EMODnet Thematic Lot n°2 – Seabed Habitats

CINEA/EMFAF/2022/3.5.2/SI2.900370

Start date of the project: 25/09/2023 - (24 months)

Phase V

Constructing EUSeaMap

User Guide

Version 3

The European Marine Observation and Data Network (EMODnet) is financed by the European Union under Regulation (EU) No 508/2014 of the European Parliament and of the Council of 15 May 2014 on the European Maritime and Fisheries Fund.

The European Marine Observation and Data Network (EMODnet) is financed by the European Union under Regulation (EU) No 508/2014 of the European Parliament and of the Council of 15 May 2014 on the European Maritime and Fisheries Fund.

Disclaimer

The information and views set out in this guide are those of the author(s) and do not necessarily reflect the official opinion of the CINEA or of the European Commission. Neither the CINEA, nor the European Commission, guarantee the accuracy of the data included in this study. Neither the CINEA, the European Commission nor any person acting on the CINEA's or on the European Commission's behalf may be held responsible for the use which may be made of the information.

Document info

Title	Constructing EUSeaMap – User Guide
Deliverable	
Author	Mickaël Vasquez (Ifremer). Mickael.Vasquez@ifremer.fr
Dissemination level	Public
Version date	15/01/2025
Version	3
Version updates	A new script: Cleaning up a habitat descriptor raster 2 new Appendices (3 and 4) More exercises in Appendix 5 (exercises 9, 10 and 11) Minor text corrections

Contents

Executive summary

In the first two phases of EMODnet, the Seabed Habitats thematic project used a commercial software, namely ARCGIS™, to produce the pan-European broad-scale seabed habitat map, EUSeaMap. A key objective of Phase 3 (2017-2021) was to develop a new GIS workflow that is robust, repeatable and transferable across marine regions, and to implement this workflow in tools based on open source technologies. As a result, most of the workflow was implemented in the form of R scripts. For technical reasons, a small part had to be implemented in the form as an ArcGIS™ ModelBuilder model. Since then, these tools have been used to develop the 2019, 2021 and 2023 versions of EUSeaMap.

This technical guide is intended to help the reader learn how to use these tools. The vocabulary and concepts used in the document are defined and the GIS workflow is described. The purpose of the scripts is documented in detail, as well as what they require as input and what they produce as output. Step-by-step hands-on training is provided. However, it is important to note that this document is intended for advanced users, i.e. that a good knowledge of the methods used for EUSeaMap, which are described in other documents, is an essential prerequisite for understanding the document and the associated scripts.

1 Introduction

The aim of this technical guide is to help the reader to use the R scripts and the model created using the ArcGIS™ ModelBuilder application developed by EMODnet Seabed Habitats[1] to produce the broad-scale seabed habitat map (referred to as "EUSeamap" in the guide) in a semi-automated manner.

The general methods used are largely based on a methodology that was originally developed by the MESH Project, the general principles of which are described in Coltman et al. (2008). It is strongly recommended that the reader read this short document before going through this one.

The way in which some concepts such as fuzzy laws or Generalized Linear Models (GLMs) are used in the context of EUSeaMap, a thorough understanding of which is a prerequisite for the understanding of the present document and related scripts, will not be developed here. Further explanation can be found in the EUSeaMap Phase 2 Technical Report Populus et al, 2017. In particular, it is particularly recommended to read Sections 2.5 to 2.7 and appendices 9 (Thresholds), 10 (Confidence) and 12.

Before describing the scripts, i.e. what they do and what their inputs and outputs are, the vocabulary and concepts used in the document are defined and the GIS workflow is described.

2 Vocabulary and concepts

A habitat is a combination of several environmental features. In habitat classifications these environmental features are formulated by what we will refer to in this document as **'habitat descriptors'**, which in the seabed section of the European classification EUNIS are at least seabed substrate types and biological zones, and in some cases other seabed features such as the energy levels induced by water movements. For example, in the Atlantic EUNIS version 2007-11 describes habitats using three habitat descriptors: energy levels, biological zones and seabed substrate types. The habitat descriptors have their own classification, e.g. in the Atlantic the habitat descriptor classification for the energy levels has 3 classes: high, moderate and low. In this document we will refer to these classes as **'habitat descriptor class'**.

The following habitat descriptors are used in EUSeaMap:

- Seabed Substrate types
- Biological Zones
- Wave-induced Energy Levels
- Current-induced Energy Levels
- Oxygen Regimes
- Salinity Levels
- Masks

Seabed substrate types and biological zones are used in all marine regions. A third habitat descriptor is then usually used, depending on the region. For example, in the Atlantic combined wave and current induced energy levels are used as a third habitat descriptor, while in the Black Sea oxygen regimes are used. Further details on which habitat descriptors are used in each region can be found in Populus et al. (2017).

All the habitat descriptor classes are described in Annex 1, together with the numerical code assigned to them in the EUSeaMap GIS workflow that produces the broad-scale habitat map.

A habitat descriptor class typically has two boundaries: an upper and a lower boundary. The upper boundary is the boundary that the class shares with its upper adjacent class in the habitat descriptor classification, and the lower boundary is the boundary that the class shares with its lower adjacent class. For example, the upper

[1] https://emodnet.ec.europa.eu/en/seabed-habitats

boundary of the Circalittoral Biological Zone is the boundary it shares with the Infralittoral, while its lower boundary is the boundary it shares with the Offshore Circalittoral.

Some habitat descriptor classes have only one boundary because they are at the top or the bottom of their classification and therefore have only one adjacent class. For example, the Infralittoral Biological Zone, the Oxic Oxygen Regime, or the High Energy Level only have a lower boundary (i.e. no upper boundary) because these classes are at the top level of the Biological Zone, Oxygen Regime and Energy Level habitat descriptor classifications respectively. Similarly, the Abyssal Biological Zone, Anoxic Oxygen Regime or Low Energy Level have only one boundary, the upper one, because they are at the bottom of their respective classifications.

3 The EUSeaMap GIS Workflow

The general workflow consists of 4 steps:

1. The habitat descriptor layers are created individually (together with their respective confidence layer).
2. These habitat descriptor raster layers are converted into polygon layers.
3. The seabed habitat layer is created by overlaying all polygon habitat descriptor layers. In parallel, the individual habitat descriptor confidence layers are overlaid to produce an overall confidence layer.
4. The seabed habitat layer is then joined to the lookup table that crosswalks the modelled numerical habitat codes and the classes of habitat classifications such as EUNIS or the MSFD broad habitat type classification[2].

In EUSeaMap, some habitat descriptor layers are provided ’out-of-the-box’, i.e. they don’t need to be created by the EUSeaMap workflow because they are created externally. An example of this in all marine regions is the habitat descriptor 'Seabed Substrate', which is created and provided directly by EMODnet Geology.

Figure 1 illustrates the workflow with an example where four habitat descriptors are considered, three of which would be produced by the workflow (HD2, 3 and 4) and one of which (HD1) would be externally created.

Sections 3.1, 3.2, 3.3.1 and 3.3.2 describe each of the three steps above.

[2] COMMISSION DECISION (EU) 2017/848

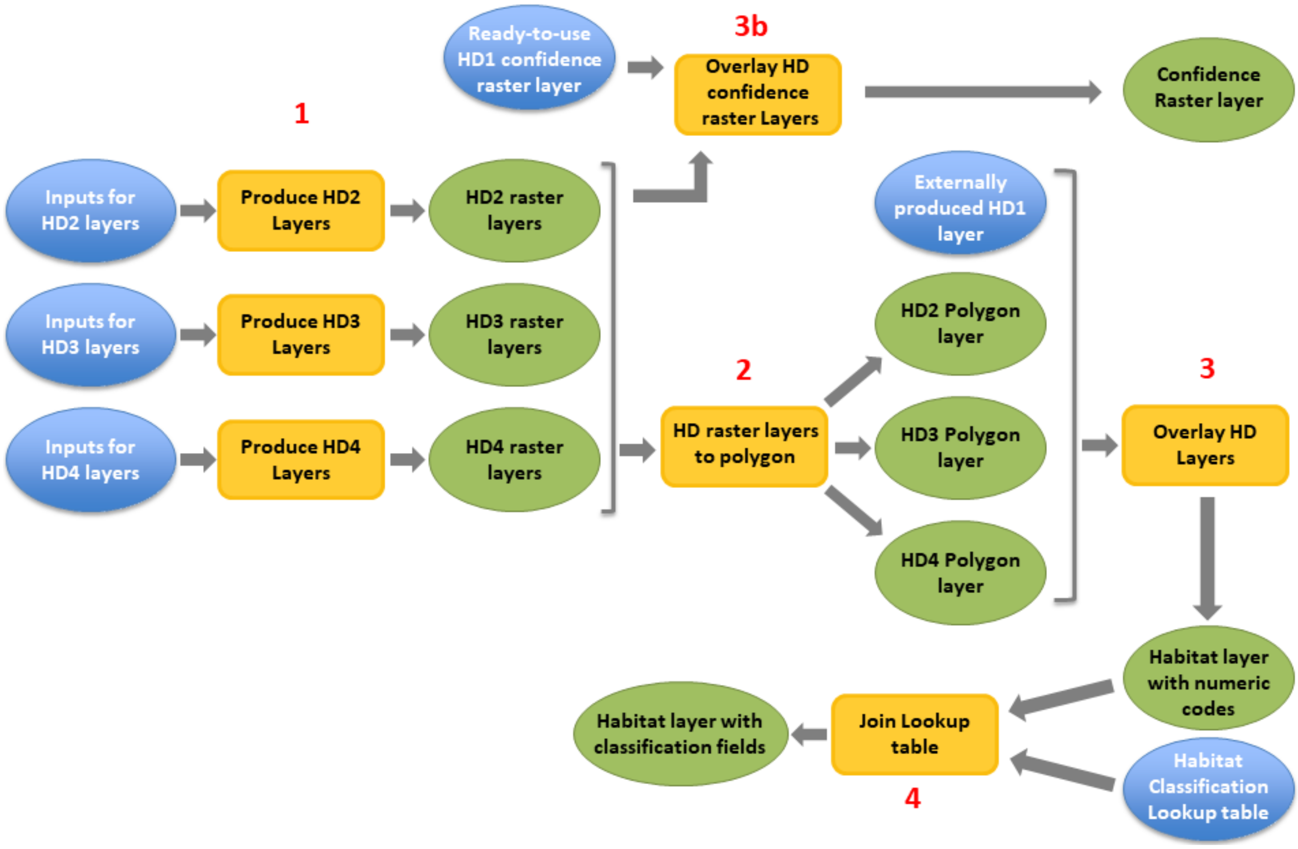


Figure 1: The general EUSeaMap GIS workflow illustrated with an example of 4 habitat descriptors (HD). Layers are produced for HD2, HD3 and HD4, while the HD1 layer is provided 'out-of-the-box' (as it is produced externally). The HD2, HD3 and HD4 layers are 1) produced as rasters (along with a confidence layer for each), 2) converted into polygons, 3) overlaid to produce the habitat layer. 3b: The HD confidence raster layers are also overlaid into a single overall confidence raster. After overlaying the HD layers, the lookup between numeric habitat codes and different classification classes is joined to the habitat layer (4).

3.1 Creation of a habitat descriptor layer and its confidence layer

3.1.1 Workflow

The production of a habitat descriptor layer and its confidence layer consists of 2 steps (Figure 2):

1. The habitat descriptor classes are modelled using spatial distribution laws. As a result, 2 raster layers are produced for each class, namely i) a spatial distribution layer and ii) a layer on the confidence in the spatial distribution.
2. All the individual class spatial distribution and confidence layers modelled in step 1 are merged into a single layer.

Sections 3.1.2 and 3.1.3 describe these two steps in more detail.

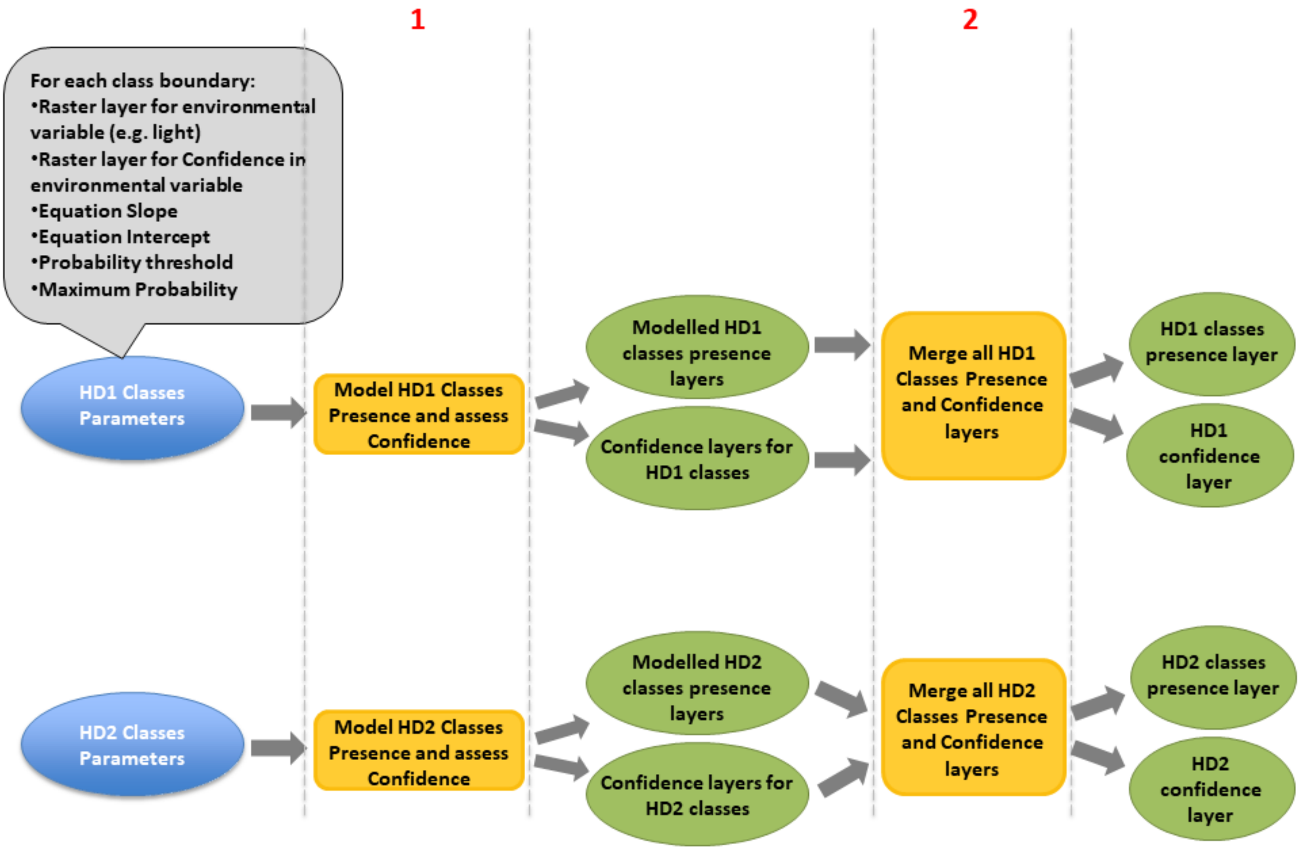


Figure 2: Production of Habitat Descriptor (HD) layers for 2 HDs, HD1 and HD2. For each HD: 1) for each HD class, presence layers are modelled and confidence is assessed (hence two output layers are produced, one for class presence and one for confidence in class presence); 2) individual HD class presence and confidence layers are merged into a single layer.

3.1.2 Step 1: modelling the habitat descriptor classes and assessing their confidence

Modelling the habitat descriptor classes: the use of spatial distribution GLM/fuzzy laws

The presence of each individual habitat descriptor class is modelled using either Generalized Linear Models (GLMs) or fuzzy laws. Details of how these GLMs and fuzzy laws are fitted in the approach developed for EUSeaMap can be found in Populus et al. (2017). In the EUSeaMap approach, GLM/fuzzy laws are used to predict the probability of the presence of a habitat descriptor class, given the value of a unique environmental predictor variable (e.g. the amount of light available at the seabed is the unique predictor variable for the presence of the infralittoral).

Here we describe the set of parameters that need to be provided to step 1 of the above workflow (see Figure 2) in order to compute a probability and presence raster using a GLM/fuzzy law.

Both GLMs (Figure 3A) and fuzzy laws (Figure 3B) allow the calculation of the probability of presence of a habitat descriptor class as a function of a predictor value.

The equation has the form:

$P(X) = e^{ax+b} / (1+e^{ax+b})$ for a GLM,

$P(X) = ax+b$ for a fuzzy law

where X is the predictor value, P(X) is the probability of occurrence of the habitat descriptor class, and a and b are the slope and the intercept of the GLM or the fuzzy function, respectively.

To model the presence of a habitat descriptor class, a GLM/fuzzy law is required for each boundary of the class, i.e. the upper boundary (if any) and the lower boundary (if any).

