EMODnet

European Marine Observation and Data Network

EMODnet Biology

EASME/EMFF/2016/006

Start date of the project: 19/04/2017 - (24 months)

EMODnet Phase III

Portfolio of modelling tools and products for European marine species and two examples of application of trait based approaches [D3.3 and D3.4]

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[1]

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the EASME or of the European Commission. Neither the EASME, nor the European Commission, guarantee the accuracy of the data included in this study. Neither the EASME, the European Commission nor any person acting on the EASME'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	Portfolio of modelling tools and products for European marine species and two examples of application of trait based approaches D4.3 Portfolio of modelling tools and products for European marine species D4.4 Two examples of application of trait based approaches
WP title	WP4 Data product creation
Task	Task 2: products constructed from one or more data sources (WP4)
Authors [affiliation]	Peter M.J. Herman [Deltares], Tom J. Webb[University of Sheffield], Charles Troupin [Univerity of Liège], Alexander Barth[Univerity of Liège], Olivier Beauchard [VLIZ], Dan Lear [MBA], Karline Soetaert [NIOZ], Lisa Sundqvist [SMHI], Lennert Schepers [VLIZ]
Dissemination level	Public

[1] The disclaimer is needed when the document is published

Contents

1 Introduction

In this report we give an overview of data products developed within EMODNET Biology, and clarify some novel modelling approaches that have been used to derive these products. We first discuss the general approach of selecting use cases based on Essential Ocean Variables. Subsequently we illustrate different products filling most of the categories mentioned. We then discuss the modelling approaches applied and end with a discussion of further developments and uses.

1.1 Sustaining GOOS Biological Essential Ocean Variables

GOOS (Global Ocean Observing System) has defined Essential Ocean Variables, a minimum set of information needed to monitor and manage the world's oceans. Within GOOS the Biology and Ecosystems Panel has identified a number of Variables on plants, animals, habitats and ecosystems that form the biological core of this system.

Currently, a set of EOVs for biology and ecosystems has been identified, and the process of validation, integration across disciplines and implementation of a mature, sustainable ocean observation programme is in progress. EMODNET Biology aims at contributing as much existing data as possible on European ecosystems to this programme. The availability of historic data is a prerequisite for validation of the approach, and moreover lends a historic perspective to future observation programmes.

The GOOS Biology and Ecosystem Essential Ocean Variables are illustrated in Figure 1. They concern six functional groups, ranging from microbes to marine mammals, and four habitat types of special importance. EMODNET data (in Biology and Habitats) cover most of these functional groups and habitats. Notable exceptions are microbes, that are poorly covered in the EMODNET databases. Very few data are available on turtles, corals and mangroves, because these are not frequently found in European waters.

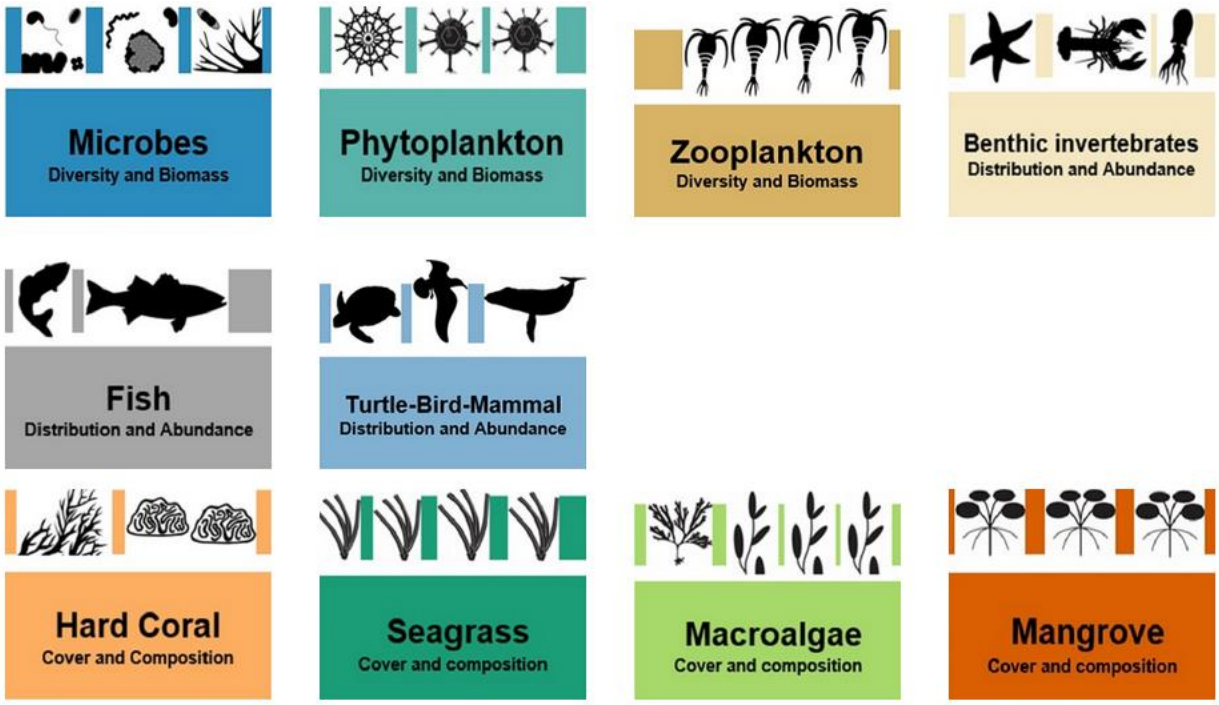


Figure 1. GOOS Essential Ocean Variables for Biology and Ecosystems

1.2 EOVs in the Atlas of European Marine Life

The EMODNET Biology Atlas of European Marine Life (http://www.emodnet.eu/launch-emodnet-atlas-european-marine-life) currently illustrates products on most of the topics covered by the EOVs. The Atlas of Marine Life conforms to the FAIR principles (Findable - Accessible - Interoperable - Reusable). All data are open and the methodologies used in the creation of the data products are shared in the EMODNET software repository (https://github.com/EMODnet). In this report we will illustrate some data products and the underlying models and methodologies, in order to document the current portfolio of applied methods. We refer for details to the description of the different products in the EMODNET portal for the Atlas (http://www.emodnet-biology.eu/about-atlas).

For *phytoplankton*, the EMODNET database contains few spatial datasets. A French coastal dataset aimed at the detection of harmful algal blooms is used as an example. Many long-term datasets are based on sustained observation at a single location. An example of the LTER site at Trieste, Italy, is used as an example. In this dataset we illustrate the use of a simple exploratory data model that increases insight in the main trends in the community. Many other long-term sustained observations are closed-data and as such not part of the EMODNET open database.

