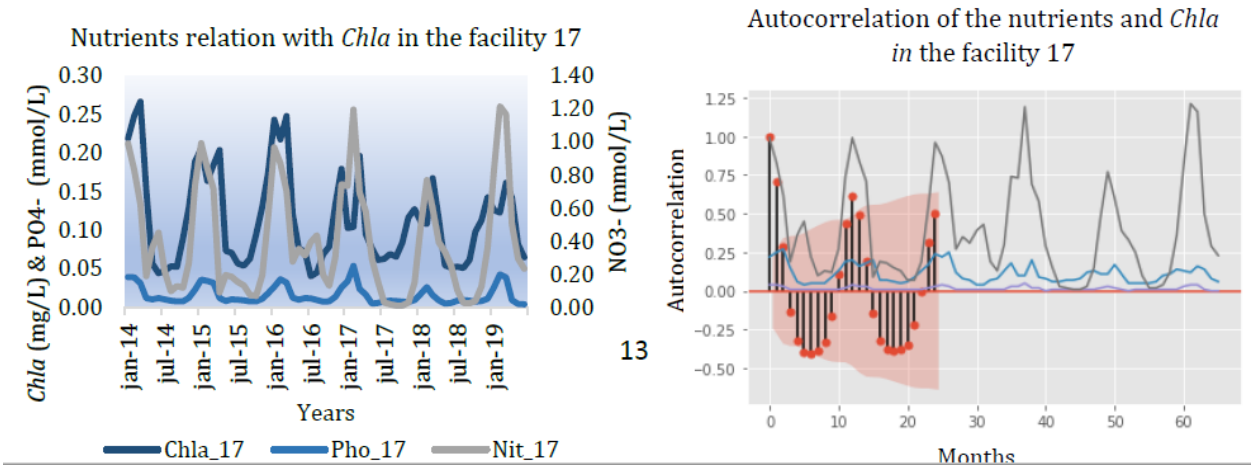

Figure 14: Correlation of Chlorophyll-a, Phosporus and Nitrate concentration in a Greek acquaculture farm

3.3.4 Ensemble forecasting

Ensemble forecasting is a method that has been used in HiSea. Instead of making a single forecast of the most likely atmospheric forcing fields, a set (or ensemble) of forecasts is produced. For our case ensemble of 10 members has been used. This set of forecasts aims to indicate the range of possible future states of the atmospheric forcing fields. Ensemble forecasting is a form of Monte Carlo analysis. The multiple simulations are conducted to account for the two usual sources of uncertainty in forecast models: (1) the errors introduced by the use of imperfect initial conditions, and (2) errors introduced because of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

imperfections in the model formulation, such as the approximate mathematical methods to solve the equations.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934