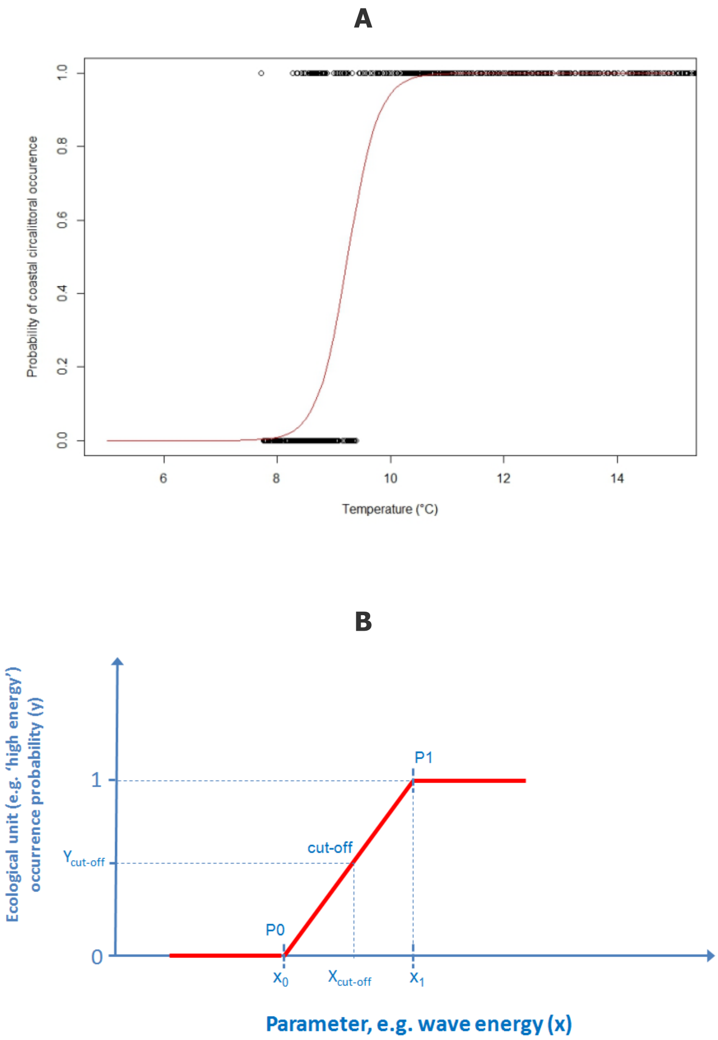


Figure 3: The 2 types of laws used to model the habitat descriptor class spatial distribution: GLM (A) and fuzzy (B) – In Populus et al, 2017

For each GLM/fuzzy law a probability cut-off (also called a threshold) is also provided so that any predicted probability value of presence can be converted to a binary value (present or absent): above the probability threshold, the habitat descriptor class is characterised as present, below it as absent. The approaches used to define these thresholds are described in Populus et al (2017).

In summary, for a GLM/fuzzy law to be used in the workflow, 3 values are required for the 2 boundaries (or the boundary if there is only one) of each habitat descriptor class: a **slope**, an **intercept** and a **probability threshold**.

Note: For background information, Annex 2 describes how to calculate the slope, intercept and probability threshold from the fuzzy law control points.

Special cases for externally-defined boundaries

Manually drawn boundaries are used in specific regions and for specific reasons. For example, in the Black Sea, Mediterranean Sea, Caspian Sea, Caribbean Sea, the choice has been made to define the deep circalittoral / bathyal and the bathyal / abyssal boundaries as slope changes. The line delineating these slope changes is semi-automatically drawn by experts. Another example is in the Baltic Sea, where the infralittoral / circalittoral boundary was defined externally based on national studies.

This type of boundary requires a special integration into EUSeaMap. The way the confidence is calculated and the type of input required is discussed in Appendix 3.

Assessing the confidence in the habitat descriptor classes

As described in Populus et al. (2017), the confidence in assigning the correct habitat descriptor class to a raster cell depends on two items:

1) The confidence **in the environmental layer values** used for each habitat descriptor class boundary: this is an input to the workflow, provided in the form of a raster layer where cell values are a categorical measure of the confidence (1=low, 2=moderate, 3=high). Some guidance on the provision of this layer can be found in Populus et al. (2017), section 2.7. Note that EMODnet Geology provides a confidence index for its polygon Seabed Substrate layer, with confidence assigned at the polygon level. Similarly, EMODnet Bathymetry provides a confidence index for its DTM, with confidence assigned at the raster cell level.

The confidence **based on the probability** predicted by the GLM/fuzzy law. For each raster cell, the workflow classifies the calculated probability (a continuous value in the range [0-1]) into one of the following categorical confidence values: 1 (for low), 2 (for moderate) or 3 (for high). The rules used to classify a probability value into a categorical confidence are described in Table 1.

Table 1: Rules used to classify a continuous probability as high, moderate or low confidence. 'probability max' is the maximum value of the probability considering the GLM/fuzzy law (typically 1, but in some cases may be less); 'range'= probability max - probability threshold

Confidence per cell	Rule
High	probability threshold + (0.6 x range) ≤ probability ≤ probability max
Moderate	probability threshold +(0.2 x range) ≤ probability < probability threshold + (0.6 x range)
Low	probability threshold ≤ probability threshold +(0.2 x range)

For example, for a GLM/fuzzy law with a maximum probability of 1, and a probability threshold of 0.5, the rules would be as described in Table 2.

Table 2: Example of rules used to classify a continuous probability as high, moderate or low confidence. In this example probability max is 1 and probability threshold is 0.5.

Confidence per cell	Rule
High	0.8 ≤ probability ≤ 1
Moderate	0.6 ≤ probability < 0.8
Low	0.5 ≤ probability < 0.6

For each habitat descriptor class boundary, the workflow automatically creates a confidence raster layer based on probability, which is done by classifying the probability as high, moderate or low confidence. For this to be possible, the user needs to provide the maximum probability as input to the script, in addition to the 3 values mentioned above (i.e. slope, intercept and probability threshold).

In a final step, the workflow creates a single confidence raster layer for each habitat descriptor class boundary by combining the confidence raster layer for the environmental variable values and the confidence raster layer based on the probability. The logic used for the combination is described in Table 3.

Table 3: logic used to combine the two types of confidence

		Confidence in values of the environmental variable		
		H	**M**	**L**
Confidence based on the probability	**H**	H	H	M
	M	M	M	L
	L	L	L	L

GIS Workfkow

The workflow for this step (figure 4) requires the following input parameters **for each class boundary**:

- The raster layer of the environmental predictor variable (e.g. seabed PAR)
- The slope of the GLM/fuzzy law
- The intercept of the GLM/fuzzy law
- The probability threshold
- The maximum probability

It produces for each habitat descriptor class:

- A categorical raster layer for the occurrence of the class. In this layer all the cells have the same value, which is assigned according to the habitat descriptor class coding convention in Annex 1.
- A categorical raster layer for confidence in the occurrence of the habitat descriptor class. The values of the cell are one of the following: 1 (low confidence), 2 (moderate confidence) and 3 (high confidence)

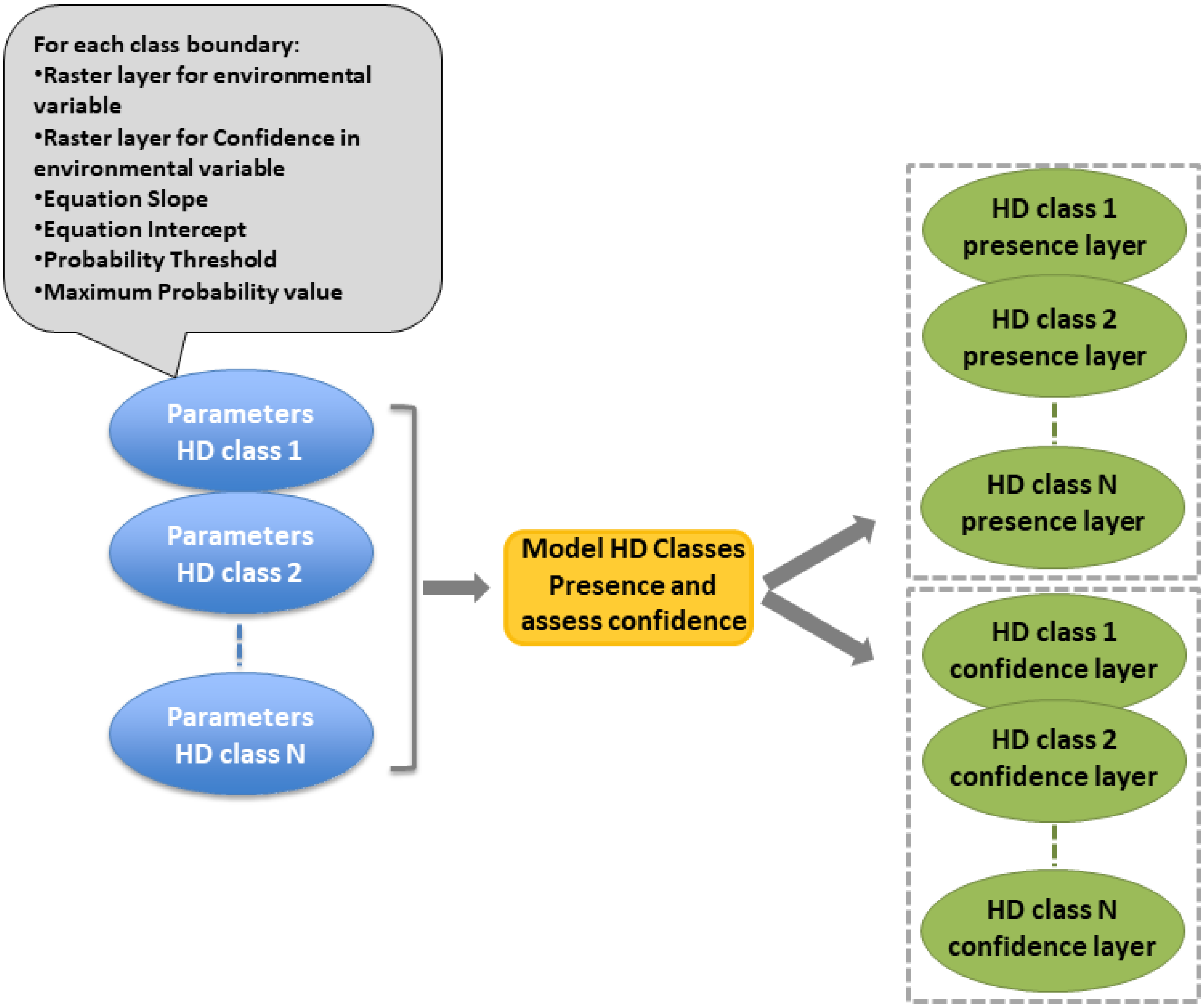


Figure 4: Workflow for the production of the individual class presence and confidence layers for a Habitat Descriptor (HD). For each class boundary, the inputs are the environmental variable raster and its confidence raster, the slope and intercept of the GLM/fuzzy law, the probability threshold and the maximum value that the probability can take considering the GLM/fuzzy law. For each class, 2 raster layers are created: a presence layer and a confidence layer.

An option: providing the slope, intercept, probability threshold and probability maximum value as a raster instead of a constant value

Within the same region, the GLM/fuzzy law slope and intercept, the probability threshold and the probability maximum value defining a habitat descriptor class boundary may vary spatially.

An example of this is shown in Figure 5. In the Black Sea, the slope and intercept values that are used for the infralittoral lower boundary are not the same whether the seabed substrate is hard or soft. In this particular case, where the GLM/fuzzy law parameters vary spatially, instead of a constant value for slope and intercept, the workflow requires a raster layer as input for both. These input rasters are constructed in a GIS software using a seabed substrate raster dataset.

See appendix 4, section 9, for a concrete application.

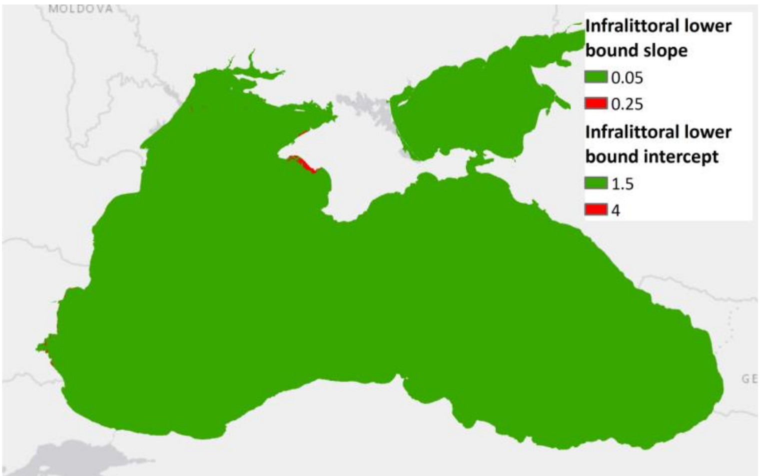


Figure 5: Example of spatially varying GLM/fuzzy slope and intercept for a habitat descriptor class boundary: the infralittoral lower boundary in the Black Sea.

3.1.3 Step 2: merging the habitat descriptor classes in a single layer

As shown in Figure 2, step 2 of the workflow merges all the habitat descriptor class layers and the confidence layers produced in step 1. The output is a single habitat descriptor raster layer containing all the classes (e.g. all biological zones) and a single habitat descriptor confidence raster layer.

3.2 Converting the habitat descriptor raster layers in polygon layers

The outputs of step 1 of the general workflow (see Figure 1) are habitat descriptor raster layers. As the creation of the habitat layer (Figure 1, step 3) involves the overlay of the habitat descriptors in vector mode, all the habitat descriptor layers need to be converted to polygon layers (Figure 1, step 2).

3.3 Creating the habitat map

The final step in the general workflow is the creation of the habitat map. In a first step the habitat descriptor layers are all combined by a geometric intersection to create the habitat layer (see Figure 1, step 3). In a second step, to the attribute table of this habitat layer is joined the lookup table that crosswalks the habitat codes resulting from the union and the classes of habitat classifications such as EUNIS or the MSFD Broad Benthic Habitat Types (see Figure 1, number 4). These two steps are described below.

3.3.1 Combining the habitat descriptor layers

General principle

All the habitat descriptor layers are combined by a geometric intersection. This results in a habitat polygon layer, the attribute table of which contains one column per habitat descriptor layer. For example, in the Mediterranean, where 3 habitat descriptors are considered, namely seabed substrate types, biological zones, and mask values, the habitat layer will contain 3 columns.

For each polygon, the values contained in these columns are summed in a new column. This results in a code that is representative of a unique combination of the habitat descriptor classes considered in the region (e.g., in the Mediterranean, representative of a seabed substrate type, a biological zone and a mask value).

Coding convention

For consistency across regions, it was agreed that all the habitat descriptor classes would be coded to two digits except, with the exception of the mask habitat descriptor classes, which are coded to one digit due to technical limitations. It was also agreed that, as shown in Figure 6, the two digits from the right would be devoted to the seabed substrate and then, from right to left, the biological zone 2 digits, the energy 2 digits, the (temperature, oxygen or salinity) 2 digits, and the mask digit.

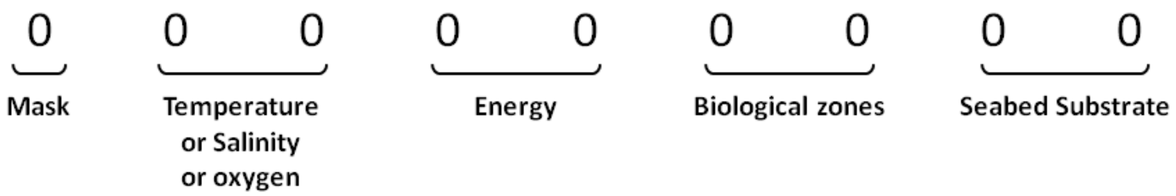


Figure 6: Coding convention for modelled habitat codes

Thus, in order for the above sum of the habitat descriptor codes to produce such coded values, each habitat descriptor code must be multiplied by the factors given in Table 4.

Table 4: Multiplication factors used for each habitat descriptor

Habitat descriptor name	Multiplication factor
Seabed substrate	1
Biological zone zones	100
Energy	10000
Oxygen, Salinity, Temperature	1000000
Mask	100000000

A few examples of codes

In the Mediterranean Sea, 200001030 means seabed substrate=30 (coarse sediment), biological zone = 10 (i.e. infralittoral) and mask = 2 (i.e. mask 1).

In the Black Sea, 100002020 means seabed substrate=20 (sand), biological zone = 20 (i.e. circalittoral) and mask = 1 (i.e. no mask).

In the Atlantic 103020 means seabed substrate=20 (sand), biological zone = 30 (i.e. deep circalittoral) and energy = 10 (i.e. low energy)

3.3.2 Joining the look-up table

The final step is to join the habitat layer and the look-up table, i.e. the table that crosswalks the digital habitat codes mentioned above and the habitat classes described in several habitat classifications (e.g. EUNIS). The format of the table that is used for EUSeaMap 2023 is in appendix 4.

modelCod	Biozone	Substrat	EUNISc	EUNISd
100001000	Infralittoral	Unknown	Na	Na
100001010	Infralittoral	Mud	A5.34	Infralittoral fine mud
100001020	Infralittoral	Sand	A5.23	Infralittoral fine sands
100001030	Infralittoral	Coarse & mixed sediment	A5.13	Infralittoral coarse sediment
100001040	Infralittoral	Coarse & mixed sediment	A5.13	Infralittoral coarse sediment
100001050	Infralittoral	Sandy Mud	A5.33	Infralittoral sandy mud
100001060	Infralittoral	Muddy Sand	A5.23	Infralittoral fine sands
100001070	Infralittoral	Rock or other hard substrata	A3	Infralittoral rock and other hard substrata
100001071	Infralittoral	Posidonia oceanica	A5.535	[Posidonia] beds
100001072	Infralittoral	Cymodocea nodosa	A5.531	[Cymodocea] beds
100002000	Circalittoral	Unknown	Na	Na
100002010	Circalittoral	Mud	A5.39	Mediterranean biocoenosis of coastal terrigenous muds
100002020	Circalittoral	Sand	A5.46	Mediterranean biocoenosis of coastal detritic bottoms
100002030	Circalittoral	Coarse & mixed sediment	A5.46	Mediterranean biocoenosis of coastal detritic bottoms
100002040	Circalittoral	Coarse & mixed sediment	A5.46	Mediterranean biocoenosis of coastal detritic bottoms
100002050	Circalittoral	Sandy Mud	A5.38	Mediterranean biocoenosis of muddy detritic bottoms
100002060	Circalittoral	Muddy Sand	A5.46	Mediterranean biocoenosis of coastal detritic bottoms
100002070	Circalittoral	Rock or other hard substrata	A4.26 or A4.32	Mediterranean coralligenous communities moderately exposed to or sheltered from hydrodynamic action
100002071	Infralittoral	Posidonia oceanica	A5.535	[Posidonia] beds
100002072	Infralittoral	Cymodocea nodosa	A5.531	[Cymodocea] beds

Figure 7: Example of look-up table. In the column "modelCod" is the modelled habitat code; the columns EUNISc and EUNISd are the EUNIS code and description that match the code

4 The R scripts & ArcGIS™ model

The R scripts and the ArcGIS™ toolbox are available on Github:

https://github.com/emodnetseabedhabitats/EUSeaMap_creation

Two R scripts are proposed for the two steps that allow the production of a habitat descriptor layer and its confidence layer (section 3.1). For the other tasks in the general workflow, i.e. converting the habitat descriptor raster layers into polygon layers (section 3.2) and combining them to produce the habitat layer (section 3.3.1), and joining the habitat layer and the look-up table (section 3.3.2), we recommend the use of ARCGIS™. The various R package functions tested for raster-to-polygon conversion did not prove sufficiently efficient for large numbers of polygons (which is the case for EUseaMap). Other tests of ad hoc R packages for intersecting polygon layers also showed some limitations (for more details see Vasquez et al. 2020, annex 4). A ARCGIS™ ModelBuilder model is proposed that allows raster-to-polygon conversion and polygon intersection to be performed together. An ARCGIS™ ModelBuilder model is also proposed for joining the habitat layer and the look-up table. Due to the large number of polygons, it is recommended that the outputs of the ARCGISTM ModelBuilder models be produced in geodatabase format rather than shapefile format.