For *zooplankton*, we used a long-term single-location observatory in Villefrance, France as an example dataset. In this dataset, the main interpretation problem was the identification of false zeroes: species that were absent from the database but were not actively looked for during the observations. As the sampling and sorting effort varied over time, comparison of different series was needed in order to distinguish false from real zeroes. Other zooplankton datasets do have strong spatial (as well as temporal) components. We used the Continuous Plankton Recorder data of the North Atlantic as an example of a long-term dataset, with which we illustrated long-term large-scale changes in species distribution. We used a dataset on zooplankton from the Baltic Sea, composed of Swedish, Finnish, German and Polish datasets, to illustrate multi-year variability and, in addition, the use of environmental variables to better interpolate the species' distributions.

An extensive data collation of *macrozoobenthos* in the North Sea, North Atlantic and Baltic Sea was made. This dataset was used to illustrate the use of biological traits as a mechanism to summarize the composition of the community. Currently, a derived product is being compiled, aiming at the estimation of benthic bioturbation potential, a variable that is of high importance for the biogeochemistry of (especially shallow) seas.

Large and systematic databases of *fish* are available through ICES and form part of the EMODNET database. We illustrate the use of biological traits to demonstrate trends of different functional types of fish over the past decades in the North Atlantic.

A database on marine *turtles* in the Azores is used to illustrate EMODNET data on this group. In general, as is the case for birds and marine mammals, the most extensive datasets are curated by groups engaging many volunteers. Recent data are rarely available for inclusion into EMODNET, but co-operation with the curating groups is sought to include their results into the Atlas. Mostly older data on *birds* in the North Sea are used as a data product illustrating time evolution in a number of typical species. Data holdings on marine mammals in EMODNET are currently too sparse to be included as a product.

Data on habitats of special significance are curated by EMODNET Habitats. This group has co-operated in the production of the Atlas of Marine Life. We illustrate their results with maps of *seagrass* along European coasts.

1.3 Modelling tools used in the EMODNET Biology Products

A large part of the work in preparing EMOCNET Biology products is invested in cleaning the data sets. Even after taxonomic checks using WoRMS (http://www.marinespecies.org/) many taxonomic problems remain, as different observers are not always consistent in the level of taxonomic resolution or chose different higher taxa. Moreover, reconciling different sampling methodologies, different units for reporting, or changing taxonomic expertise in the course of a long-lasting time series, all cause breaks in the time series or spatial patterns that require close consideration before a product can be presented. We consider this work as an essential part of the EMODNET Biology task.

In preparing the products, we also had to recur to different forms of modelling in order to make useful products. The basic method for spatial dataset is gridding using DIVA (ref). Where spatial coverage of the data is sufficient, this is our method of preference, as it adds as little interpretation and modelling as possible to the output. We assume that EMODNET data products will be used by others in more sophisticated modelling applications, for the calculation of indicators or other derived statistics, and that the provision of well-controlled data is the essence of the EMODNET contribution to that work.

We did add several modelling components to the product, where this was needed to obtain interpretable products. A first example concerns the improvement of gridded maps of zooplankton in the Baltic. We used data on salinity and other environmental variables to interpolate between sparsely distributed sampling points across the Baltic. A second example concerns a first-stage interpretation of time series of phytoplankton (LTER site of Trieste), where we used multivariate analysis to illustrate how different species contributed towards the two dominant trends in the community, one related to seasonality and the other to long-term changes caused by eutrophication pressure. A third example concerns the modelling of thermal sensitivity of marine taxa. These approaches will be presented as part of the portfolio in section 3.

2 Portfolio of data products

2.1 Overview

An overview of EMODNET Biology data products is provided in Table 1, based on Lear et al. (subm.). Note that in this table a number of data products (e.g. maps of turtles and bird species) have been lumped in the section ‘abundance aps of various species or groups’. The overview shows that most products can serve as input to the management of diverse issues. This general-purpose nature of the EMODNET products is very important. It implies that the products have minimal interpretation in their generation and can be used in diverse application with dedicated methods. The fact that further processing may be needed for practical applications is one of the main reasons to make the data processing fully transparent, with full availability of the intermediate processing steps and data sets.

Full details of the data products and graphical representations are available from http://www.emodnet-biology.eu/about-atlas.

Lear et al. (subm.) discuss the integration level of the different products, as well as the audience and potential use of the products in full detail. This discussion will not be repeated here.

Table 1*:* Portfolio of EMODNET Biology data products. Based on Lear et al. (subm.)

Product	Supported EOV + Audience	Impact/Importance	Issues
Gridded abundance maps of various species/groups	All EOVs	Evolution and distribution of key species/species groups.	Marine Spatial planning, Fisheries, Habitat specifications
Neural network modelling of Baltic zooplankton abundances	Zooplankton abundance	Predicted gridded abundance maps of zooplankton in Baltic Sea, based on multiple sampling campaigns and using environmental information for abundance modelling	Marine Spatial planning, Eutrophication, Invasive Species
Distribution of fish living modes in European seas	Fish abundance and distribution	Distributions of the main functional types of benthic and bentho-pelagic fish species.	Marine Habitats specification Fisheries policies
Distribution of benthic macroinvertebrate living modes in European seas	Benthic invertebrate abundance and distribution	Distributions of the main functional types of benthic macroinvertebrates	Marine Habitats specification Fisheries policies, related to fishing pressure on benthos Eutrophication
Thermal affinities for European marine species	All EOVs	Shows temperature and vulnerability to temperature change of species.	Global change effects on marine ecosystems
OOPS - Copepods: ICES Operational Oceanographic Products and Services - Gridded Copepod abundance data	Zooplankton biomass and diversity	Zooplankton data of Continuous Plankton Recorder in N.Atlantic, used to illustrate temporal change in spatial patterns. ICES uses this product in their Operational Oceanographic Products and Services (OOPS)	Global change, regime shifts in marine ecosystems
Invasive marine species occurring in European marine harbours	All EOVs	Use EMODNET Biology and EUROBIS occurrence data of invasive species to check for false negatives in samples of invasive species in harbours	Basis for ballast water policy, Exemption policy from ballast water checks

Temporal trend of invasive species Marenzellaria in the Baltic Sea	Benthic invertebrate abundance and distribution	Temporal trend of invasive species	Invasive species
Phytoplankton community analysis in the Northern and Middle Adriatic	Phytoplankton community composition	Show temporal patterns in species and species groups, and interpret in terms of seasonality and long-term trend	Eutrophication Pollution
Temporal trend of algal toxicity along the French coast	Phytoplankton community composition	Show spatio-temporal distribution of toxic algae along French coast	Eutrophication Pollution
Long term zooplankton time series analysis from Villefranche, Western Mediterranean	Zooplankton biomass and diversity	Show temporal patterns in species and species groups, and interpret in terms of seasonality and long-term trend	Eutrophication, Pollution

2.2 Main data processing steps

Table 2 illustrates, for the different data products, what type of data sets the product was based on. It also specifies the necessary data clean-up and modelling steps that were needed to prepare the product. Data sets vary from well-prepared combined data sets that were curated by external parties, over compilation of individual researchers' data set into combined products, to the application of a 'big data' approach on the complete OBIS database.