The scripts and the ARCGIS™ ModelBuilder models are documented in section 4.3 (Script for the production of habitat descriptor layers) and 4.4 (Creating the habitat layer and overall confidence). Figure 8 shows the general workflow with references in brakets to the subsection describing the different tasks.

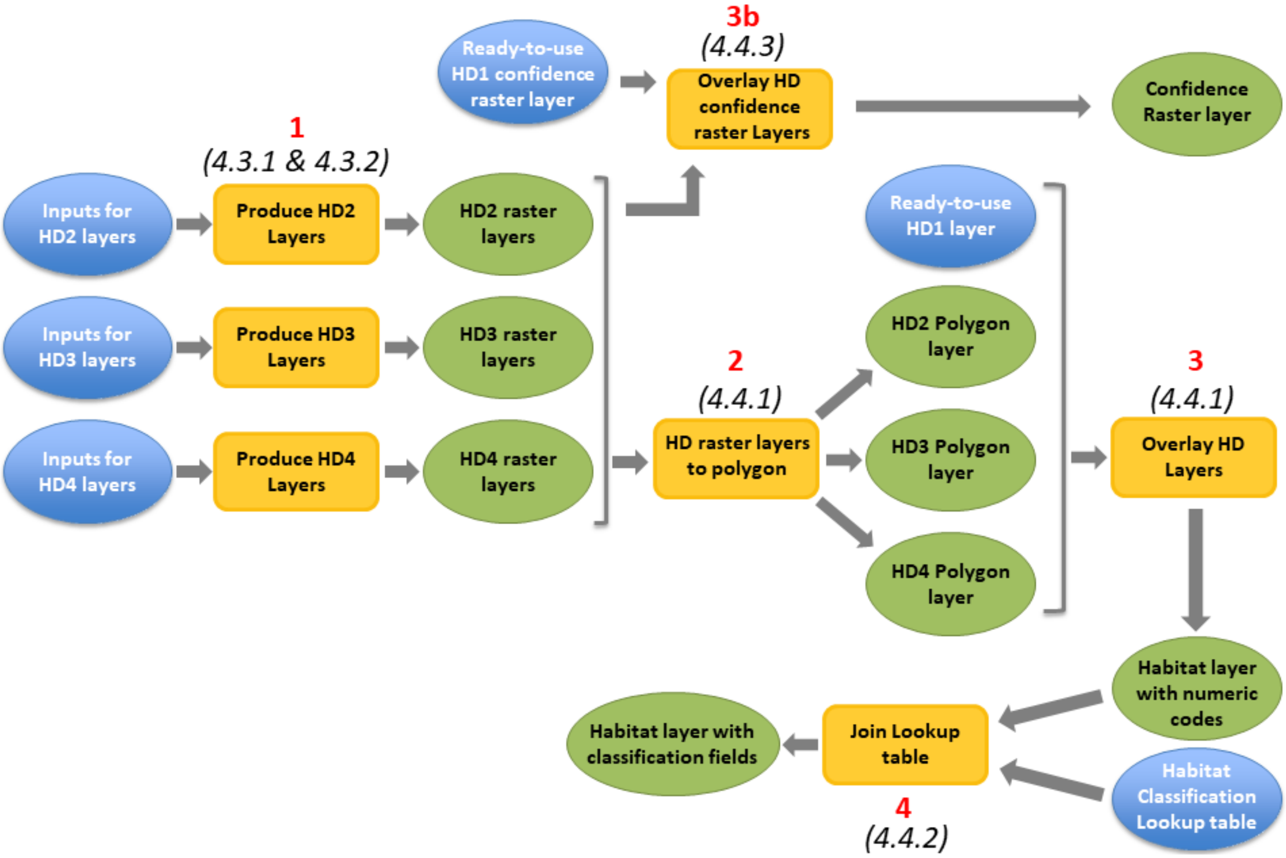


Figure 8: General workflow and subsection (in brackets) that documents the script or ModelBuilder model corresponding to the various tasks

4.1 Prerequisites

4.1.1 System requirements

RAM/Memory: the R package *terra* is memory hungry. Minimum requirement is 16 GB, ideal 32 GB.

4.1.2 Software Requirements

R

The scripts are written in the R language. Therefore, R has to be installed.

R packages

The R package terra is required

ArcGIS™

ARCGIS™ ModelBuilder models require ArcGIS™ 10.0 or higher. No extension to native ArcGIS™ is required.

4.2 General principles

4.2.1 A critical requirement: raster inputs must all be spatially consistent

Since most of the scripts perform raster overlay operations, a key requirement is that all input rasters must be fully spatially consistent, i.e. strictly the same spatial extent and resolution. If this requirement is not met, the script will fail and return an error message such as "different extent" or "different resolution".

4.2.2 Formulation of the R script inputs via configuration files

The R scripts were designed to be run by non-R experts, i.e. as a "black box". Most of the inputs passed to the scripts are formulated via configuration files in csv (comma separated values) format (Figure 9). In general, very few parameters, such as the working directory, need to be formulated directly in the script (Figure 10).

shortName	longName	code	upper_boundary_equation_type	upper_boundary_variable	upper_boundary_slope	upper_boundary_intercept
infra	Infralittoral	10				
shallowCirca	Shallow circalittoral	20	1	SeabedLightEnergy.tif	-0.926	2.102
deepCirca	Deep circalittoral	30	1	SeabedLightPercentage.tif	-2000	1.5

Figure 9: Screenshot of a csv configuration file in which script inputs are formulated

```
#------------------------- script parametrisation --------------------------------------------------
#chunk size and max memory to load for processings
chunksize<-1e+07
maxmemory<-1e+09

#working directory
workingDirectory<-"D:/travail/euseamap_phase3/WP2_modelling_thresholds_confidence/Modeling/r_draft/medsea_small_with_new_codif"

#the config csv file that describes all the bits that have to be merged
config_csvFileName<-"medsea_habitat_descriptor_merging_inputs.csv"

#the name of the habitat descriptor as it appears in the config csv fine, column "habitat_descriptor_shortName"
habitat_descriptor_shortName<-"biozones"
#----------------------------------------------------------------------------------------------------
```

Figure 10: Example of parameters that are formulated in a script. In this example, there are 5 parameters (workingDirectory, config_csvFileName, habitat_descriptor_shortName)

In the script description sheets the inputs passed via the configuration file are referred to as 'Inputs' and the parameters formulated in the script are referred to as 'script parameters'.

4.2.3 A strict folder structure

For a given study area (e.g. Mediterranean Sea), the user must follow a standard folder structure (Figure 11) in order to run the script.

The root folder is the **working directory**. Below this are the following folders:

- **config_files**: folder where the script expects the csv configuration files to be stored.
- **input**: folder where the script expects the input files to be stored
- **output**: folder where the script writes the output files

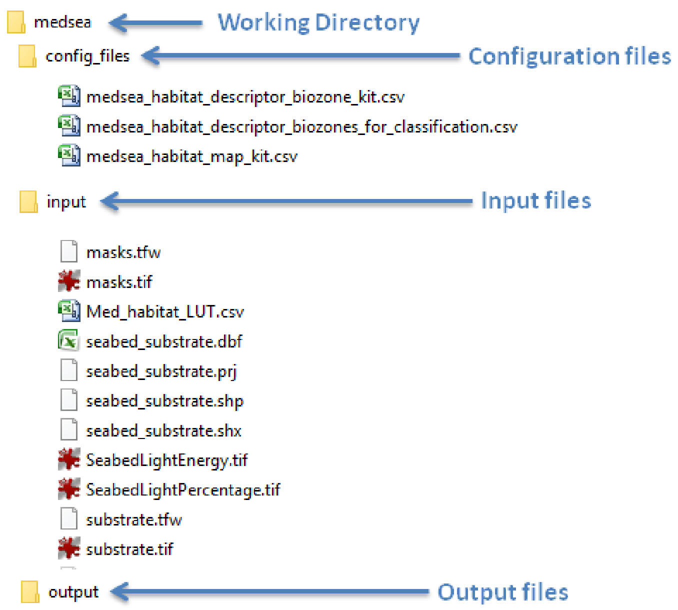


Figure 11: The folder structure

4.2.4 Changing the maximum amount of memory used by terra

The "terra" R package uses a lot of memory as it performs all calculations in memory. All the scripts that use the terra package have a script parameter, 'maxmemory', which specifies the maximum amount of RAM (in GB) that terra is allowed to use when processing a raster dataset. Depending on the amount of RAM available on your computer, you may wish to increase this value.

4.3 Scripts for the production of habitat descriptor layers

4.3.1 Modelling the habitat descriptor classes

<table>
<tr><td>Script name</td><td>habitat_descriptor_modelling.R</td></tr>
<tr><td>Author</td><td>Mickaël Vasquez (Ifremer)</td></tr>
<tr><td>Date</td><td>26/05/2023</td></tr>
<tr><td>Descr.</td><td>Produces rasters of the spatial distribution of the classes of a habitat descriptor according to the modelling law parameters that characterise the class boundaries. The script produces a spatial distribution raster for each class whose input parameters are described in the csv configuration file. This configuration file specifies for each class boundary the predictor (i.e. the environmental variable) raster, the equation type (GLM or fuzzy), the slope and intercept of the equation, the probability threshold and the maximum probability of the law. See section 3.1 for more details on these parameters.

For each class, the script will produce a spatial distribution raster, a confidence raster, a confidence raster based on probability and, optionally, a probability raster.

If it's a GLM, the equation used for the probability raster calculation is

$P(\text{predictor}) = e^{(\text{slope x predictor + intercept})} / (1+e^{\text{slope x predictor + intercept}})$

If it's a fuzzy law, it is

P(predictor) = slope x predictor + intercept</td></tr>
<tr><td>Workflow</td><td>Figure 2, (see number 1) and Figure 4</td></tr>
<tr><td>Inputs</td><td>All inputs are described in a csv configuration file. The csv file contains one row for each habitat descriptor class. If the habitat descriptor class has only one boundary (see section 2 for further explanation of boundaries), the csv file cells for the missing boundary parameters are left blank.

Important: all input raster files must be located in the input folder

<table>
<tr><th>Column Name</th><th>Type*</th><th>Description</th></tr>
<tr><td>shortName</td><td>C</td><td>The habitat descriptor class short name. Crucial. it will be used in the name of the output rasters. Must not contain special characters</td></tr>
<tr><td>longName</td><td>C</td><td>The habitat descriptor class short name. Only used in the messages displayed by the script</td></tr>
<tr><td>code</td><td>N</td><td>The habitat descriptor class code that will be given to the output class raster.</td></tr>
</table>
</td></tr>
</table>

			Must follow the conventions (see annex 1)
	upper_boundary_equation_type	N Or C	For the class upper boundary, Integer value that indicates the type of modelling law 1 = fuzzy 2 = GLM 3 = fuzzy for externally defined boundary If the type of modelling law varies spatially across the study area, a raster name must be provided instead of an integer value. The raster cells will indicate, via the same values as above, if the type is fuzzy or GLM
	upper_boundary_variable	C	For the class upper boundary, the raster file name for the environmental variable
	upper_boundary_confidence_in_variable	C	For the class upper boundary, the raster file name for the confidence in the environmental variable
	upper_boundary_slope	N Or C	For the class upper boundary, the slope of the GLM/fuzzy law. If the slope varies spatially across the study area, a raster name must be provided instead of a numeric value. The raster cells will indicate the value of the slope
	upper_boundary_intercept	N Or C	For the class upper boundary, the intercept of the GLM/fuzzy law. If the intercept varies spatially across the study area, a raster name must be provided instead of a numeric value. The raster cells will indicate the value of the intercept
	upper_boundary_threshold	N Or C	For the class upper boundary, the probability threshold. If the probability threshold varies spatially across the study area, a raster name must be provided instead of a numeric value. The raster cells will indicate the value of the threshold
	upper_boundary_max_probability	N Or	For the class upper boundary, the maximum value that the probability can have (in some cases may be less than 1)

		C	If the maximum probability varies spatially across the study area, a raster name must be provided instead of a numeric value. The raster cells will indicate the value of the maximum probability
	lower_boundary_equation_type	N Or C	For the class lower boundary, Integer value that indicates the type of modelling law 1 = fuzzy 2 = GLM 3 = fuzzy for externally defined boundary If the type of modelling law varies spatially across the study area, a raster name must be provided instead of an integer value. The raster cells will indicate, via the same values as above, if the type is fuzzy or GLM
	lower_boundary_variable	C	For the class lower boundary, driver raster name
	lower_boundary_confidence_in_variable	C	For the class lower boundary, the raster file name for the confidence in the environmental variable
	lower_boundary_slope	N Or C	For the class lower boundary, the slope of the GLM/fuzzy law. If the slope varies spatially across the study area, a raster name must be provided instead of a numeric value. The raster cells will indicate the value of the slope
	lower_boundary_intercept	N Or C	For the class lower boundary, the intercept of the GLM/fuzzy law. If the intercept varies spatially across the study area, a raster name must be provided instead of a numeric value. The raster cells will indicate the value of the intercept
	lower_boundary_threshold	N Or C	For the class lower boundary, the probability threshold. If the probability threshold varies spatially across the study area, a raster name must be provided instead of a numeric value. The raster cells will indicate the value of the threshold

<table>
<tr><td></td><td>
<table>
<tr><td>lower_boundary_max_probability</td><td>N
Or
C</td><td>For the class lower boundary, the maximum value that the probability can have (in some cases may be less than 1)
If the maximum probability varies spatially across the study area, a raster name must be provided instead of a numeric value. The raster cells will indicate the value of the maximum probability</td></tr>
<tr><td colspan="3">(*) C=character, N=numerical</td></tr>
</table>
</td></tr>
<tr><td>Outputs</td><td>
For each habitat descriptor class described in the csv file (i.e. each row of the configuration file), the outputs are as follows:

A spatial distribution raster, the cells of which have the value indicated in the csv file column “code”. The name of the tif file is the class short name (indicated in the csv file column “shortName”)
A confidence raster, obtained by combining the raster layer of the confidence in values of the environmental variable and the raster layer of the confidence based on the probability. The name of the tif file is the class short name (indicated in the csv file column “shortName”) + “_confidence_overall”
A raster for the confidence based on the probability. The name of the tif file is the class short name (indicated in the csv file column “shortName”) + “_confidence_based_on_proba.tif”
Optionally, A probability raster. The name of the tif file is the class short name (indicated in the csv file column “shortName”) + “_proba”

All the outputs will be created in the folder ‘output’.
</td></tr>
<tr><td>Scripts parameters</td><td>
<table>
<tr><th>Name</th><th>Type*</th><th>Description</th></tr>
<tr><td>maxmemory</td><td>N</td><td>See section 4.2.4Error! Reference source not found.</td></tr>
<tr><td>workingDirectory</td><td>C</td><td>Full path to the working directory</td></tr>
<tr><td>config_csvFileName</td><td>C</td><td>Name of the configuration name</td></tr>
<tr><td>habitat_descriptor_probability_rasters_as_output</td><td>B</td><td>If TRUE for each habitat descriptor class a raster file will be created for the probability rasters. If FALSE, no file will be created</td></tr>
<tr><td colspan="3">(*) C=character, N=numerical , B=Boolean</td></tr>
</table>
</td></tr>
<tr><td>How to run it?</td><td>

Create a configuration file and put it in the Config_files folder
Open the script in RStudio
Edit the script parameters described above
Click Button ‘Source’ in RStudio

</td></tr>
</table>

4.3.2 Merging the individual habitat descriptor classes in a single layer

Script name	habitat_descriptor_merging_classes.R
Author	Mickaël Vasquez
Date	14/04/2023
Descr.	For one or more habitat descriptors • All the spatial distribution rasters for each individual class are merged into a single raster layer • All the confidence rasters for each individual class are merged into a single raster layer • All the confidence based on probability rasters for each individual class are merged into a single raster layer • All the habitat presence probability rasters for each individual class are merged into a single raster layer
Workflow	Figure 2, (see number 2)
Inputs	All inputs are described via a csv configuration file. The csv file may describe the inputs for several habitat descriptors (i.e. no need to prepare a csv file for each habitat descriptor). For each habitat descriptor, the csv file contains one row for each raster layer that that will be merged in the output layer (so one row per habitat descriptor class to be merged).

Column Name	Type*	Description
habitat_descriptor_shortName	C	The name of the habitat descriptor to which the class is attached. **Crucial**. it will be used in the name of the output rasters. **Must not** contain special characters
class_fileName	C	The name of the input raster file that characterises the habitat descriptor class spatial distribution
proba_fileName	C	The name of the input raster file that characterises the habitat descriptor class presence probability
based_on_proba_confidence_fileName	C	The name of the input raster file that characterises the confidence in the habitat descriptor class occurrence based on the probability
overall_confidence_fileName	C	The name of the input raster file that characterises the overall confidence in the habitat descriptor class occurrence

<table>
<tr><td></td><td>

folder	C	The name of the folder that contains the input raster file. Classically the output folder

(*) C=character, N=numerical

</td></tr>
<tr><td>Outputs</td><td>

For each habitat descriptor described in the csv file, creates (all the outputs will be created in the folder 'output'):

- A spatial distribution raster, the name of which is what is indicated in the csv file in the column "habitat_descriptor_shortName"
- A probability raster, result of all the class probability rasters merging, the name of which is what is indicated in the csv file in the column "habitat_descriptor_shortName" + "_proba"
- A confidence based on probability raster, result of all the class confidence based on probability rasters merging, the name of which is what is indicated in the csv file in the column "habitat_descriptor_shortName" + "_confidence_based_on_proba"
- An overall confidence raster, result of all the class overall confidence raster rasters merging, the name of which is what is indicated in the csv file in the column "habitat_descriptor_shortName" + "_confidence_overall"

</td></tr>
<tr><td>Scripts parameters</td><td>

Name	Type*	Description
maxmemory	N	See section 4.2.4
workingDirectory	C	Full path to the working directory
config_csvFileName	C	Name of the configuration name
output_habitat_descriptor_raster	B	If TRUE a raster file will be created for the spatial distribution. If FALSE no file will be created
output_overall_confidence	B	If TRUE a raster file will be created for the overall confidence. If FALSE no file will be created
output_confidence_based_on_proba	B	If TRUE a raster file will be created for the confidence based on probability. If FALSE no file will be created
output_probability_raster	B	If TRUE a raster file will be created for the probability. If FALSE no file will be created

(*) C=character, N=numerical, B=boolean

</td></tr>
<tr><td>How to run it?</td><td>

- Create a configuration file and put it in the Config_files folder
- Open the script in RStudio
- Edit the script parameters described above
- Click Button 'Source' in RStudio

</td></tr>
</table>

4.3.3 Cleaning up a habitat descriptor raster

A habitat descriptor raster may need to be cleaned up, e.g. after modelling a group of raster cells may have been misclassified. This is done by the script.

<table>
<tr><td>Script name</td><td>habitat_descriptor_calculating_combined_energy.R</td></tr>
<tr><td>Author</td><td>Mickaël Vasquez</td></tr>
<tr><td>Date</td><td>26/06/2023</td></tr>
<tr><td>Descr.</td><td>A shapefile is provided as input, containing a series of polygon within which the raster has to be cleaned up.

3 types of action are possible:
1) within the polygon the cell values are removed (i.e. replaced by NAs)
2) within the polygon the cell values of a given value are replaced with another value
3) within the polygon a sieve filter is applied (to reduce the 'salt-and-pepper' effect). See this page for further detail on that function of the terra package.

The attributes of the shapefile are the following:
- TYPE: describe the action to be done (-1: remove values inside the polygon, 1: replace a pixel value with another value; 2: apply a sieve filter)
- IN_VAL: for TYPE=1 only. Value to be replaced
- OUT_VAL: for TYPE=1 only. new value
- MAX_SIZE: for TYPE=2 only. Corresponds to the 'threshold' parameter of the terra package function sieve</td></tr>
<tr><td>Outputs</td><td>A raster dataset, which is the corrected version of the input raster</td></tr>
<tr><td>Scripts parameters</td><td>
<table>
<tr><th>Name</th><th>Type*</th><th>Description</th></tr>
<tr><td>workingDirectory</td><td>C</td><td>Full link to the working directory (i.e. the directory that contains the habitat shapefile)</td></tr>
<tr><td>clean_up_shape_file</td><td>C</td><td>Path to the input shapefile that contains the polygons and describes the action to be done</td></tr>
<tr><td>input_raster_file</td><td>C</td><td>Path to the raster to be corrected</td></tr>
<tr><td>output_raster_file</td><td>C</td><td>Path to output raster to be corrected</td></tr>
<tr><td colspan="3">(*) C=character, B=Boolean</td></tr>
</table>
</td></tr>
<tr><td>How to run it?</td><td>• Open the script in RStudio
• Edit the script parameters described above
• Click Button 'Source' in RStudio</td></tr>
</table>

4.4 Creating the habitat layer and overall confidence

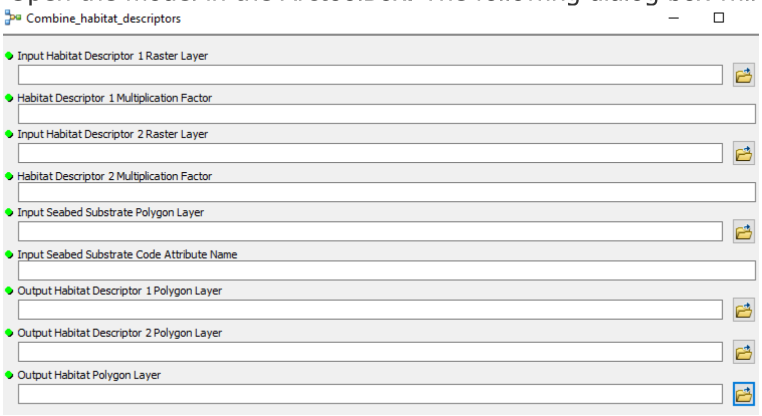
4.4.1 Combining the habitat descriptor layers

<table>
<tr><td>ArcGIS™ Model Builder Model name</td><td>Combine_habitat_descriptors</td></tr>
<tr><td>Author</td><td>Mickaël Vasquez</td></tr>
<tr><td>Date</td><td>04/2021</td></tr>
<tr><td>Descr.</td><td>Requires ArcGIS™ 10.0 (or higher). Creates the habitat layer, which includes:
1) Converting the modelled habitat descriptor rasters into polygon layers
2) Overlaying (by intersection) all the habitat descriptor polygon layers, including the seabed substrate (which is expected to be already in the form of a polygon layer)
3) Calculating the habitat code, which is the sum of: seabed substrate code + (habitat descriptor 1 code x habitat descriptor 1 multiplication factor) + (habitat descriptor 2 code x habitat descriptor 2 multiplication factor). For the multiplication factor, the user may use the convention elaborated for EUSeaMap (see table 4)

Note: the model assumes that there are 2 modelled habitat descriptors, as this is typically the case in EUSeaMap. If there are more habitat descriptors, the model will need to be edited.</td></tr>
<tr><td>Worklow reference</td><td>Figure 1, (numbers 2 and 3)</td></tr>
<tr><td>Inputs</td><td>
<table>
<tr><th>Model Variable name</th><th>Type</th><th>Description</th></tr>
<tr><td>Input Seabed Substrate Polygon Layer</td><td>Feature class</td><td>The input Seabed Substrate polygon layer</td></tr>
<tr><td>Input Seabed Substrate Code Attribute Name</td><td>String</td><td>The name of the seabed substrate polygon layer attribute that describes the seabed substrate code</td></tr>
<tr><td>Input Habitat Descriptor 1 Raster Layer</td><td>Raster dataset</td><td>The input habitat descriptor 1 raster layer</td></tr>
<tr><td>Habitat Descriptor 1 Multiplication Factor</td><td>Variant</td><td>The multiplication factor for the Habitat Descriptor 1. The user may use the convention elaborated for EUSeaMap (see table 4)</td></tr>
<tr><td>Input Habitat Descriptor 2 Raster Layer</td><td>Raster dataset</td><td>The input habitat descriptor 2 raster layer</td></tr>
<tr><td>Habitat Descriptor 2 Multiplication Factor</td><td>Variant</td><td>The multiplication factor for the habitat descriptor 2. The user may use the convention elaborated for EUSeaMap (see table 4)</td></tr>
</table>
</td></tr>
</table>

<table>
<tr><td rowspan="4">Outputs</td><td>Model Variable name</td><td>Type</td><td>Description</td></tr>
<tr><td>Output Habitat Descriptor 1 Polygon Layer</td><td>Feature class</td><td>The habitat descriptor 1 polygon obtained by the raster-to-polygon conversion. is recommended to create a geodatabase feature class rather than a shapefile</td></tr>
<tr><td>Output Habitat Descriptor 2 Polygon Layer</td><td>Feature class</td><td>The habitat descriptor 2 polygon obtained by the raster-to-polygon conversion. It is recommended to create a geodatabase feature class rather than a shapefile</td></tr>
<tr><td>Output Habitat Polygon Layer</td><td>Feature class</td><td>The habitat polygon obtained by intersection of habitat descriptor 1 polygon layer, the habitat descriptor 2 polygon layer, and the seabed substrate. It contains an attribute, “hab_code”, the values of which are the sum of: seabed substrate code + (habitat descriptor 1 code x habitat descriptor 1 Multiplication Factor) + (habitat descriptor 1 code x habitat descriptor 1 Multiplication Factor). It is recommended to create a geodatabase feature class rather than a shapefile</td></tr>
<tr><td>How to run it?</td><td colspan="3">

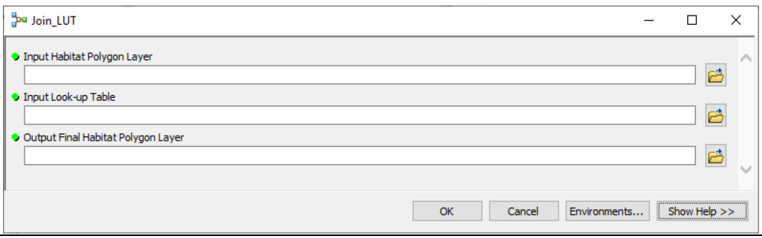
- Open the model in the ArctoolBox. The following dialog box will appear



If an update of the tool is required, (e.g. add an habitat descriptor), edit it</td></tr>
<tr><td>Main ArcGIS™ tools used</td><td colspan="3">

- Raster to Polygon (Conversion)
- Intersect (Analysis)</td></tr>
</table>

4.4.2 Joining the habitat polygon layer and the look-up table that crosswalks the modelled habitat codes and habitat classifications

<table>
<tr><td>ArcGIS™ Model Builder Model name</td><td>Join_LUT</td></tr>
<tr><td>Author</td><td>Mickaël Vasquez</td></tr>
<tr><td>Date</td><td>09/2023</td></tr>
<tr><td>Descr.</td><td>Joins the habitat shapefile or geodatabase feature class produced by the model “Combine_habitat_descriptors” (see section 4.4.1), in which habitats are described via a code, to a look-up table (in dbase or geodatabase table format) that crosswalks each habitat code with various habitat classification (e.g. EUNIS).</td></tr>
<tr><td>Worklow reference</td><td>Figure 1, (see number 4)</td></tr>
<tr><td>Outputs</td><td>A shapefile or (preferably) a geodatabase feature class containing the attributes of the look-up table</td></tr>
<tr><td>Scripts parameters</td><td>
<table>
<tr><th>Model Variable name</th><th>Type</th><th>Description</th></tr>
<tr><td>Input Habitat Polygon Layer</td><td>C</td><td>The Habitat polygon layer. It is recommended to use a geodatabase feature class rather than a shapefile</td></tr>
<tr><td>Input Look-up Table</td><td>C</td><td>Look-up table. In dbase or geodatabase table format</td></tr>
<tr><td>Output Final Habitat Polygon Layer</td><td>C</td><td>The output Habitat polygon layer</td></tr>
</table>
</td></tr>
<tr><td>How to run it?</td><td>
Open the model in the ActoolBox. The following dialog box will appear


</td></tr>
<tr><td>Main ArcGIS™ tools used</td><td>Multipart To Singlepart, Join Field, Repair Geometry</td></tr>
</table>

4.4.3 Creating the habitat confidence layers

<table>
<tr><td>Script name</td><td>habitat_map_calculating_confidence.R</td></tr>
<tr><td>Author</td><td>Mickaël Vasquez</td></tr>
<tr><td>Date</td><td>05/04/2023</td></tr>
<tr><td>Descr.</td><td>For one or more habitat classifications, create the habitat confidence layer, which is the combination of the individual confidence layers of the habitat descriptors that are involved in the habitat classification. For example, in the EUNIS classification the habitat descriptors that are involved (seabed substrate, biological zone and energy level) are not the same as the ones involved in the MSFD broad habitat types classification (seabed substrate and biological zone only). For each cell, the result value is the minimum confidence value of the input habitat descriptor confidence layers.</td></tr>
<tr><td>Workflow</td><td>Figure 1, (see number 3b). In figure 1, in the interest of readability only one output is represented. But there can be more than one depending on the number of classifications.</td></tr>
<tr><td>Inputs</td><td>All inputs are described via a csv configuration file. The csv file may describe the inputs for several habitat classifications (i.e. no need to prepare a csv file for each classification). For each habitat classification, the csv file contains one row for each habitat descriptor confidence raster layer that that will be combined in the output layer (so one row per habitat descriptor confidence to be merged).
<table>
<tr><th>Column Name</th><th>Type*</th><th>Description</th></tr>
<tr><td>habitat_classification_name</td><td>C</td><td>The name of the habitat classification for which the habitat descriptor is used. Crucial. it will be used in the name of the output raster. Must not contain special characters</td></tr>
<tr><td>confidence_fileName</td><td>C</td><td>The name of the habitat descriptor confidence raster file</td></tr>
<tr><td>folder</td><td>C</td><td>The path to the folder that contains the file, relatively to the working directory</td></tr>
<tr><td colspan="3">(*) C=character, N=numerical</td></tr>
</table></td></tr>
<tr><td>Outputs</td><td>For each habitat classification described in the csv file, creates (in the folder ‘output’) the overall confidence raster, the name of which is what is indicated in the csv file in the column “habitat_classification_name”+” _confidence_overall”</td></tr>
<tr><td>Scripts parameters</td><td><table>
<tr><th>Name</th><th>Type*</th><th>Description</th></tr>
<tr><td>maxmemory</td><td>N</td><td>See section Error! Reference source not found.</td></tr>
<tr><td>workingDirectory</td><td>C</td><td>Full path to the working directory</td></tr>
</table></td></tr>
</table>

	config_csvFileName	C	Name of the configuration name
	(*) C=character, N=numerical		
How to run it?	• Create a configuration file and put it in the Config_files folder • Open the script in RStudio • Edit the script parameters described above • Click Button 'Source' in RStudio		

4.5 Additional script: combining the wave-induced and current-induced habitat descriptor layers

Not shown in figure 1 (because specific to the energy habitat descriptor, which itself is specific to the EUNIS 2007 classification and to the Atlantic region) is the step that comprises combining the wave-induced and current-induced habitat descriptor layers and their respective confidence layers. The script performs this intermediate task.

Script name	habitat_descriptor_calculating_combined_energy.R
Author	Mickaël Vasquez
Date	03/05/2023
Descr.	Calculates the overall energy level layer by combining the current-induced energy level layer and the wave-induced energy level layer. The rule is to keep the maximum value of the two layers is kept, e.g. if for a cell the current is moderate and the wave is low, then the combined energy is moderate. When it comes to the confidence, for each cell the confidence is that of the habitat descriptor that contributes to the overall energy level, i.e. if for a cell the combined energy is that of wave (because wave energy level > current energy level), then the combined confidence is that of wave. If for a cell the combined energy level is that of wave and current (because wave energy level = current energy level), the confidence is the average confidence of wave energy level and current energy level, rounded up. In addition to the full-coverage energy level confidence raster, a confidence raster specific to EUNIS 2007-11 is created; In this raster, only cells where the seabed substrate is rock (i.e. cell value = 70) and the biozone is infralittoral or circalittoral (i.e. cell value ≤ 30) are assigned a confidence value (i.e. all other cells are assigned NA).
Worklow reference	Not shown in figure 1, but the outputs of figure 1/step 1 are individual habitat descriptor layers for current and wave. These 2 habitat descriptor layers then need to be combined in order to produce a wave/current-induced energy layer, which together with the other habitat descriptor layers (biological zones, seabed substrate) is then provided as input to figure 1/step 2, Similarly, figure 1/step 3b requires as input a combined wave/current confidence layer.
Outputs	3 raster files: • One for the combined wave/current energy

	• One for the confidence in the combined wave/current energy • One for the confidence in the combined wave/current energy for the EUNIS 2007-11 classification
Scripts parameters	see table below
How to run it?	• Open the script in RStudio • Edit the script parameters described above • Click Button 'Source' in RStudio

Name	Type*	Description
workingDirectory	C	*Full link to the working directory (i.e. the directory that contains the habitat shapefile)*
current_fileName	C	path to the input raster file corresponding to the current-induced energy
current_confidence_fileName	C	path to the input confidence raster file corresponding to the current-induced energy
wave_fileName	C	path to the input raster file corresponding to the current-induced energy
wave_confidence_fileName	C	path to the input confidence raster file corresponding to the current-induced energy
biozone_raster_fileName	C	path to the input raster file corresponding to the biozones. Will be used to create of the EUNIS2007-11 confidence raster
substrate_raster_fileName	C	path to the input raster file corresponding to the seabed substrate. Will be used to create of the EUNIS2007-11 confidence raster
ouput_fileName	C	path to the output energy file name
ouput_confidence_fileName	C	path to the output confidence file
ouput_EUNIS_confidence_fileName	C	path to the output EUNIS2007-11 confidence file
calculateConfidence	**B**	Is it required to calculate confidence? If FALSE, the confidence layer is not calculated

(*) C=character, B=Boolean

5 References

Coltman, N., Golding, N., Verling, E., 2008. Developing a broadscale predictive EUNIS habitat map for the MESH study area. 16 pp. Available online at www.emodnet-seabedhabitats.eu/pdf/MESH%20EUNIS%20model.pdf

Populus J., Vasquez M., Albrecht J., Manca E., Agnesi S., Al Hamdani Z., Andersen J., Annunziatellis A., Bekkby T., Bruschi A., Doncheva V., Drakopoulou V., Duncan G., Inghilesi R., Kyriakidou C., Lalli F., Lillis H., Mo G., Muresan M., Salomidi M., Sakellariou D., Simboura M., Teaca A., Tezcan D., Todorova V. and Tunesi L., 2017. EUSeaMap, a European broad-scale seabed habitat map. 174p. http://doi.org/10.13155/49975