Table 2*:* Data processing and modelling for the EMODNET Biology product portfolio

Product	Data base	Clean-up steps	Modelling
Gridded abundance maps of various species/groups	Compiled data bases curated by external parties (e.g. JNCC)	None	DIVA gridding
Neural network modelling of Baltic zooplankton abundances	National databases of Swedish, Finnish, Polish and German national data centers	Within and between data sets: taxonomic coherence, units of measurement, field methods	DIVA with neural network model for influence of external variables
Distribution of fish living modes in European seas	ICES database, curated by ICES	Elimination of species not properly sampled by the applied methods	Compilation of trait database. Grouping of species in

			functional groups. DIVA gridding
Distribution of benthic macroinvertebrate living modes in European seas	Data sets delivered by individual researchers or researcher groups	Taxonomic coherence. Elimination of data not sampled with compatible devices. Unit conversions.	Compilation of trait database. Grouping of species in functional groups. DIVA gridding
Thermal affinities for European marine species	OBIS occurrences database	None	Derivation of thermal affinities based on occurrence patterns. Calculation of sensitivities to temperature change
OOPS - Copepods: ICES Operational Oceanographic Products and Services - Gridded Copepod abundance data	Continuous Plankton Recorder	None	Grouping of data per season and time period. DIVA gridding. Basic time series analysis per species.
Invasive marine species occurring in European marine harbours	OBIS occurrences database	None, except selection of appropriate species and spatial extent	Qualitative comparison of OBIS results to harbour database.
Temporal trend of invasive species Marenzellaria in the Baltic Sea	Swedish, Polish, German, Finnish national data bases	Unit conversions.	Grouping of data per time period. DIVA gridding
Phytoplankton community analysis in the Northern and Middle Adriatic	Single data set of Long Term Ecological Research Site	Taxonomic coherence within a single long-term data set. Lumping of species to appropriate level of resolution	Multivariate time series analysis
Temporal trend of algal toxicity along the French coast	IFREMER data set, curated by IFREMER	Selection of HABs from phytoplankton database, using trait information in WoRMS	DIVA gridding
Long term zooplankton time series analysis from Villefranche, Western Mediterranean	Diverse datasets within a single institution	Taxonomic coherence. Lumping of species. Distinguishing between false and real zeroes. Combination of some data sets	None. Cleaned time series are shown.

3 Portfolio of data modelling approaches

In this section we describe the data modelling approaches that were applied in the current EMODNET Biology products. In principle, we have a much broader portfolio of methods, expertise and software available, but the aim of this section is not to review the current literature.

In selecting methods for application in EMODNET products, we have attempted to minimize the amount of interpretation and model-based filling of gaps in the observations. Wherever possible, we have used simple gridding of the available data to show in the first place what the available data tell us about fauna, flora and ecosystems. However, in a number of applications this was not sufficient. Either data were too sparsely distributed in space or time, or too complicated in taxonomic resolution to be easily interpreted. In those cases we have enriched the products with modelling approaches. An example is the Baltic zooplankton dataset, where sampling stations are unequally distributed over the Baltic Sea, and large areas are left for interpolation. Knowing that salinity is one of the major structuring factors for community composition in the Baltic, we found that using this information greatly improved the reliability of the gridded maps. We subsequently involved other environmental variables to test if further refinement was possible.

3.1 DIVA gridding

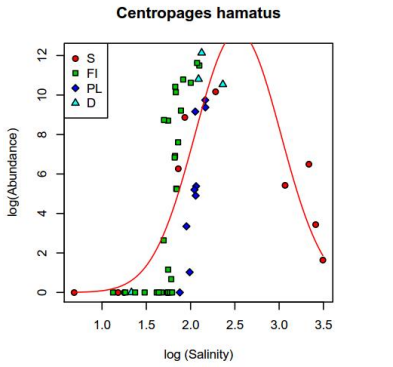
The preparation of gridded maps with the DIVA (Data-Interpolating Variational Analysis) software (https://github.com/gher-ulg/DIVAnd.jl/) is the *de facto* standard in the different EMODNET lots. In EMODNET Biology, we followed this standard. DIVA is extensively described in several publications and will not be detailed here. In short, DIVA is a method designed to efficiently interpolate in situ measurements onto a regular grid. In general terms, the variational inverse methods aim to derive a continuous field which is close to the observations (it should not necessarily pass through all observations because different types of errors affect them) and “smooth” (i.e. small first and second derivatives), as it has to represent climatological fields.

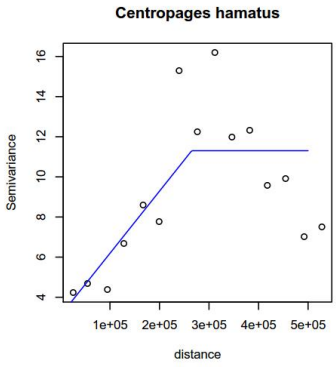
3.2 Kriging with dependence on a single environmental factor

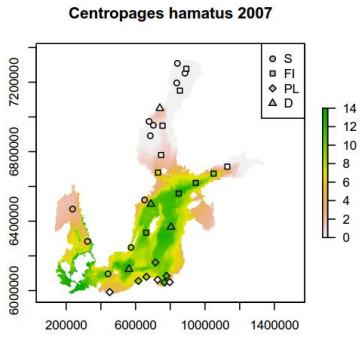
For the gridding of Baltic zooplankton a problem was posed by the application of DIVA as a gridding method. The spatial gaps between the observations are in some places relatively large. DIVA indicated large uncertainty of the interpolation values in these areas. That results, following standard procedures, in blanking the areas where the gridding is uncertain.

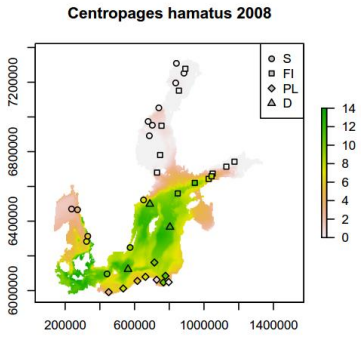
In a first modelling effort, we applied kriging with a linearizable co-factor. In this approach, first a non-linear, but linearizable response model of the abundance of the species on salinity was fitted by GLM. Subsequently, the residuals of the observation from this model were subjected to kriging interpolation and added to the response model. The resulting interpolation values take into account the spatial distribution of salinity, which is a master factor for the spatial distribution of species in the Baltic Sea. An example of this approach is shown in Figure 3. The response model is illustrated in the top-left graph. The fitted variogram is illustrated in the top right graph. The other pictures in Figure 3 illustrate the interpolated distribution of the species in subsequent years. For most species, we see relatively little variation in the spatial distribution from year to year, and in general a good correspondence between the interpolation and the observations. Note that the actual observation points were no interpolation points, so that deviation between observation and interpolation is, in principle, possible. However, the interpolation will be drawn to the observations. This may result in strongly varying patterns from year to year, even if the salinity field is relatively constant. Such fluctuations were, however, rarely seen.