Vasquez M., Manca E., Inghilesi R., Martin S., Agnesi S., Al Hamdani Z., Annunziatellis A., Bekkby T., Pesch R., Askew N., Bentes L., Castle L., Doncheva V., Drakopoulou V., Gonçalves J., Laamanen L., Lillis H., Loukaidi V., McGrath F., Mo G., Monteiro P., Muresan M., O'Keeffe E., Populus J., Pinder J., Ridgeway A., Sakellariou D., Simboura M., Teaca A., Tempera F., Todorova V., Tunesi L., Virtanen E. (2020). EUSeaMap 2019, A European broad-scale seabed habitat map, technical report . EASME/EMFF/2018/1.3.1.8/Lot2/SI2.810241– EMODnet Thematic Lot n° 2 – Seabed Habitats. https://doi.org/10.13155/74782

6 Appendix 1: the habitat descriptors used for EUSeaMap, their classes and codes

Seabed substrate

2-digit code	Name
0	Seabed
10	Fine mud
20	Sand
30	Coarse substrate
40	Mixed sediment
50	Sandy mud
60	Muddy sand
70	Rock or other hard substrata
97	Sediment
98	Sandy mud or Muddy sand
99	Fine mud or Sandy mud or Muddy sand
71	[Posidonia oceanica] meadows
72	[Posidonia oceanica] "Barrier-reef"
73	Dead mattes of [Posidonia oceanica]
74	Coralligenous platforms
75	Facies with [Vermetus] spp. of the infralittoral algae biocenosis
76	Facies with [Ficopomatus enigmaticus] of the euryhaline and/or eurythermal lagoon biocenosis
1	Biogenic substrate
2	Worm reefs
77	[Sabellaria alveolata] reefs

2-digit code	Name
78	[Sabellaria spinulosa] reefs
79	[Serpula vermicularis] reefs
3	Bivalve reefs
80	[Limaria hians] beds
81	[Ostrea edulis] beds
4	Mussel beds
82	[Mytilus edulis] beds
83	[Modiolus modiolus] beds
84	dominated by zebra mussel
85	dominated by valve snails
86	[Hiatella arctica] beds
5	Cold water coral reefs
6	Tropical coral reefs
87	[Lophelia pertusa] reefs
88	[Solenosmilia variabilis] reefs
89	Shell gravel
90	Peat bottoms
91	Biogenic detritic bottoms
92	dominated by [Mytilaster lineatus] and [Amphibalanus improvisus]
93	dominated by [Dreissena grimmi] and corophiids

Energy

2-digit code	Name
0	No Energy data
10	Low energy
20	Moderate energy
30	High energy

Oxygen

2-digit code	Name
0	No oxygen data
10	oxic
20	suboxic
30	anoxic

Salinity

2-digit code	Name
0	No salinity data
10	Oligohaline
20	Mesohaline
30	Polyhaline
60	Euhaline

Mask

1-digit code	Name
1	No mask
2	Mask 1
3	Mask 2

Biological zones

2-digit code	Name
0	No information
10	Infralittoral
20	Shallow circalittoral
30	Deep circalittoral
40	Bathyal
50	Abyssal
52	Atlanto-Arctic mid bathyal
60	Upper bathyal
61	Atlantic upper bathyal in North Arctic
62	Atlanto-Arctic upper bathyal
63	Atlantic upper bathyal
64	Atlanto-Mediterranean upper bathyal
65	Mid bathyal
66	Atlanto-Arctic upper bathyal in North arctic
67	Arctic mid bathyal
68	Atlantic mid bathyal
69	Atlanto-Mediterranean mid bathyal
70	Lower bathyal
71	Arctic lower bathyal
72	Atlanto-Arctic lower bathyal

2-digit code	Name
73	Atlantic lower bathyal
74	Atlanto-Mediterranean lower bathyal
75	Upper abyssal
76	Arctic upper abyssal
77	Atlanto-Arctic upper abyssal
78	Atlantic upper abyssal
79	Atlanto-Mediterranean upper abyssal
80	Mid abyssal
81	Arctic mid abyssal
82	Atlanto-Arctic mid abyssal
83	Atlantic mid abyssal
84	Atlanto-Mediterranean mid abyssal
85	Lower abyssal
86	Arctic lower abyssal
87	Atlanto-Arctic lower abyssal
88	Atlantic lower abyssal
89	Atlanto-Mediterranean lower abyssal
91	Atlantic upper bathyal in South Arctic
96	Atlanto-Arctic upper bathyal in South Arctic

7 Appendix 2: calculating the slope, intercept and probability threshold for fuzzy laws

Figure 2.1 illustrates a fuzzy law for one of the two boundaries of a habitat descriptor class (e.g. lower boundary of the class 'shallow circalittoral'). In abscissa are the driver values (e.g. temperature). In ordinate is the probability of occurrence for the habitat descriptor class. The shape of the fuzzy function is governed by two control points, P0 and P1. P0 (x0,0) indicates where the probability begins to increase above 0. P1(x1,1) is the point where the probability starts to be 1.

In-between is a simple straight line, whose slope a and intercept b are defined as:

a = 1 / (x1-x0)

b = -x0 / (x1-x0)

As with a GLM approach, a probability threshold (or cut-off) value has to be worked out. It is the probability value above which the habitat descriptor class will be classified as present and below which it will be classified as absent.

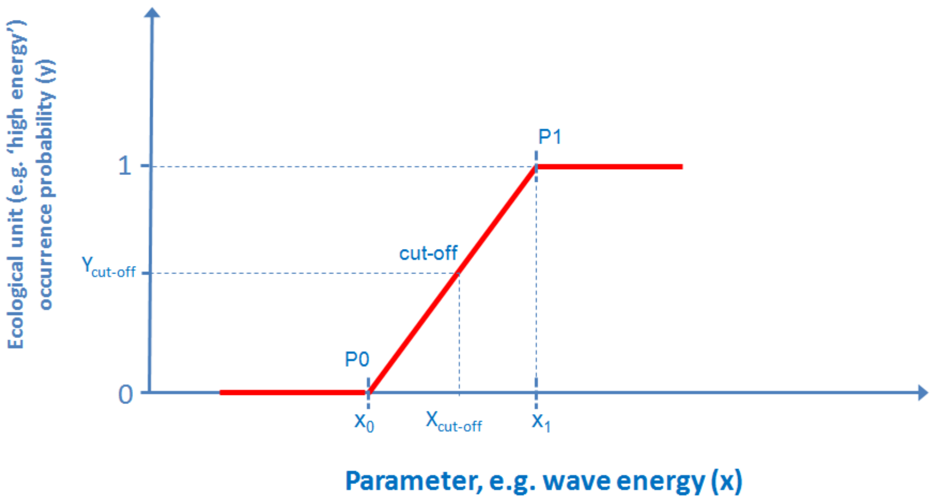


Figure 2.1. The fuzzy classifier shape is governed by 2 control points P0 and P1. The cut-off is the point whose y coordinate is the probability value above which the habitat descriptor class will be classified as present and below which it will be classified as absent.

Usually, in a threshold analysis, the values that are worked out are X_0, X_1 and $x_{cut\text{-}off}$, while the GIS workflow requires as input the slope, the intercept and the probability threshold ($Y_{cut\text{-}off}$ in figure 2.1). An excel file that is provided in the supplemental material will facilitate the calculation of these items from x_0, x_1 and $X_{cut\text{-}off}$. The name of the file is "slope_intercept_fuzzy_calculation.xlsx".

	X1	X0	threshold	slope	intercept	probability threshold
Infra lower boundary	2.27	1.19	1.82	0.93	-1.10	0.58
shallow circa upper boundary	1.19	2.27	1.82	-0.93	2.10	0.42
shallow circa lower boundary	0.00075	0.00025	0.0005	2000.00	-0.50	0.50
deep circa upper boundary	0.00025	0.00075	0.0005	-2000.00	1.50	0.50

1 / (X1-X0) -X0 / (X1-X0) threshold x slope + intercept

Figure 2.2. Screenshot of the excel file provided for the calculation of the slope, intercept and probability threshold from x_0, x_1 and $X_{cut\text{-}off}$

8 Appendix 3: Externally defined boundaries

Externally defined boundaries are used in specific regions and for specific reasons. For example, in the Black Sea, Mediterranean Sea, Caspian Sea, Caribbean Sea, the choice has been made to define the deep circalittoral / bathyal and the bathyal / abyssal boundaries as slope changes. The line delineating these slope changes is semi-automatically drawn by experts. Other examples are the Baltic and the Adriatic Sea, where the infralittoral / circalittoral boundary was defined externally based on local studies.

This type of boundary requires a special integration into EUSeaMap. This appendix deals with the way the confidence is calculated and the type of input required.

Calculation of the confidence in the classification

The rule that governs the way confidence in the classification is addressed in EUSeaMap is always the same: the closer we are to the boundary of a habitat descriptor class, the less confidence we have in the classification. Conversely, the further we are from the boundary the more confidence we have in the classification. To implement this rule when the boundary is provided to the workflow 'out of the box' (i.e. not calculated by the workflow itself), we use the distance to the boundary as the input variable and an adapted fuzzy classifier to model the presence of the habitat descriptor class and to provide a measure of the confidence in the classification (fig. 3.1): at the boundary (i.e. distance=0), the probability is set to 0.5. Up to a certain distance X_1 (e.g. 0.04 decimal degrees), the probability increases linearly up to 1. Above that distance, the probability is always 1.

As a result:

- Slope = $0.5/X_1$
- Intercept = 0.5

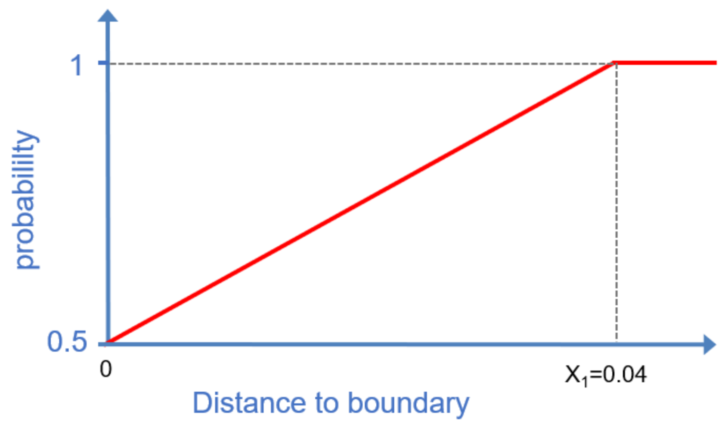


Figure 3.1. fuzzy classifier used when an 'out of the box' boundary is provided to the workflow

Input variable

The input variable is the distance to the boundary. For a habitat descriptor class **upper** boundary, the raster cells that are within and below the class have distance values, while raster cells that are above the class have value 0. For a habitat descriptor class **lower** boundary, the raster cells that are within and above the class have distance values, while raster cells that are below the class have value 0.

For example (fig. 3.2), if the boundary is the bathyal **upper** boundary, the raster cells that are within and below the bathyal (i.e. within the bathyal or the abyssal) have distance values, while raster cells that are above (i.e. within the deep circalittoral, the shallow circalittoral or the infralittoral), the bathyal have value 0. For example, if the boundary is the bathyal **lower** boundary, the raster cells that are within and above the bathyal (i.e. within the deep circalittoral, the shallow circalittoral or the infralittoral) have distance values, while raster cells that are below (i.e. within the bathyal or the abyssal), the bathyal have value 0.

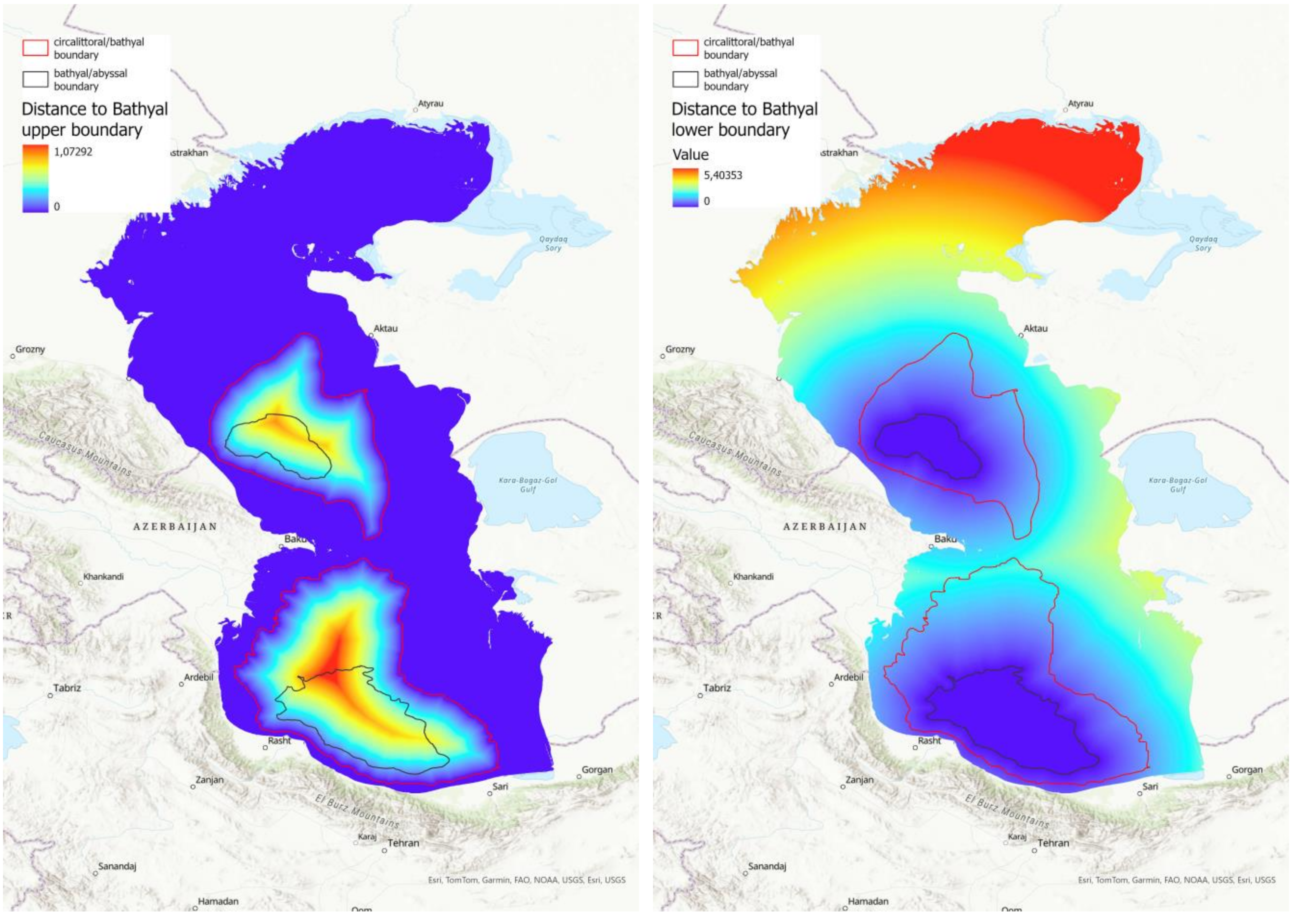


Figure 3.2. distance input variable for the bathyal upper boundary (left) and the bathyal lower boundary (right)

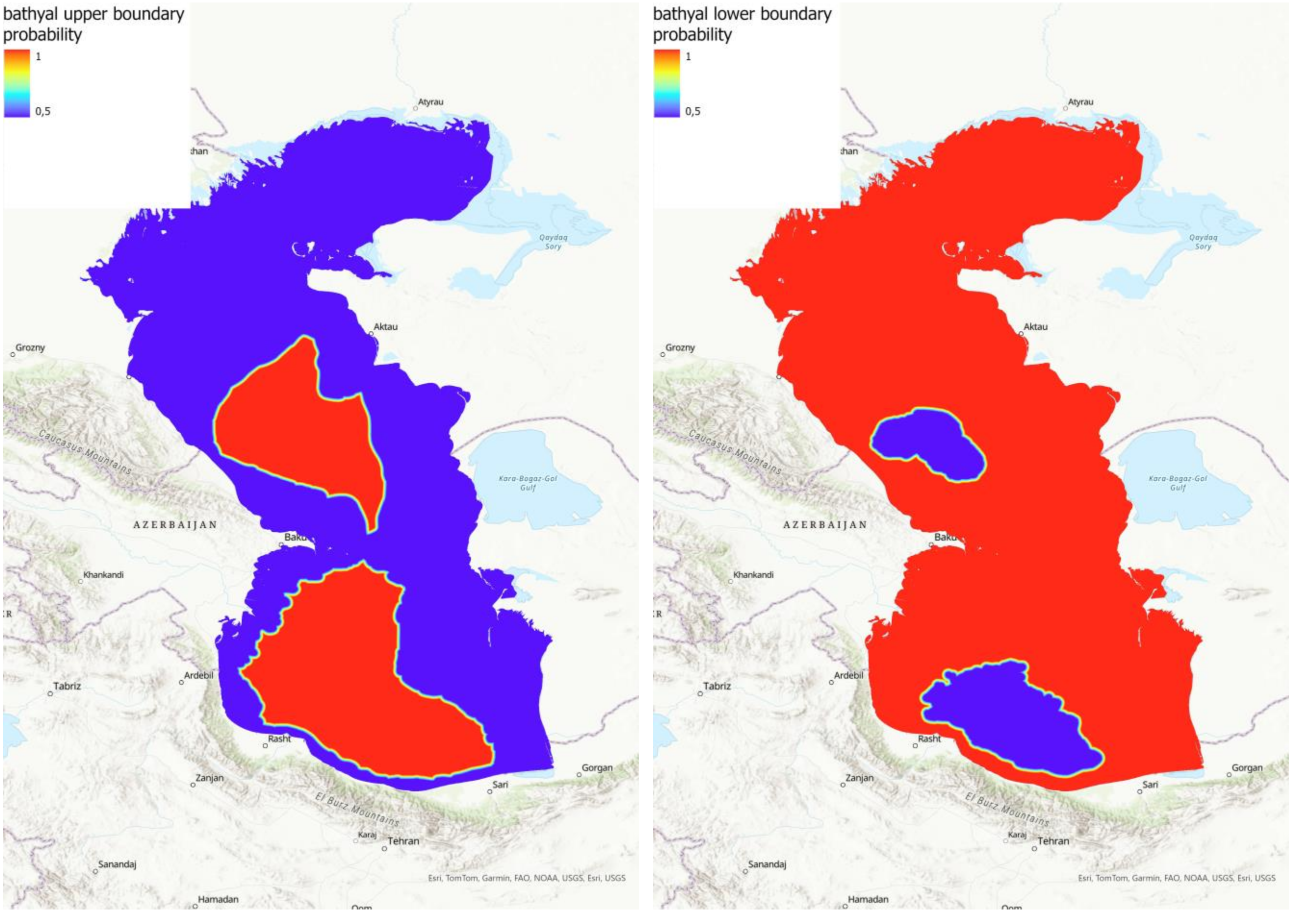


Figure 3.3. bathyal upper boundary probability (left) and bathyal lower boundary probability (right) when using the fuzzy classifier of fig. 3.1 with input raster datasets of figure 3.2

9 Appendix 4: habitat attribute table format

Field Name	Description
Oxygen	Oxygenation level at the seabed. Used in the Black Sea only. Possible values are in Appendix 1.
Salinity	Salinity level at the seabed. Used in the Baltic Sea only. Possible values are in Appendix 1.
Energy	Combined current-and wave-induced energy at the seabed. Possible values are in Appendix 1.
Biozone	Biological zones. Possible values are in Appendix 1.
Substrate	Seabed substrate type. Possible values are in Appendix 1.
EUNIScomb	Habitat description using EUNIS 2007-2011 code (e.g. 'A5.35'). Na where EUNIS 2007-2011 is not applicable.
EUNIScombD	EUNIS 2007-2011 full description (e.g. 'A5.35: Circalittoral sandy mud'). Na where EUNIS 2007-2011 is not applicable.
Allcomb	Habitat description using EUNIS 2007-2011 code (e.g. 'A5.35') where EUNIS 2007-2011 is applicable, or other unpublished classification where 2007-2011 is not applicable.
AllcombD	Habitat description using EUNIS 2007-2011 full description (e.g. A5.35: Circalittoral sandy mud) where EUNIS 2007-2011 is applicable. Description using other unpublished classification (e.g. 'Low energy infralittoral seabed') where 2007-2011 is not applicable.
SalcombD	Used in the Baltic Sea only. Habitat description using an unpublished habitat classification that describes salinity level (e.g. 'Deep circalittoral mixed sediments in Oligohaline').
MSFD_BBHT	Habitat description using the MSFD Benthic Broad Habitat types (as defined in COMMISSION DECISION (EU) 2017/848). Na where the classification is not applicable.
EUNIS2019C	Habitat description using EUNIS 2022 code (e.g. 'MB23'). Na where EUNIS 2022 is not applicable.
EUNIS2019D	Habitat description using EUNIS 2022 full description (e.g. 'MB23: Baltic infralittoral biogenic habitat'). Na where EUNIS 2022 is not applicable.
All2019D	Habitat description using EUNIS 2022 full description (e.g. 'MB23: Baltic infralittoral biogenic habitat') where EUNIS 2022 is applicable, or other unpublished classification (e.g. 'Baltic infralittoral seabed') where EUNIS 2022 is not applicable.

All2019DL2	Habitat description using EUNIS 2022 description at level 2 (e.g. 'MB5: Infralittoral sand'), or other unpublished classification (e.g. 'Infralittoral seabed') where EUNIS 2022 is not applicable.
RegionalD	Habitat description using a published regional classification. Used in the Baltic (HELCOM HUB classification) and the Mediterranean Sea (revised classification of Mediterranean habitat types). Na when the classification is not applicable.
Val_comm	Comment.

10 Appendix 5: Hands-on training

In exercises 1 to 3 we will model the spatial distribution of the following habitat descriptor classes:

- Biological zones
- Wave-induced energy levels
- Current-induced energy levels

Their respective confidence raster layers will also be created.

In exercise 4, the class spatial distribution layers will be merged so that all classes of a habitat descriptor are combined into one layer.

In exercise 5, the habitat descriptor layers for current-induced energy levels and wave-induced energy levels will be combined in order to create a combined wave/current-induced energy level layer.

In exercise 6, the habitat layer will be created by combining the 3 habitat descriptor layers 'seabed substrates', 'energy levels' and 'biological zones'. In exercise 7 this layer will be joined to a table that crosswalks the modelled habitats with two standard classifications (EUNIS 2007-11 and MSFD broad habitat types). In exercise 8, the confidence layers will be created.

Exercise 9 is an example of using rasters instead of constant values for the slope and the intercept.

Exercise 10 is an example of using externally-defined boundaries.

Exercise 11 is an example of cleaning up a habitat descriptor raster.

The R scripts and the ArcGIS™ toolbox are available are all available on Github:

https://github.com/emodnetseabedhabitats/EUSeaMap_creation

The input data used for the hands-on training is available here:

https://cloud.ifremer.fr/index.php/s/zl51b605IyN6lDM

1 - Modelling the biological zone classes

1.1 – Understanding the csv configuration file

The csv configuration file describes all the elements required by the script to model a set of habitat descriptor classes. It is therefore essential to become familiar with the organisation of the file. The meaning of the file columns is documented in the guide, section 4.3.1.

- Open the csv configuration file (in the 'config_files' folder, file habitat_descriptor_modelling_biozones.csv). This file describes all the elements required by the script to model the biological zones.

<u>Questions</u>

- For the Infralittoral lower boundary, what is the name of the raster file used as environmental variable?

Answer: variable_seabed_par.tif

- For the Infralittoral lower boundary, what is the name of the raster file used as confidence in the environmental variable?

Answer: confidence_seabed_par.tif

- For the Shallow Circalittoral upper boundary, what is the name of the raster file used as environmental variable?

Answer: variable_seabed_par.tif. Note: it is the same as for the infralittoral lower boundary, which is normal because the infralittoral lower boundary and the shallow circalittoral upper boundary are … the same boundary, i.e. the boundary between the infralittoral and the shallow circalittoral

- For the Shallow Circalittoral lower boundary, what is the name of the raster file used as environmental variable?

Answer: variable_wavebase.tif. Note: this variable is the ratio ʎ/h, where ʎ=wave wavelength and h=depth

- For the Shallow Circalittoral lower boundary, what is the name of the raster file used as confidence in the environmental variable?

Answer: confidence_wavebase.tif

- What is the type of equation used for the infralittoral lower boundary?

Answer: the value is 2, which means that the equation type is GLM

- What is the type of equation used for the abyssal upper boundary?

Answer: the value is 1, which means that the equation type is fuzzy

- why is the max probability for the shallow circalitttoral 0.69 upper boundary and not 1?

Answer: the max probability will be for PAR at seabed=0 mol.pho.m^{-2}.d^{-2}

As slope=-1.076 and intercept=0.777

P(X=0) = exp(-1.076*0+0.777)/ (1+exp(-1.076*0+0.777)) = 0.69

- Check that in the csv file all the input parameters (class code, equation type, slope, intercept, probability thresholds, max probability) defined for each biological zone boundary are the same as those listed in table 3.1.

Was there any error?

1.2 – Exploring the input data

- In a GIS software, open all the input raster files (i.e. *variable* rasters and *confidence in variable* rasters) listed in the csv configuration file and take a look at them individually (remember, they are all in the 'input' folder).

- In the confidence in variable rasters, what are the possible values? What do they mean?

Answer: for a raster cell, the possible values of confidence in the value of the environmental variable are 1, 2 and 3, meaning Low, Moderate and High confidence, respectively.

1.3 – Running the script

- Open habitat_descriptor_modelling.R in Rstudio.

- Assign the script variable `workingDirectory` the path to the working directory (which is the folder containing the folders 'config_files', 'input', and 'ouput ').

- Assign the script variable `config_csvFileName` the name of the configuration file, i.e. 'habitat_descriptor_modelling_biozones.csv'

```
config_csvFileName<-"habitat_descriptor_modelling_biozones.csv"
```

- In Rstudio, click the 'Source' button. The script starts to run. Depending on the capacity of your computer, it may take some time to complete. When it has finished, the Rstudio console will display a message mentioning "Completed in…" followed by the duration.

Table 5.1: variable, classification method and parameters (slope, intercept, probability threshold and max probability) used as input to the script for each biological zone boundary. Other parameters (variable threshold, fuzzy X1 and fuzzy X0) are provided for background information

Habitat descriptor class boundary	Class code	Equation type	Variable	Variable Threshold	Fuzzy X1	Fuzzy X0	Slope	Intercept	Probability threshold	Max probability
Infralittoral Lower Boundary	10	GLM	PAR at seabed (mol.pho.m^{-2}.d^{-1})	0.7	-	-	1.076	-0.777	0.49	1
Shallow Circalittoral Upper Boundary	20	GLM	PAR at seabed (mol.pho.m^{-2}.d^{-1})	0.7	-	-	-1.076	0.777	0.51	0.69
Shallow Circalittoral Lower Boundary	20	GLM	Wave base ratio	1.5	-	-	19.2	-28.7	0.41	1
Deep Circalittoral Upper Boundary	30	GLM	Wave base ratio	1.5	-	-	-19.2	28.7	0.59	1
Deep Circalittoral Lower Boundary	30	Fuzzy	Depth (m)	-200	-180	-220	0.025	5.5	0.5	1
Upper Bathyal Upper Boundary	60	Fuzzy	Depth (m)	-200	-220	-180	-0.025	-4.5	0.5	1
Upper Bathyal Lower Boundary	60	Fuzzy	Depth (m)	-1300	-1017	-1583	0.0018	2.797	0.5	1
Lower Bathyal Upper Boundary	70	Fuzzy	Depth (m)	-1300	-1583	-1017	-0.0018	-1.797	0.5	1
Lower Bathyal Lower Boundary	70	Fuzzy	Depth (m)	-2200	-1912	-2488	0.0017	4.319	0.5	1
Abyssal Upper Boundary	75	Fuzzy	Depth (m)	-2200	-2488	-1912	-0.0017	-3.319	0.5	1

1.4 – Exploring the results

a) Check that there are as many spatial distribution rasters created as there are biological zones listed in the file 'habitat_descriptor_modelling_biozones.csv'. The tif files have been named according to what is written in the 'shortName' column of the csv file. In addition, 2 rasters have been created for each biological zone, i.e. one for the confidence based on the probability (the name of the tif is in the form 'shortName'_confidence_based_on_proba) and another one for the overall confidence (the name of the tif is in the form 'shortName'_confidence_overall), the latter being the combination of i) the former and ii) the raster for confidence in the environmental variable provided as input in the csv file (columns 'upper_boundary_confidence_in_variable' and 'lower_boundary_confidence_in_variable'). The output folder should contain the following files:

Spatial distribution files	Confidence based on probability	Overall confidence
Infra.tif	Infra_confidence_based_on_proba.tif	Infra_confidence_overall.tif
Shallow_Circa.tif	Shallow_Circa_confidence_based_on_proba.tif	Shallow_Circa_confidence_overall.tif
Deep_Circa.tif	Deep_Circa_confidence_based_on_proba.tif	Deep_Circa_confidence_overall.tif
Upper_Bathyal.tif	Upper_Bathyal_confidence_based_on_proba.tif	Upper_Bathyal_confidence_overall.tif
Lower_Bathyal.tif	Lower_Bathyal_confidence_based_on_proba.tif	Lower_Bathyal_confidence_overall.tif
Abyssal.tif	Abyssal_confidence_based_on_proba.tif	Abyssal_confidence_overall.tif

b) Take some time to view these rasters in a GIS software. Check that the code assigned to each class spatial distribution raster cell matches the code given in the 'code' column of the csv configuration file.

2 - Modelling the current-induced energy classes

2.1 – Examining the csv configuration file

- Open the csv configuration file (in the folder 'config_files', file habitat_descriptor_modelling_currents.csv). This file describes all the elements required by the script to model the current energy levels. The meaning of the file columns is documented in the guide, section 4.3.1.

- Check that all the input parameters (class code, equation type, slope, intercept, probability thresholds, max probability) defined for each energy level boundary are the same as those listed in table 3.2.

Was there any error?

Table 5.2: variable, classification method and parameters (slope, intercept, probability threshold and max probability) used as input to the script for each current level boundary. Other parameters (variable threshold, fuzzy X1 and fuzzy X0) are provided for background information

Habitat descriptor class boundary	Class code	Equation type	Variable	Variable Threshold	Fuzzy X1	Fuzzy X0	Slope	Intercept	Probability threshold	Max probability
Current High Lower Boundary	10	Fuzzy	Current-induced kinetic energy ($N.m^{-2}$)	400	450	350	0.01	-3.5	0.5	1
Current Moderate Upper Boundary	20	Fuzzy	Current-induced kinetic energy ($N.m^{-2}$)	400	350	450	-0.01	4.5	0.5	1
Current Moderate Lower Boundary	20	Fuzzy	Current-induced kinetic energy ($N.m^{-2}$)	80	100	60	0.025	-1.5	0.5	1
Current Low Upper Boundary	30	Fuzzy	Current-induced kinetic energy ($N.m^{-2}$)	80	60	100	-0.025	2.5	0.5	1

2.2 – Exploring the input data

Open in a GIS software all the input raster files (i.e. *variable* rasters and *confidence in variable* rasters) listed in the csv configuration file and take a look at them individually (remember, they are all in the 'input' folder).

2.3 – Running the script

- Open habitat_descriptor_modelling.R in Rstudio.

- Assign the script variable `workingDirectory` the path to the working directory (which is the folder that contains the 'config_files', 'input', and 'ouput ').

- Assign the script variable `config_csvFileName` the name of the configuration file, i.e. 'habitat_descriptor_modelling_currents.csv'

```
config_csvFileName<-"habitat_descriptor_modelling_currents.csv"
```

- In Rstudio, click the 'Source' button. The script starts to run. Depending on the capacity of your computer, it may take some time to complete. When it is finished, the Rstudio console will display a message saying "Completed in…" followed by the duration.

2.4 – Exploring the results

a) Check that all the rasters have been created. The 'output' folder should contain the following files:

Spatial distribution files	Confidence based on probability	Overall confidence
Currents_High.tif	Currents_High_confidence_based_on_proba.tif	Currents_High_confidence_overall.tif
Currents_Moderate.tif	Currents_Moderate_confidence_based_on_proba.tif	Currents_Moderate_confidence_overall.tif
Currents_low.tif	Currents_Low_confidence_based_on_proba.tif	Currents_Low_confidence_overall.tif

b) Take some time to view these rasters in a GIS software. Check that the code assigned to each class spatial distribution raster cell matches the code given in the 'code' column of the csv configuration file.

3 - Modelling the wave-induced energy classes

3.1 –Examining the csv configuration file

- Open the csv configuration file (in the folder 'config_files', file habitat_descriptor_modelling_wave.csv). This file describes all the items required by the script to model the wave energy levels. The meaning of the file columns is documented in the guide, section 4.3.1.

- Check that all the input parameters (class code, equation type, slope, intercept, probability thresholds, max probability) defined for each energy level boundary are the same as those listed in table 3.3.

Was there any error?

Table 5.3: variable, classification method and parameters (slope, intercept, probability threshold and max probability) used as input to the script for each wave level boundary. Other parameters (variable threshold, fuzzy X1 and fuzzy X0) are provided for background information

Habitat descriptor class boundary	Class code	Equation type	Variable	Variable Threshold	Fuzzy X1	Fuzzy X0	Slope	Intercept	Probability threshold	Max probability
Wave High Lower Boundary	10	Fuzzy	Wave-induced kinetic energy ($N.m^{-2}$)	22	27	17	0.1	-1.7	0.5	1
Wave Moderate Upper Boundary	20	Fuzzy	Wave-induced kinetic energy ($N.m^{-2}$)	22	17	27	-0.1	2.7	0.5	1
Wave Moderate Lower Boundary	20	Fuzzy	Wave-induced kinetic energy ($N.m^{-2}$)	7.6	10.6	4.6	0.16666667	-0.76666667	0.5	1
Wave Low Upper Boundary	30	Fuzzy	Wave-induced kinetic energy ($N.m^{-2}$)	7.6	4.6	10.6	-0.16666667	1.76666667	0.5	1

3.2 – Exploring the input data

Open in a GIS software all the input raster files (i.e. *variable* rasters and *confidence in variable* rasters) listed in the csv configuration file and take a look at them individually (reminder: they are all in the 'input' folder).

3.3 – Running the script

- Open habitat_descriptor_modelling.R in Rstudio.

- Assign the script variable `workingDirectory` the path to the working directory (which is the folder that contains the 'config_files', 'input', and 'ouput '.

- Assign the script variable `config_csvFileName` the name of the configuration file, i.e. 'habitat_descriptor_modelling_wave.csv'

```
config_csvFileName<-"habitat_descriptor_modelling_wave.csv"
```

- In Rstudio, click the 'Source' button. The script starts to run. Depending on the capacity of your computer, it may take some time to complete. When it is finished, the Rstudio console will display a message saying "Completed in…" followed by the duration.

3.4 – Exploring the results

a) Check that all the rasters have been created. The 'output' folder should contain the following files:

Spatial distribution files	Confidence based on probability	Overall confidence
Wave_High.tif	Wave_High_confidence_based_on_proba.tif	Wave_High_confidence_overall.tif
Wave_Moderate.tif	Wave_Moderate_confidence_based_on_proba.tif	Wave_Moderate_confidence_overall.tif
Wave_Low.tif	Wave_Low_confidence_based_on_proba.tif	Wave_Low_confidence_overall.tif

b) Take some time to view these rasters in a GIS software. Check that the code assigned to each class spatial distribution raster cell matches the code given in the 'code' column of the csv configuration file.

4 – Merging all the habitat descriptor classes to create one single raster per habitat descriptor

The outputs of exercises 1 to 3 are individual spatial distribution raster datasets and confidence raster datasets for each class of the 3 habitat descriptors biological zones, current-induced energy levels and wave-induced energy levels. Here for the 3 habitat descriptors we will merge all the class raster datasets into a single raster dataset.