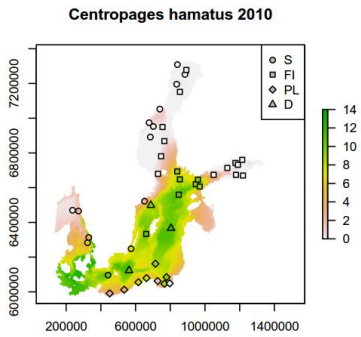
A disadvantage of this approach is that it does not use DIVA gridding and in that sense deviates from most EMODNET products. DIVA interpolation has a number of distinct advantages for application in areas such as the Baltic Sea, as it does not interpolate across islands and other restrictions for the water. Also, consistency between different products increases the comparability and usability of EMODNET products. We therefore developed an approach using DIVA, but taking into account external variables to guide the interpolation.

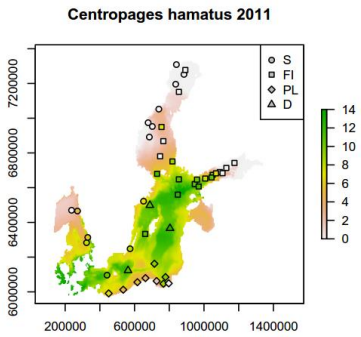


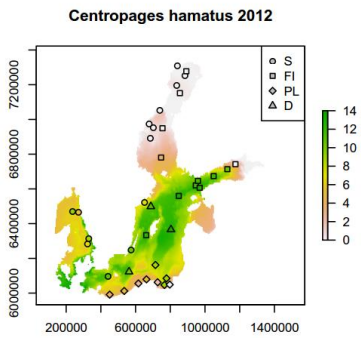












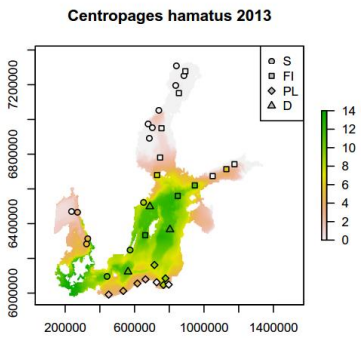


Figure 3. Interpolation of the abundance of Centropages hamatus *in different years in the Baltic. The top left figure shows the fitted response curve between abundance and salinity; the top right curve shows the variogram fitting. The other panels give the interpolated abundance for subsequent years. Observation points have a different shape dependent on the data source (four different countries) and are filled with the same colour scheme as the interpolation*

3.3 DIVA interpolation using information from environmental variables

DIVAnd is essentially a mono-variate reconstruction method (i.e., one variable is interpolated at a time). We describe here how the method can be extended to use other related variables (called the *covariables*) to help improve the interpolation.

The multivariate analysis of a variable x with a list of covariable z_1, z_2, … is expressed as follows:

$$x = x' + f(z_1, z_2, \dots, W_1, b_1, W_2, b_2, \dots)$$

where f is a non-linear function of the known covariables and unknown parameters W_1, b_1, W_2, b_2, …. The structure of the function f is given here by a neural network. The multilayer perceptron, a class of feedforward artificial neural network, is employed. Such a model consists of one input layer, one output layer, and at least one hidden layer.

The field x' is also unknown. Its spatial structure is constrained by DIVAnd.

3.3.1 Neural network

For every location j, the initial values of vector v^1 are the co-variables at the location j. This vector is linearly transformed by a weight matrix W_k and an offset vector b_k, and then a non-linear activation function (here rectified linear unit, ReLU) is applied to each element element of the resulting vector (except for the last step).

$$v_j^{(k+1)} = g^{(k+1)} + f^{(k)}(W_k v_j^{(k)} + b_k)$$

Here, the weights W_k and b_k do not depend on space directly, but the longitudes and latitudes are selected as covariables.

3.3.2 Experiments with synthetic observations

The test is prepared as follows (Figure 4):

1. Create a series of random field which are the "co-variables".
2. Create synthetic observations by combining these covariables to generate the "true field".
3. Sample these fields at random locations to get the "synthetic observations".
4. Perturb these co-variables as they are not perfectly known in practice.
5. Try to recover the true field from the synthetic observation using the imperfectly known covariables using the neural network.

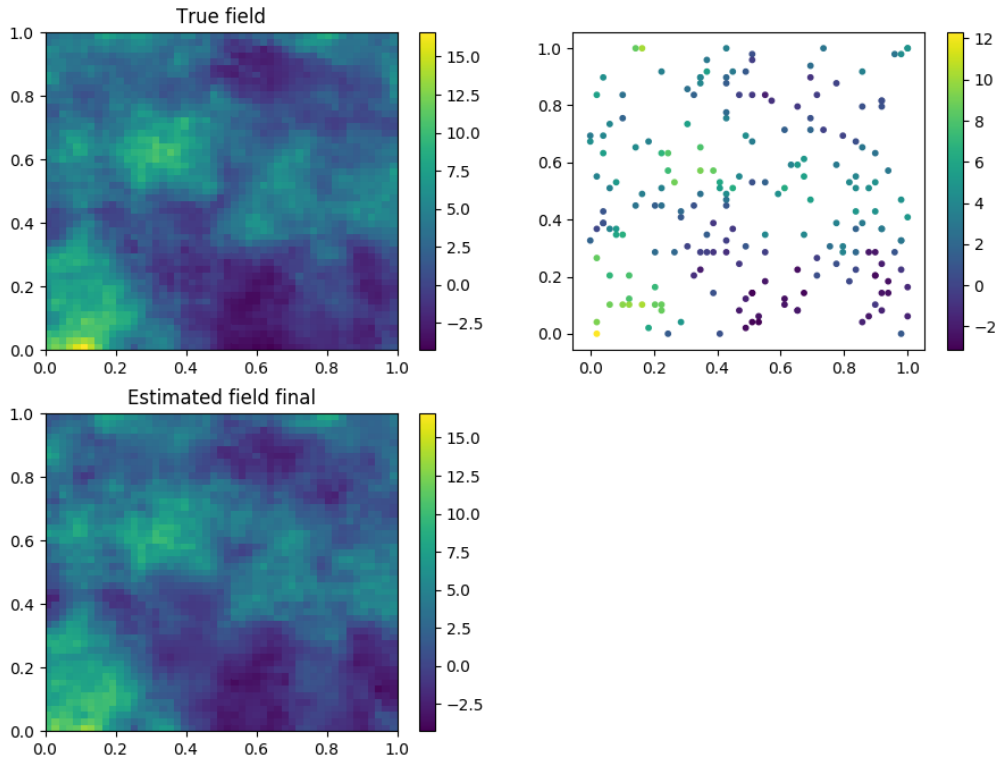


Figure 4. Illustration of the test with synthetic data. The True Field is calculated as a function of a number of environmental variables. The field is sampled at random locations, and the sampled values are perturbed. Subsequently the interpolation routing re-estimates the field, yielding the estimated field as shown in the lowermost picture. This estimated field corresponds closely to the true field above.

3.3.3 Logistic regression problem

We use the equations:
p(Model 1 | obs.) = 0.9 x 0.7 x (1 − 0.1) = 0.57
p(Model 2 | obs.) = 0.3 x 0.1 x (1 − 0.3) = 0.02

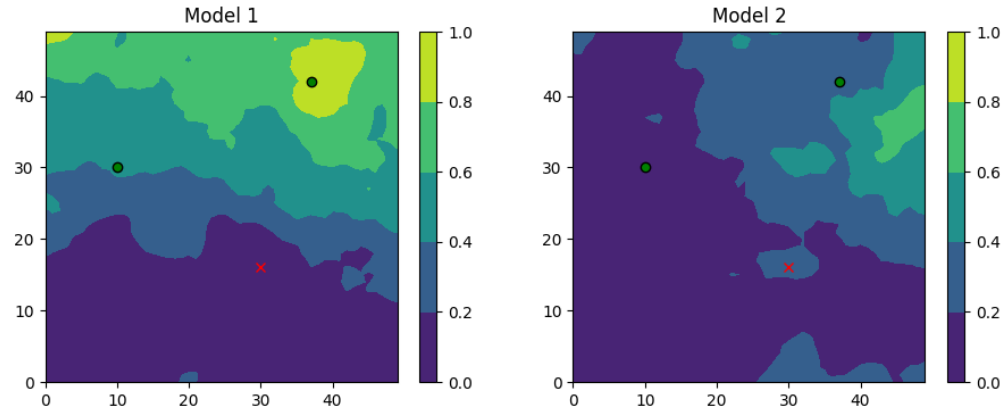


Figure 5. Illustration of the models used in the logistic regression problem

This leads to a similar problem as the previous case, except that here the true field represents a probability of occurrence. Here the synthetic observations are binary (occurrence or not). The cost-function to minimize is based on the negative log-likelihood (i.e. find the model which maximize the probability of the observations).

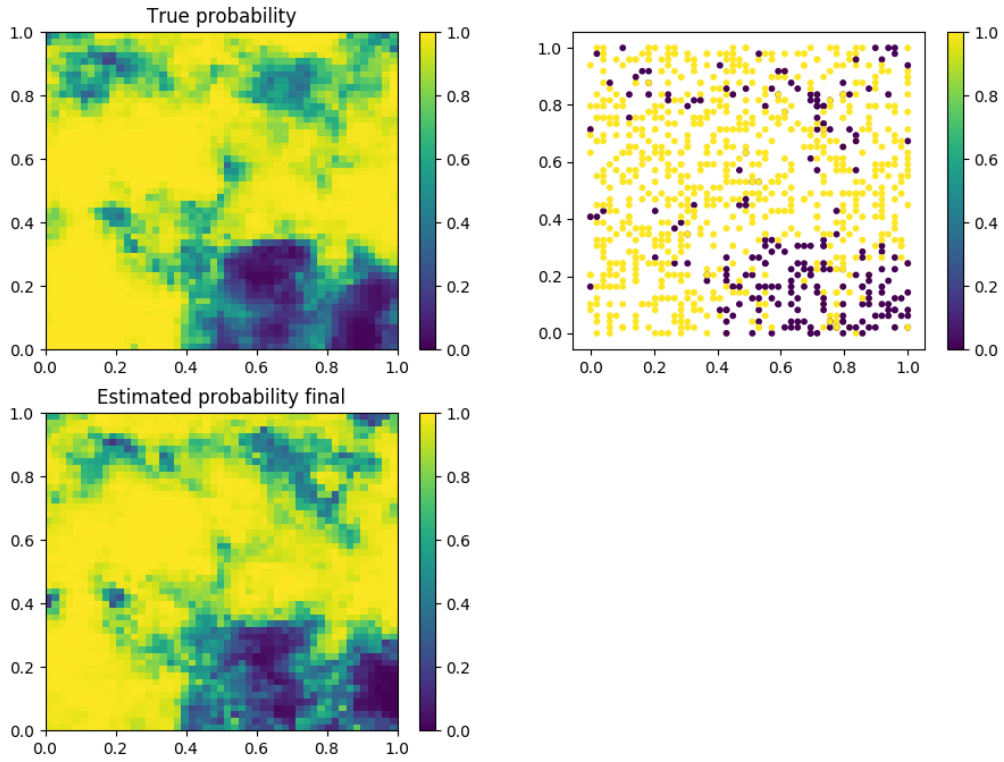


Figure 6. Logistic regression problem. The True field is expressed as a probability of occurrence. The samples are binary (present/absent) depending on the probability of occurrence. Based on the samples, the field is reconstructed as the estimated field.

3.3.4 Application to real data

A realistic application is developed with the objective to generate a gridded data product for 40 zooplankton species. The tools employed are DIVAnd for the spatial interpolation and Knet (in Julia) as the neural network library.

The main dataset consists of zooplankton observations in the Baltic Sea, obtained from four national data collections (Sweden, Finland, Germany and Poland), as used before in the kriging with co-variables. In contrast to the previous analysis, the neural network uses several co-variables as input:

- Dissolved oxygen concentration (from EMODnet Chemistry)
- Salinity (from SeaDataCloud)
- Temperature (from SeaDataCloud)
- Chlorophyll concentration (from MODIS Aqua satellite)
- Bathymetry (from GEBCO)
- the distance from coast (from GSFC, NASA)
- the position (latitude and longitude) and the year.

The fields represent the yearly average abundance. For every species the correlation length and signal to noise ratio are estimated using the spatial variability of the observations.

3.3.5 Results

The interpolated fields (here we show as an example *Acartia (Acanthacartia) bifilosa* for the year 2007) show good agreement with the observations and the cross-validation data points. The observations (inside white circles) are overlaid on the gridded field to allow a direct, visual comparison.

In addition, complex spatial dependencies could be learned from the covariables.