The script that does this is 'habitat_descriptor_merging_classes.R', documented in the guide, section 4.3.2.

4.1 – Examining the csv configuration file

- Open the csv configuration file (in the 'config_files' folder, file habitat_descriptor_merging_inputs.csv). This file describes all the elements required by the script to create the habitat descriptor rasters by merging their corresponding class rasters. The meaning of the columns in the file is documented in the guide, section 4.3.2.

We want to create 3 rasters for each habitat descriptor, i.e for

- The spatial distribution
- The confidence based on probability
- The overall confidence

We do not want to merge the continuous probability rasters. That's why the column 'proba_fileName' of the csv file is empty.

We need to give the script the names of the raster files created in exercises 1, 2 and 3 as input so that it can merge them. That's what the 'class_fileName', 'based_on_proba_confidence_fileName' and 'overall_confidence_fileName' columns are for.

Indeed, for each habitat descriptor (whose name is in the column 'habitat_descriptor_shortName'), the rasters created in exercises 1, 2 and 3 are listed in the columns 'class_fileName' (class spatial distribution rasters), 'based_on_proba_confidence_fileName' (rasters for confidence based on probability), and 'overall_confidence_fileName' (rasters for overall confidence). The 'folder' column contains the name of the folder containing all these files. As we created them all using the habitat_descriptor_modelling.R script, they are all in the 'output' folder.

4.2 – Running the script

- Open habitat_descriptor_merging_classes.R in Rstudio.

- Assign the script variable `workingDirectory` the path to the working directory (which is the folder that contains the 'config_files', 'input', and 'ouput ').

- Assign the script variable `config_csvFileName` the name of the configuration file, i.e. 'habitat_descriptor_merging_inputs.csv'

```
config_csvFileName<-"habitat_descriptor_merging_inputs.csv"
```

- Set the 4 boolean variables as follows:

```
output_habitat_descriptor_raster<-TRUE

output_overall_confidence<-TRUE

output_confidence_based_on_proba<-TRUE
```

```
output_probability_raster<-FALSE
```

- In Rstudio, click the 'Source' button. The script starts to run. Depending on the capacity of your computer, it may take some time to complete. When it is finished, the Rstudio console will display a message saying "Completed in…" followed by the duration.

4.3 – Exploring the results

a) Check that all the rasters have been created.

3 rasters have been created for each habitat descriptor listed in the csv file, one for the spatial distribution, one for the confidence based on probability, and a last one for the overall confidence that the modelled class is correct.

As a result, the 'output' folder should contain the following files:

Spatial distribution files	Confidence based on probability	Overall confidence
Biozones.tif	Biozones_confidence_based_on_proba.tif	Biozones_confidence_overall.tif
Currents.tif	Currents_confidence_based_on_proba.tif	Currents_confidence_overall.tif
Wave.tif	Wave_confidence_based_on_proba.tif	Wave_confidence_overall.tif

b) Take some time to view these rasters in a GIS software.

5 - combining the wave-induced and current-induced habitat descriptor layers

In exercises 1 to 4 we have created the raster layers for the habitat descriptors biological zones, wave-induced and current-induced energy levels. Before creating the habitat layer, we need to combine the latter two to create a single layer for the combined current/wave energy levels. The confidence raster layers are also combined. In addition to the full-coverage energy level confidence raster, a EUNIS 2007-11 specific confidence raster is created. In this raster only cells where the seabed substrate is rock and the biozone is infralittoral, shallow circalittoral or deep circalittoral are assigned a confidence value. This requires the input of a raster version of the seabed substrate (created by converting the seabed substrate polygon dataset to a raster using e.g. QGIS) and the biozone raster created earlier (see Section 1). The script 'habitat_descriptor_calculating_combined_energy.R' (see guide, section 4.3.3) performs these tasks.

5.1 – Running the script

- Open habitat_descriptor_calculating_combined_energy.R in Rstudio.

- Assign to the script variable `workingDirectory` the path to the working directory (which is the folder that contains the 'config_files', 'input', and 'ouput ').

- Assign as follows the R variables that indicate the path to the input rasters:

```
#path to the input raster file corresponding to the current-induced energy
current_fileName<-"output/Currents.tif"
#path to the input confidence raster file name corresponding to the current-induced energy
current_confidence_fileName<-"output/Currents_confidence_overall.tif"
#path to the input raster file corresponding to the current-induced energy
wave_fileName<-"output/Wave.tif"
#path to the input confidence raster file name corresponding to the current-induced energy
wave_confidence_fileName<-"output/Wave_confidence_overall.tif"
#path to the input raster file corresponding to the biozones. Will be used for the creation of the EUNIS2007-11 confidence raster
#(as EUNIS2007-11 only considers energy levels for rock substrate, only cells where the seabed substrate is rock are assigned a confidence value)
biozone_raster_fileName<-"output/Biozones.tif"
#path to the input raster file corresponding to the seabed substrate. Will be used for the creation of the EUNIS2007-11 confidence raster
#(as EUNIS2007-11 only considers energy levels for rock substrate, only cells where the biozone is infralittoral or circalittoral are assigned a confidence value)
substrate_raster_fileName<-"input/seabed_substrate.tif"
```

- Assign as follows the R variables that indicate the path to the output rasters:

```
#path to the output energy file
ouput_fileName<-"output/Energy.tif"
#path to the output confidence file
ouput_confidence_fileName<-"output/Energy_confidence_overall.tif"
#path  to the output EUNIS2007-11 confidence file
```

```
ouput_EUNIS_confidence_fileName<-"output/Energy_confidence_overall_EUNIS2007-
11.tif"
```

- Set the following R boolean variable to TRUE for the script to output the combined confidence raster layer as well.

```
calculateConfidence<-TRUE
```

- In Rstudio, click the 'Source' button. The script starts to run. Depending on the capacity of your computer, it may take some time to complete. When it is finished, the Rstudio console will display a message saying "Completed in…" followed by the duration.

5.2 – Exploring the results

a) Check that all the rasters have been created. There should be 3 files in the 'output' folder: Energy.tif and Energy_confidence_overall.tif, and Energy_confidence_overall_EUNIS2007-11.tif.

b) Take some time to view these rasters in a GIS software. Note in particular the difference between the full-coverage energy level confidence raster and the a EUNIS 2007-11 specific confidence raster. If you overlay the EUNIS 2007-11 specific confidence raster, the seabed substrate raster and the biozone raster you will note that in the EUNIS 2007-11 specific confidence raster there are values only where there is rock (i.e. seabed substrate = 70) and where the biozone is infralittoral, shallow circalittoral or deep circalittoral (i.e. biozone ≤ 30).

6 - Combining the habitat descriptor layers

Now it is time to create the habitat map. The first step is to combine the habitat descriptor layer 'seabed substrate', which is a polygon layer provided by EMODnet Geology, with the 2 habitat descriptor layers that we have created in exercises 1 to 5, i.e. the biological zones and the combined wave/current-induced energy levels. In order to preserve all the spatial detail of the seabed substrate layer, the combination is done in polygon mode. Therefore, the raster layers must be converted to polygons prior to combination. The combination is then performed by intersecting the layers.

Unfortunately, this is not done using a R script. The reason for this is that R packages have not proven to be efficient for either raster-to-polygon transformation or intersection of polygon layers when the number of polygons is large, which is usually the case for EUSeaMap. Therefore, ArcGIS™ is used here. A ModelBuilder model has been developed. It is documented in the guide, section 4.4.1.

6.1 – Examining the seabed substrate layer

In a GIS software, open the 'seabed_substrate.shp' shapefile that is in the input folder.

- Look at the attribute table. It contains many attributes, one of which is the code for the substrate types according to the EUSeaMap coding convention (see appendix 1).

- Create a symbology according to the attribute named 'EUSeaMap' to display the different substrate types in different colours.

6.2 – Creating a geodatabase

The performance of ArcGIS, especially when joining attribute tables (which will be done in the next section), is significantly better when geodabase feature classes are used instead of shapefiles.

Create a File Geodatabase in the output folder and name it *habitat_output.gdb*.

6.3 – Running the ArcGIS™ modelbuider model

The ModelBuilder model runs under ArcGIS™ 10.0 (or higher). It has also been successfully tested with ArcGIS Pro 3.3.

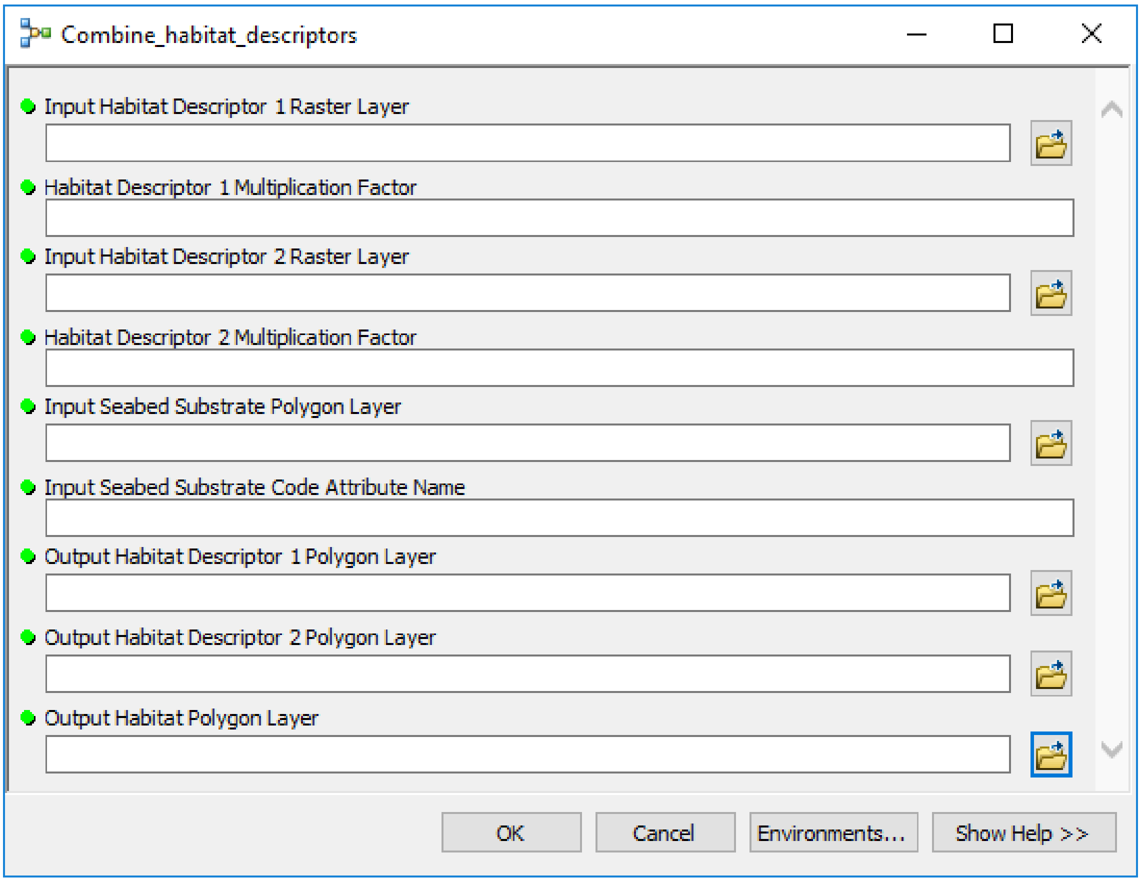
The ArcGIS™ modelbuider model:

1) converts the biological zone and energy raster layers into polygon layers

2) performs an intersection of the biological zone, energy raster and seabed substrate polygon layers

3) calculates a habitat code in a new column in the polygon layer resulting from the intersection. The calculation will be a sum of the biological zone, energy raster and seabed substrate codes. In order for the habitat code to follow the EUSeaMap coding convention (see guide, section 3.3.1), the biological zone code is multiplied by 100, the energy level code is multiplied by 10000, and the seabed substrate code is left as it is before the sum.

- In ArcGIS™, open (i.e. double-click) the 'Combine_habitat_descriptors' tool from the toolbox 'Toolbox_EUSeaMap_ArcGIS10.tbx'. The window shown below will open. The window entries are documented in the guide, section 4.4.1.



Input Habitat Descriptor 1 Raster Layer: select the file *Biozones.tif* (folder 'output')

Habitat Descriptor 1 Multiplication Factor: enter *100*

Input Habitat Descriptor 2 Raster Layer: select the file *energy.tif* (folder 'output')

Habitat Descriptor 2 Multiplication Factor: enter *10000*

Input Seabed Substrate Polygon Layer: select the file *seabed_substrate.shp* (folder 'input)

Input Seabed Substrate Code Attribute Name: enter *EUSeaMap*

Output Habitat Descriptor 1 Polygon Layer: we will create the feature class *biozones* in the *habitat_output.gdb* geodatabase

Output Habitat Descriptor 2 Polygon Layer: we will create the feature class *energy_levels* in the *habitat_output.gdb* geodatabase

Output Habitat Polygon Layer: we will create the feature class *habitats* in the *habitat_output.gdb* geodatabase

Click OK. It may take some time to generate the output, depending on the capacity of your computer.

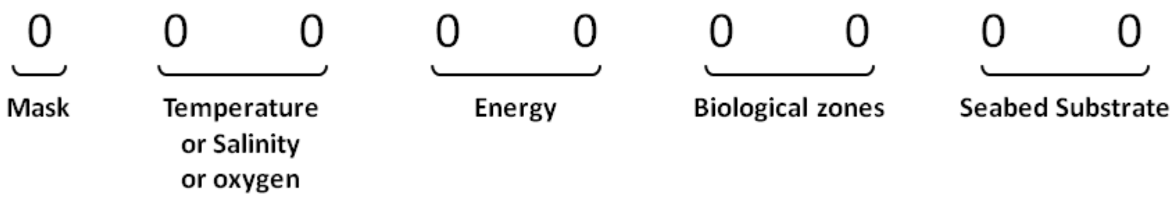
6.4 – Exploring the results

- In ArcGIS, open the feature class *habitats* that we have just created in the geodatabase. Each polygon corresponds to an intersection of a biological zone, an energy level and a seabed substrate.

- Open the attribute table. Look at the values in the attribute 'hab_code'

- What does the code 203070 mean in terms of biological zone, energy level and seabed substrate?

Reminder: the coding convention is as follows:



Answer: 203070 means 20 Energy=20 (i.e. 'Moderate'), Biological zone=30 (i.e. 'Deep circalittoral') and seabed substrate=70 (i.e. 'Rock)

- Make a symbology according to the attribute named 'hab_code'' to display the various habitat types in different colors.

6.5 – Subsidiary question

The input biological zone and energy level layers were in raster form. Only the seabed substrate layer was in polygon form. Why didn't we just rasterise the seabeb substrate layer and do the combination in raster mode rather than in polygon mode? It would have been much easier and faster!

Answer: doing the combination in raster mode would have indeed been much easier and faster, but it would have required converting the seabed substrate polygon layer to raster at a fixed resolution (the same resolution as the energy and the biological zone layers, i.e. approx. 100m), and as a result some spatial detail would have been lost.

7 - Joining the habitat polygon layer and the look-up table that crosswalks the modelled habitat codes and habitat classifications

The last step for creating the habitat map is to crosswalk the modeled habitat code with habitat types described in two habitat classications: EUNIS 2007-11 and the MSFD broad habitat types. This is performed by the script 'habitat_map_joining_LUT.R', documented in the guide, section 4.4.2.

It will require as input the *habitat* feature class created in the previous exercise and a look-up table, whose function is to crosswalk each modelled codes with a EUNIS class and a MSFD broad habitat type. Note that the look-up table is a simplified version of the table that is used for EUSeaMap, whose format is in Appendix 4. The output will be a new feature class, the columns of which will be the ones that are in the look-up table.

7.1 – Exploring the look-up table

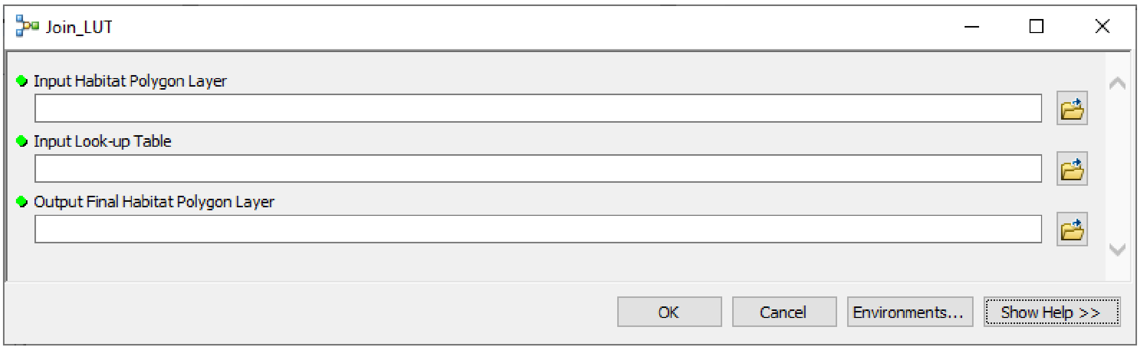
Open the file *LUT.dbf* that is in the folder 'input'. The column contains all possible codes in our study area. Then there are several other columns that crosswalk the code with several items, including the EUNIS 2007-11 class (column 'EUNISCombD') and the MSFD broad habitat types (column 'MSFD_BBHT').

7.2 – Running the ArcGIS™ modelbuider model

The ModelBuilder model runs under ArcGIS™ 10.0 (or higher). It has also been successfully tested with ArcGIS Pro 3.3.

The ArcGIS™ modelbuider model joins the look-up table that crosswalks the habitat layer created in the previous exercise and the look-up table. Additionnaly it runs a Multipart To Singlepart and a Repair Geometry.

- In ArcGIS™, open (i.e. double-click) the tool 'Join_LUT' from the toolbox 'Toolbox_EUSeaMap_ArcGIS10.tbx'. The window shown below will open. The window entries are documented in the guide, section 4.4.2.