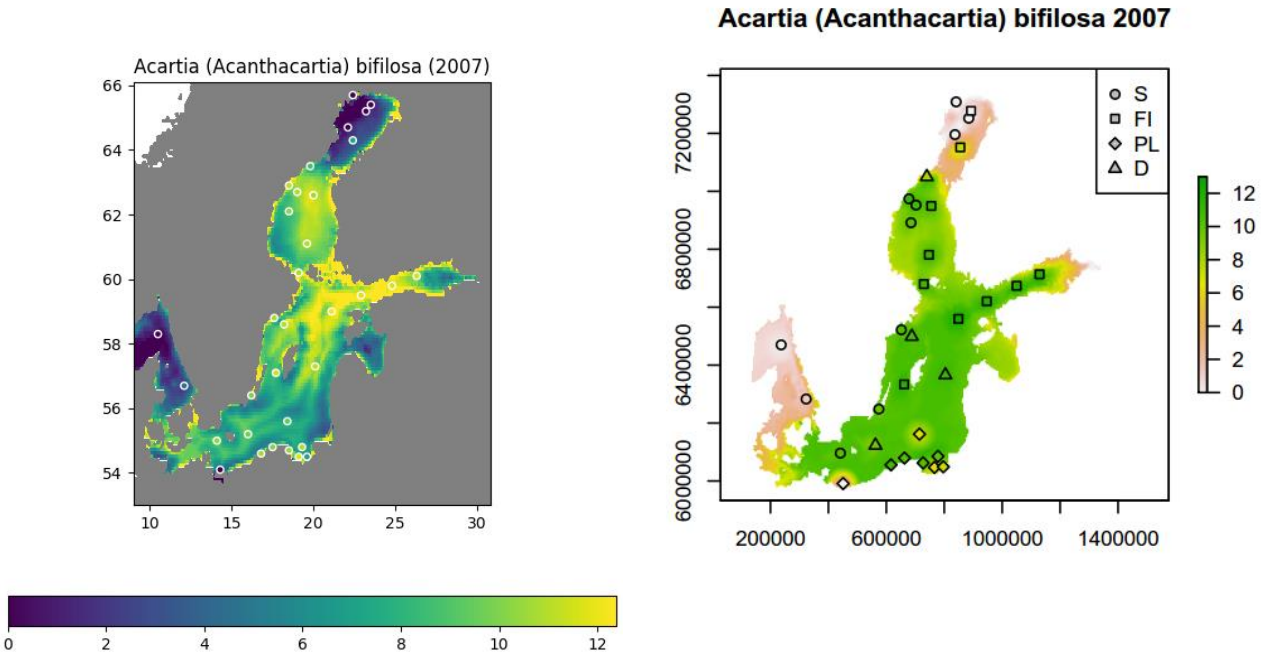


Figure 7. Left: the result of the field estimated using DIVA with NN. For comparison the field estimated with kriging using salinity as a co-variable is shown right. Note the difference in the colour scales used for the two figures.

For comparison, we show the result of the kriging with salinity as a co-variable for the same species and year in Figure 7. Please note the different color scheme in this figure. In comparison with the kriging result, the interpolated field of the NeuralNetwork – DIVA shows more differentiation, especially in the Baltic proper where the qualities of the water are followed closer. Many points have about equal correspondence with both interpolation methods, but sample points close to the German and Polish coast suggest that the features revealed by the DIVA interpolation coincide much better with the variations in the observations. This was generally observed in the results, adding credibility to the results of this interpolation.

We conclude that the Neural network can extract non-linear relationships useful to generated gridded data products. We present essentially a multivariate extension to DIV And where the dependency to other variables ("covariables") are estimated from the observations. Tests with synthetic data show that the underlying true field can be reconstructed from observations, even when the covariables are not perfectly known. The technique was also to abundance of 40 zooplankton species in the Baltic. The gridded dataset for all 40 species is available at http://www.emodnet-biology.eu/.

3.4 Summarizing temporal trends in multi-species time series data

Time series of plankton, such as the LTER series at Trieste, contain a large number of taxa, with a large diversity of trends in time. Presenting such time series in a compact but informative way is a major challenge. We developed an interactive R Shiny application for the time series, in which we show in one page the time evolution of the yearly variation in species abundance over the many years of the time series, and the seasonal pattern of occurrence, averaged over the years.

In a second page we summarize the information making use of a basic multivariate analysis. Based on a Principal Component Analysis (PCA) of the double-sqrt transformed abundances, we plot the many species as arrows in the biplot, to which we add in addition the centroids of the years and of the months. This picture summarizes the two main trends in the data: the long-term change in species composition, a process mostly determined by eutrophication status of these coastal waters, and the seasonal pattern of species succession. By selecting one taxon, the arrow of the taxon is highlighted in green and compared to all other arrows in grey. By selecting a group (e.g. all diatoms), all taxa belonging to the group are highlighted.

Figure 8 shows an example for the genus Cyclotella. The time series of yearly observations clearly shows that the abundance of the genus has been increasing over time, especially since 2009. The genus is mainly dominant in May-June.

In the multivariate plot, the arrow of the genus clearly points in the direction of the later years in the time series (time trend goes from right to left on the plot. The arrow is parallel to the x-axis, indicating that it occurs in the middle of the seasonal trend. The months are indicated in blue. Spring is on top, Autumn is below, summer is in the middle.

The interactive tool is available at http://www.emodnet-biology.eu/phytoplankton-community-analysis-northern-adriatic

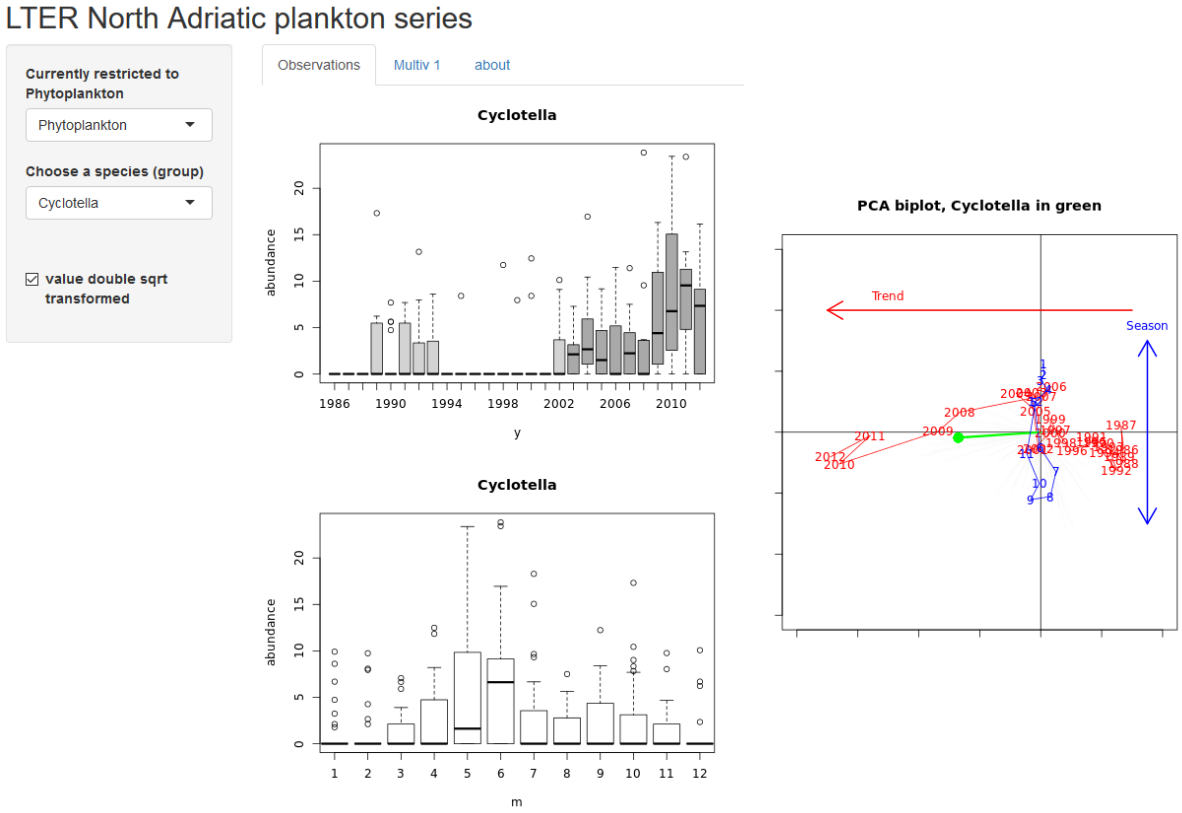


Figure 8. Example output of one taxon of the LTER Trieste dataset

4 The use of species traits to summarize large databases

Not all species are similar. They are characterized by a number of traits, e.g. size, age at maturity, number of offspring produced, type of locomotion, type of feeding, etc. These traits largely determine under which environmental conditions they will thrive. Species that mature early and produce a large number of rapidly dispersing offspring, as an example, are well adapted to temporary environments, e.g. patches that have become vacant after some disturbance. Their offspring will easily reach these patches, and develop fast to dominate the community in early stages. However, traits are correlated, and species cannot be good at everything. The rapidly dispersing, fast growing species will typically be poor competitors when resources are scarce, thus will be outcompeted by other species when the disturbed patch that they occupied early is maturing into a more stable community.

The main advantages of trait-based approaches are multifold. By abstracting from species identity, they facilitate comparison across biogeographic zones, as different species with similar traits may occupy similar niches in different zones. Second, many species traits are related to the type of temporal variability dominating their environment. As an example, in macrobenthos three distinct groups were found that are characterized by their differential response to stress patterns in their environment. The first one comprises species resistant to physical stress in natural conditions through a strong mobility, a short life cycle and a high offspring survival probability. The second one is composed of opportunistic species, also with a relatively short life span, among which many are pioneer species which are not very habitat-specific. The third group is composed of species which require many years or even decades to achieve a minimum of reproductive success. Because the trait-defined groups are characterized by their response to the temporal pattern of stress in their environment, they have potential use as indicators of human stress to

marine environments. Population survival in the two first groups was shown to be unaffected by bottom trawling, whereas the response of the third species group is generally negative. Hence, these products provide information on the benthic ecology, but also on ecosystem vulnerability to human pressures.

Application of trait-based analysis requires the compilation of life-history and living-mode characteristics of many hundreds of species. Although much information is available, often in very old references, the compilation is a long and painstaking effort. We compiled this information for many species of macrobenthos and fish. The coverage of European species is far from complete, but most dominant species in the North Atlantic realm are covered. As it is easier to find information on frequently occurring species, even a trait database that covers only a quarter of the species may easily cover 90% of the abundance. Extension to cover all species is impossible, as no life-history information is available in the literature on rare to very rare species.

Within the Atlas of European Marine Life, we have produced three trait-based products. For macrobenthos and fish we compiled trait information from the literature and applied it to distinguish different trait-based types of species. We produced maps of their distribution, as well as maps of the spatial distribution of the individual trait modalities. For a third product, thermal vulnerability of marine species, we adopted a different approach. We derived the thermal tolerance of many species based on their occurrence records in OBIS.

4.1 Benthic functional traits

The data product is based on 11 different datasets, that have been carefully merged into one database for macrobenthos in the North Atlantic European waters. The different occurrence lists were merged, and the validity of taxonomic names was checked with the World Register of Marine Species (WoRMS). Two measures were considered per geographic location: individual organism density and number of taxa, in two separated data frames.

Individual organism counts were summed per combination of location, sampling gear, sampling surface area, year, month and taxon. Sampling effort per location and sampling gear was calculated by summing the sampling surface areas of the different samples (when a location was sampled several times) and divided by the number of sampling times. This enabled the calculation of individual organism densities per sampling gear and per location, and finally per location (expressed in number of individuals per squared meters).

The number of taxa per location was calculated by successive averaging, firstly per combination of sampling gear, sampling surface area, year and month; then, per combination of sampling gear, sampling surface area and year, and so on until finally averaging per location.

Trait products are derived from the multiplication of the sample locations × taxa matrix by the taxa × trait modalities matrix through the community weighted mean procedure (CWM). For a given community and within a trait, the modality score is the percentage of species expressing the considered modality; all modality scores within that trait sum to one.

Finally, all the final outcomes were interpolated by the DIVA (Data-Interpolating Variational Analysis) tool to create gridded output maps.

Figure 9 shows the basic outcome of the analysis: maps of the prevalence of the three different types in the North-Atlantic European seas. Note that values in the Mediterranean are only based on a few French samples – outside the French coast the interpolation values are not reliable.

We see a high prevalence of the resistant type in the northern Baltic Sea. This species-poor community distorts the scale somewhat. Outside this area, the resistant type is mainly dominant in shallow, wave-swept areas, mostly coastal but also occurring on shallow areas such as the Dogger Bank. The Resilient type has a fairly homogeneous distribution, but is less abundant in deeper areas of the North Sea. Its

occurrence increases towards northern seas. The vulnerable type is most abundance in the deeper parts of the North Sea. It is rare or absent in the Baltic, and also not very dominant in northern seas.

Apart from these maps, the product also encompasses interpolated maps of the prevalence of all modalities of the different traits.

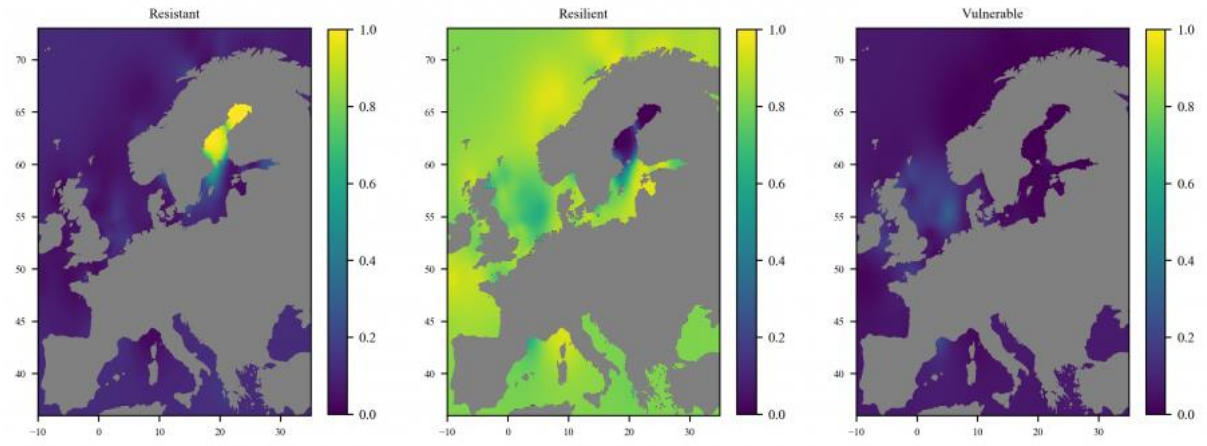


Figure 9. Distribution of the three main functional types of macrobenthos, based on DIVA interpolation and trait-based classification

4.2 Fish functional traits

This data product is based on the International Bottom Trawl Survey database maintained by ICES and part of EMODNET. Species occurrences were selected in order to maximize the spatial extent over the European northwest shelf. Therefore, data older than the year 2000 were discarded since species lists were not completely reported in all regional protocols before that year. Some areas (e.g. north Spain) are still missing as the species lists are still partially reported. In other areas (e.g. Baltic Sea, Baltic International Trawl Survey), lists were completely opened only after 2010.