Input Habitat Polygon Layer: select the feature class *habitats* (geodatabase *habitat_output.gdb* geodatabase)

Input Look-up Table: select the file *LUT.dbf* (folder 'input')

Output Final Habitat Polygon Layer: we will create the feature class *habitats_final* in the *habitat_output.gdb* geodatabase

Click OK. It may take some time to generate the output, depending on the capacity of your computer.

7.2 – Exploring the results

- In a GIS software, open the feature class *habitats_final*. It has the same polygons as the habitat layer created in the previous exercise. The attribute table, though, is quite different. Take a look at it.

- Make a symbology according to the attribute named "EUNIScombD" to display the various EUNIS habitat types in different colors.

- duplicate the layer in your software and with this layer make a symbology according to the attribute named "MSFD_BBHT" to display the various MSFD broad habitat types in different colors.

8 - Creating the habitat confidence layers

The final step is to create the habitat confidence raster layers. As the habitat layer is the result of the combination of habitat descriptor layers, the habitat confidence layer is the combination of the confidence in the habitat descriptor layers. Two habitat confidence layers are produced here because each habitat is proposed in two classifications, namely EUNIS 2007-11 and the MSFD broad habitat types, and these two classifications do not include the same habitat descriptors: EUNIS 2007-11 includes seabed substrates, biological zones and energy levels; the MSFD broad habitat types include only seabed substrates and biological zones . Therefore, the confidence raster layer for EUNIS 2007-11 will be the combination of the confidence layers for seabed substrates, biological zones and energy levels, while the confidence raster layer for the MSFD broad habitat types will be the combination of the confidence layers for seabed substrates and biological zones.

This is done using the 'habitat_map_calculating_confidence.R' script documented in the guide, section 4.4.3.

8.1 – Examining the csv configuration file

- Open the csv configuration file (in the folder 'config_files', file habitat_calculating_habitat_confidence.csv). This file describes all the elements required by the script to create confidence rasters by merging their corresponding habitat descriptor confidence. The meaning of the columns in the files is documented in the guide, section 4.4.3. Basically, the csv file describes the habitat descriptor confidence rasters to be combined. for each classification.

- Check that the names listed in the 'confidence_fileName' column are correct. The biological zone confidence raster is the one created in exercise 4. The energy confidence raster is the one created in exercise 5 for the EUNIS2007-11 habitat classification.

8.2 – Running the script

- Open habitat_map_calculating_confidence.R in Rstudio.

- Assign the script variable `workingDirectory` the path to the working directory (which is the folder that contains the 'config_files', 'input', and 'ouput ').

- Assign the script variable `config_csvFileName` the name of the configuration file, i.e. 'habitat_calculating_habitat_confidence.csv'

- In Rstudio, click the 'Source' button. The script starts to run. Depending on your computer capacity, it may take some time before it is finished. When it is finished, a message in the Rstudio console will say "Completed in…" followed by the duration.

8.3 – Exploring the results

- Check that all the raster have been created in the folder 'output'. There should be two of them, namely

- habitat_eunis2007_confidence_overall.tif for the EUNIS 2007-11 classification
- habitat_MSFD_BBHT_confidence_overall.tif for the MSFD benthic broad habitat types

- Take some time to look at these rasters in a GIS software. You can compare them. There are patterns in habitat_eunis2007_confidence_overall.tif that are not present in habitat_MSFD_BBHT_confidence_overall.tif. This is because the EUNIS 2007-11 classification takes energy into account, whereas the MSFD BBHT classification doesn't.

9 – Back to modelling biological zone classes - An example of a boundary with parameters that vary spatially

Download and unzip the data used for this exercise, available here:

https://cloud.ifremer.fr/index.php/s/1Ew2ITX2q9BAaM3

As mentioned at the end of section 3.1.2, the slope, intercept, and other parameters used to set a habitat descriptor class boundary can vary spatially. For these spatially-varying value parameters, a raster file is used in the configuration file instead of a constant value. In this exercise, we will model the biological zone classes in an area located at the entrance to the Baltic Sea. The particularity here is for the shallow circalittoral/deep circalittoral boundary, where bathymetry is used in some areas (i.e. in euhaline and polyhaline waters) and the probability of the halocline meeting the seabed is used in other areas (i.e. in mesohaline waters). For that boundary it is therefore necessary to input the script with a composite raster that has values for bathymetry where bathymetry is used and values for the probability of the halocline hitting the seabed where this variable is used. Similarly, the confidence in the variable raster is a composite of the confidence in the bathymetry (where bathymetry is used) and the confidence in the halocline hitting the seabed (where this variable is used). In addition, the slope and intercept also vary spatially, so the script is given a raster of slopes and a raster of intercepts instead of a constant value of slope and intercept. The parameters used are summarised in table 3.4.

Table 5.4: variable, classification method and parameters (slope, intercept, probability threshold and max probability) used as input to the script for each biological zone boundary. Other parameters (variable threshold, fuzzy X1 and fuzzy X0) are provided for background information

Habitat descriptor class boundary	Class code	Equation type	Variable	Variable Threshold	Fuzzy X1	Fuzzy X0	Slope	Intercept	Probability threshold	Max probability
Infralittoral Lower Boundary	10	Fuzzy	PAR at seabed ($mol.pho.m^{-2}.d^{-1}$)	0.07	0.1	0.04	16.6666667	-0.6666667	0.5	1
Shallow Circalittoral Upper Boundary	20	Fuzzy	PAR at seabed ($mol.pho.m^{-2}.d^{-1}$)	0.07	0.04	0.1	-16.6666667	1.6666667	0.5	1
Shallow Circalittoral Lower Boundary **In Euhaline and polyhaline waters**	20	Fuzzy	Bathymetry (m)	-30	-25	-35	0.1	3.5	0.5	1
Shallow Circalittoral Lower Boundary **In mesohaline waters**	20	Fuzzy	Probability of the halocline meeting the seabed	0.95	0.9	1	-10	10	0.5	1
Deep Circalittoral Upper Boundary **In Euhaline and polyhaline waters**	30	Fuzzy	Bathymetry (m)	-30	-35	-25	-0.1	-2.5	0.5	1
Deep Circalittoral Upper Boundary **In mesohaline waters**	30	Fuzzy	Probability of the halocline meeting the seabed	0.95	1	0.9	10	-9	0.5	1

9.1 – Examining the csv configuration file used for modelling the biological zone classes

- Open the csv configuration file (in the folder 'config_files', file habitat_descriptor_modelling_biozones.csv).

- Can you say what is the difference between what is filled in for the shallow circalittoral lower boundary slope / intercept and the deep circalittoral upper boundary slope / intercept and what was filled in for the boundary slope / intercept in the previous exercises?

Answer: tif file names are fiiled in instead of numerical values

9.2 – Exploring the input data

- Open in a GIS software all the input raster files that are used for the shallow circalittoral lower boundary and the deep circalittoral upper boundary, i.e.

    - for the former the files that are specified in the columns lower_boundary_variable, lower_boundary_confidence_in_variable, lower_boundary_slope, and lower_boundary_intercept,

    - and, for the latter the files that are specified in the columns upper_boundary_variable, upper_boundary_confidence_in_variable, upper_boundary_slope, and upper_boundary_intercept.

- For the 4 rasters used for the slope and intercept, use a symbology based on unique values. How many unique values are there in each of the 4 rasters?

Answer: 2 values. One for the cells where the bathymetry is used, one for the cells where the probability of the halocline meeting the seabed is used.

9.3 – creating the biological zone class rasters

- Open habitat_descriptor_modelling.R in Rstudio.

- Assign to the script variable `workingDirectory` the path to the working directory (which is the folder that contains the 'config_files', 'input', and 'ouput ').

- Assign to the script variable `config_csvFileName` the name of the configuration file, i.e. 'habitat_descriptor_modelling_biozones.csv'

```
config_csvFileName<-"habitat_descriptor_modelling_biozones.csv"
```

- In Rstudio, click the 'Source' button. The script starts to run. Depending on the capacity of your computer, it may take some time to complete. When it has finished, the Rstudio console will display a message mentioning "Completed in…" followed by the duration.

- Open the csv configuration file habitat_descriptor_merging_inputs.csv located in the folder 'config_files', check what is inside and close it.

- Open habitat_descriptor_merging_classes.R in Rstudio.

- Assign to the script variable `workingDirectory` the path to the working directory (which is the folder that contains the 'config_files', 'input', and 'ouput ').

- Assign to the script variable `config_csvFileName` the name of the configuration file, i.e. 'habitat_descriptor_merging_inputs.csv'

```
config_csvFileName<-"habitat_descriptor_merging_inputs.csv"
```

- Set the 4 boolean variables as follows:

```
output_habitat_descriptor_raster<-TRUE
output_overall_confidence<-TRUE
output_confidence_based_on_proba<-TRUE
output_probability_raster<-FALSE
```

- In Rstudio, click the 'Source' button. The script starts to run. Depending on the capacity of your computer, it may take some time to complete. When it is finished, the Rstudio console will display a message saying "Completed in…" followed by the duration.

9.4 – Exploring the results

- Check that all the rasters have been created, namely
    - Biozones.tif
    - Biozones_confidence_based_on_proba.tif
    - Biozones_confidence_overall.tif

In this exercise we have used rasters instead of constant values for the slope and the intercept. Rasters can be used as well instead of constant values for:

- The equation type

- The threshold

- The maximum probability

10 –Modelling biological zone classes - An example of externally defined boundary

The way in which externally defined boundaries are integrated into EUSeaMap is discussed in Appendix 3.

Download and unzip the data used for this exercise, available here:

https://cloud.ifremer.fr/index.php/s/Dw00xavKGvaIJyS

In this exercise, we will model the biological zone classes of the Caspian Sea. The particularity here is for the deep circalittoral/bathyal and bathyal/abyssal boundaries, which have been externally defined by experts. For these boundaries a distance X_1 of 0.04 decimal degrees is used, which corresponds to a fuzzy classifier slope of 12.5 and an intercept of 0.5. The parameters used are summarised in table 3.5.

Table 5.5: variable, classification method and parameters (slope, intercept, probability threshold and max probability) used as input to the script for each biological zone boundary. Other parameters (variable threshold, fuzzy X1 and fuzzy X0) are provided for background information

Habitat descriptor class boundary	Class code	Equation type	Variable	Variable Threshold	Fuzzy X1	Fuzzy X0	Slope	Intercept	Probability threshold	Max probability
Infralittoral Lower Boundary	10	Fuzzy	PAR at seabed ($mol.pho.m^{-2}.d^{-1}$)	0.15	0.2	0.1	10	-1	0.5	1
Shallow Circalittoral Upper Boundary	20	Fuzzy	PAR at seabed ($mol.pho.m^{-2}.d^{-1}$)	0.15	0.1	0.2	-10	2	0.5	1
Shallow Circalittoral Lower Boundary	20	Fuzzy	Bathymetry (m)	-50	-40	-60	0.05	3	0.5	1
Deep Circalittoral Upper Boundary	30	Fuzzy	Bathymetry (m)	-50	-60	-40	-0.05	-2	0.5	1
Deep Circalittoral Lower Boundary	30	Fuzzy for externally defined boundary	Distance to boundary (dd)	0	0.04	-	12.5	0.5	0.5	1
Bathyal Upper boundary	40	Fuzzy for externally defined boundary	Distance to boundary (dd)	0	0.04	-	12.5	0.5	0.5	1
Bathyal Lower boundary	40	Fuzzy for externally defined boundary	Distance to boundary (dd)	0	0.04	-	12.5	0.5	0.5	1
Abyssal Upper boundary	50	Fuzzy for externally defined boundary	Distance to boundary (dd)	0	0.04	-	12.5	0.5	0.5	1

10.1 – Exploring the input data

- Open in a GIS software the file bathyal_abyssal.shp. This shapefile contains the bathyal and abyssal polygons whose boundaries with respectively the deep circalittoral and the bathyal have been externally semi-automatically delineated by experts.

- From these boundaries, the required input rasters have been created. These are the following:

File	Description
dist_to_deepCirca_lowerbound.tif	Distance to the deep circalittoral lower boundary
dist_to_Bathyal_upperbound.tif	Distance to the bathyal upper boundary
dist_to_Bathyal_lowerbound.tif	Distance to the bathyal lower boundary
dist_to_abyssal_upperbound.tif	Distance to the abyssal upper boundary

Open these files in a GIS software. As a reminder, the inputs must strictly comply with these requirements:

- For a habitat descriptor class **upper** boundary, the raster cells that are within and below the class have a value of distance to the boundary, while raster cells that are above the class have a value of 0.
- For a habitat descriptor class **lower** boundary, the raster cells that are within and above the class have a value of distance to the boundary, while raster cells that are below the class have a value of 0.

Check the following:

File	Cells with distance value	Cells with value 0
dist_to_deepCirca_lowerbound.tif	Those of infralittoral, shallow circalittoral, and deep circalittoral	Those of bathyal and abyssal
dist_to_Bathyal_upperbound.tif	Those of bathyal and abyssal	Those of infralittoral, shallow circalittoral, and deep circalittoral
dist_to_Bathyal_lowerbound.tif	Those of infralittoral, shallow circalittoral, deep circalittoral and bathyal	Those of abyssal
dist_to_abyssal_upperbound.tif	Those of abyssal	Those of infralittoral, shallow circalittoral, deep circalittoral and bathyal

10.2 – Examining the csv configuration file used for modelling the biological zone classes

- Open the csv configuration file (in the folder 'config_files', file habitat_descriptor_modelling_biozones.csv). As you can see, for the deep circalittoral lower boundary, the bathyal upper and lower boundary and the abyssal upper boundary the value for the parameter upper_boundary_equation_type is 3 (reminder: 1=fuzzy, 2=GLM, 3=fuzzy for externally defined boundary).

- For these boundaries, the raster of confidence in the variable used is that of the confidence in the bathymetry. This is because the boundaries were defined based on changes in seabed slope, and seabed slope is a derivative of bathymetry.

10.3 – creating the biological zone classes

- Open habitat_descriptor_modelling.R in Rstudio.

- Assign to the script variable `workingDirectory` the path to the working directory (which is the folder that contains the 'config_files', 'input', and 'ouput ').

- Assign to the script variable `config_csvFileName` the name of the configuration file, i.e. 'habitat_descriptor_modelling_biozones.csv'

```
config_csvFileName<-"habitat_descriptor_modelling_biozones.csv"
```

- In Rstudio, click the 'Source' button. The script starts to run. Depending on the capacity of your computer, it may take some time to complete. When it has finished, the Rstudio console will display a message mentioning "Completed in…" followed by the duration.

10.4 - Exploring the results

Open the following output rasters:

- bathyal_proba.tif
- bathyal_confidence_based_on_proba.tif
- bathyal_confidence_overall.tif

Explore these results.

10.5 - Merging all the biological zone class rasters

- Open the csv configuration file habitat_descriptor_merging_inputs.csv located in the folder 'config_files', check what is inside and close it.

- Open habitat_descriptor_merging_classes.R in Rstudio.

- Assign to the script variable `workingDirectory` the path to the working directory (which is the folder that contains the 'config_files', 'input', and 'ouput ').

- Assign to the script variable `config_csvFileName` the name of the configuration file, i.e. 'habitat_descriptor_merging_inputs.csv'

```
config_csvFileName<-"habitat_descriptor_merging_inputs.csv"
```

- Set the 4 boolean variables as follows:

```
output_habitat_descriptor_raster<-TRUE

output_overall_confidence<-TRUE

output_confidence_based_on_proba<-TRUE

output_probability_raster<-FALSE
```

- In Rstudio, click the 'Source' button. The script starts to run. Depending on the capacity of your computer, it may take some time to complete. When it is finished, the Rstudio console will display a message saying "Completed in…" followed by the duration.

10.6 – Exploring the results

- Check that all the rasters have been created, namely
    - Biozones.tif
    - Biozones_confidence_based_on_proba.tif
    - Biozones_confidence_overall.tif

11 – Cleaning up a biological zone raster dataset

Download and unzip the data used for this exercise, available here:

https://cloud.ifremer.fr/index.php/s/fWi2qxabhQoWil2

In this exercise we will use the script that allows to clean up a habitat descriptor raster dataset. The script is documented in section 4.3.3. The 3 types of action will be used, i.e. 1) some cells will be removed, 2) some cell values will be replaced with other values, and 3) a sieve will be applied in an area to reduce noise.

11.1 – Exploring the input data

The inputs are 1) the raster dataset to be cleaned up (biozones.tif, which is in the 'output' folder) and 2) the polygon shapefile (cleanup_biozones.shp, which is in the 'input' folder) containing the areas where the cleaning is to be done and describing the type of action to be taken within each polygon. Open these files in a GIS software. Look at the attribute table. For 3 polygons, the value of the TYPE attribute is -1, which means that there the cell values will be given NA. For 2 polygons, the value of the TYPE attribute is 1, which means that within the polygons the cell values will be replaced with other values. In these 2 areas we have a situation where errors in the bathymetry have led to critical misclassification (Deep circalittoral instead of Shallow circalittoral). In both cases, the IN_VAL and OUT_VAL attributes indicate that within the polygons the cell value of cells with a value of 30 (i.e. Deep circalittoral) will be replaced with a value of 20 (i.e. Shallow circalittoral). For one polygon, the value of the TYPE attribute is 2, which means that within the polygon a sieve will be applied to reduce noise. The MAX_SIZE indicates that the threshold value of the sieve terra package function is 100.

11.2 – Cleaning up the biological zone raster

- Open habitat_descriptor_clean_up.R in Rstudio.

- Assign to the script variable `workingDirectory` the path to the working directory (which is the folder that contains the 'config_files', 'input', and 'ouput ').

- Assign to the following variables as

```
clean_up_shape_file<-"input/cleanup_biozones.shp"
input_raster_file<-"output/biozones.tif"
output_raster_file<-"output/biozones_corrected.tif"
```

- In Rstudio, click the 'Source' button. The script starts to run. Depending on the capacity of your computer, it may take some time to complete. When it is finished, the Rstudio console will display a message saying "Completed in…" followed by the duration.

11.3 – Exploring the results

Open the file biozones_corrected.tif in a GIS software. Compare with the file biozones.tif where there are polygons in cleanup_biozones.shp