The product displays the spatio-temporal distribution of four types of fish, based on a multivariate analysis of eight life history traits. Four main living modes were identified. The first and second groups comprise small species with a short life cycle. Species from group one differ from those from group two by a reduced fecundity compensated by either internal incubation or benthic embryonic development that increase juvenile survival, whereas species from group two are characterised by higher fecundity and juvenile mortality (unattended pelagic eggs). By contrast, groups three and four comprise large and long-lived species. Species from these groups also differ in fecundity, higher in group three, and juvenile survival, higher in group four. Group four is exclusively composed of elasmobranchs (rays, squates and sharks) which either internally incubate a few offspring that they release as adult-miniatures or release large eggs in strongly protective cases fixed on the sea floor. Although these data products provide information on fish species community ecology, they also provide indication on ecosystem vulnerability through the distribution of the two last species groups given the time that these species require to achieve their life cycle, including reproductive success and trophic control; besides, many of these large species have an important commercial value.

Average spatial group distributions are completed by sliding series of temporal windows restricted to three successive years in order to detect potential structural changes. Complementary maps display scores for each of the 37 trait modalities aggregated per spatial location.

Figure 10 shows an average over the entire sampling period of the occurrence of the four types.

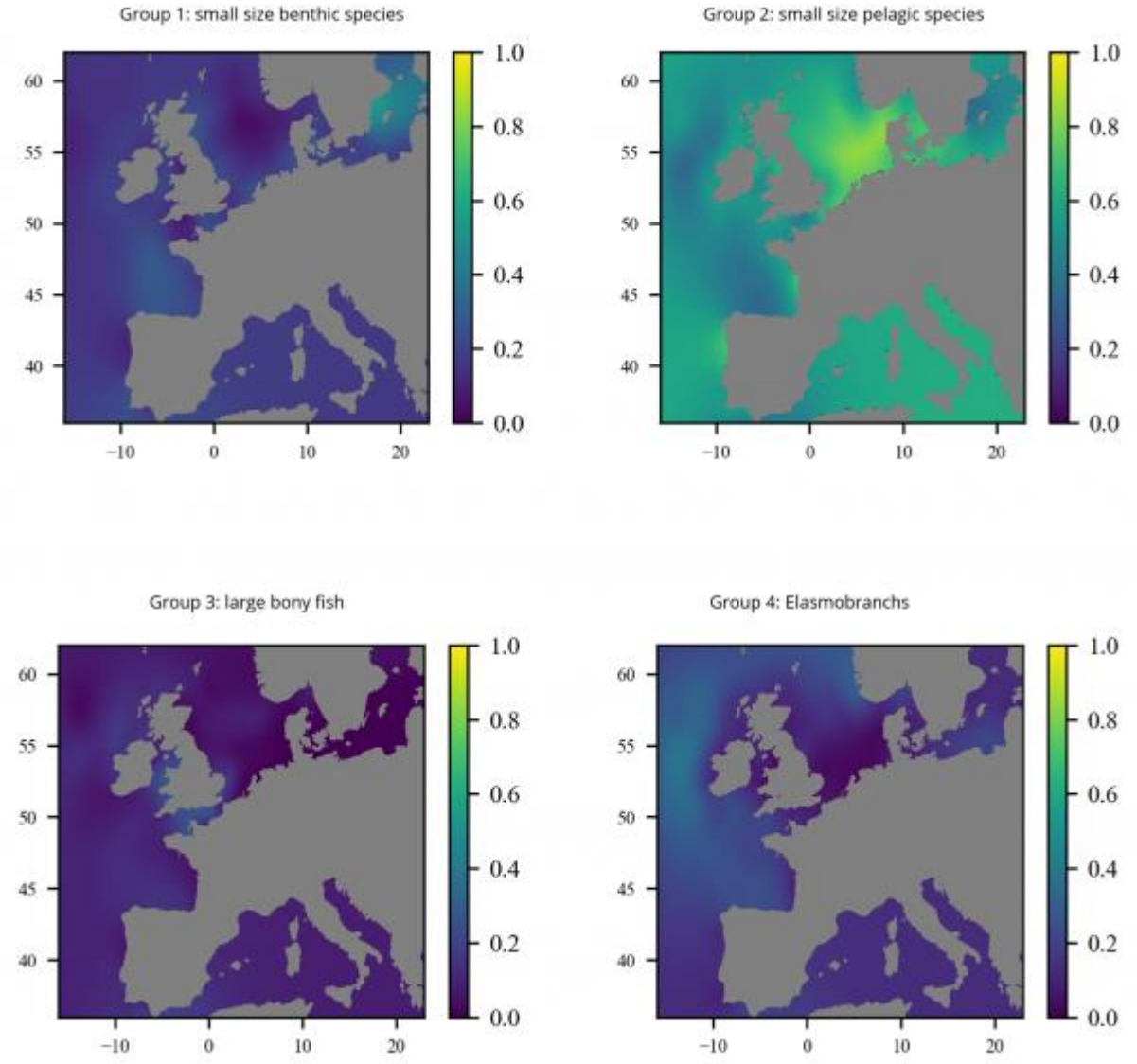


Figure 10. Occurrence of the different fish trait-based types in European waters. Note that data for the Mediterranean are very restricted and interpolations there are unsure.

4.3 Thermal affinities for European marine species

This product differs from the other trait-based products in that the traits of the species are derived from their spatial occurrence, rather than from external literature sources. Thermal affinities were derived for all European marine species, by matching occurrence records from OBIS to gridded temperature products. These species-level thermal affinities were then used to produce assemblage-level averages on a 0.5° grid covering European seas, separately for benthos, zooplankton, fish and other functional groups. Finally these gridded assemblage-level averages were compared to current and projected future sea temperatures to identify areas of high climate vulnerability for each functional group.

The product is a table with the thermal affinities of each species, which can be grouped by functional group to obtain 'community level thermal affinities'. With this information, we can compare the functional group-level temperature affinities to current and projected future temperature and create maps.

As an example, the map in Figure 11 shows the difference between mean zooplankton thermal affinity based on mean sea surface temperature (SST) and expected maximum SST in 2050:

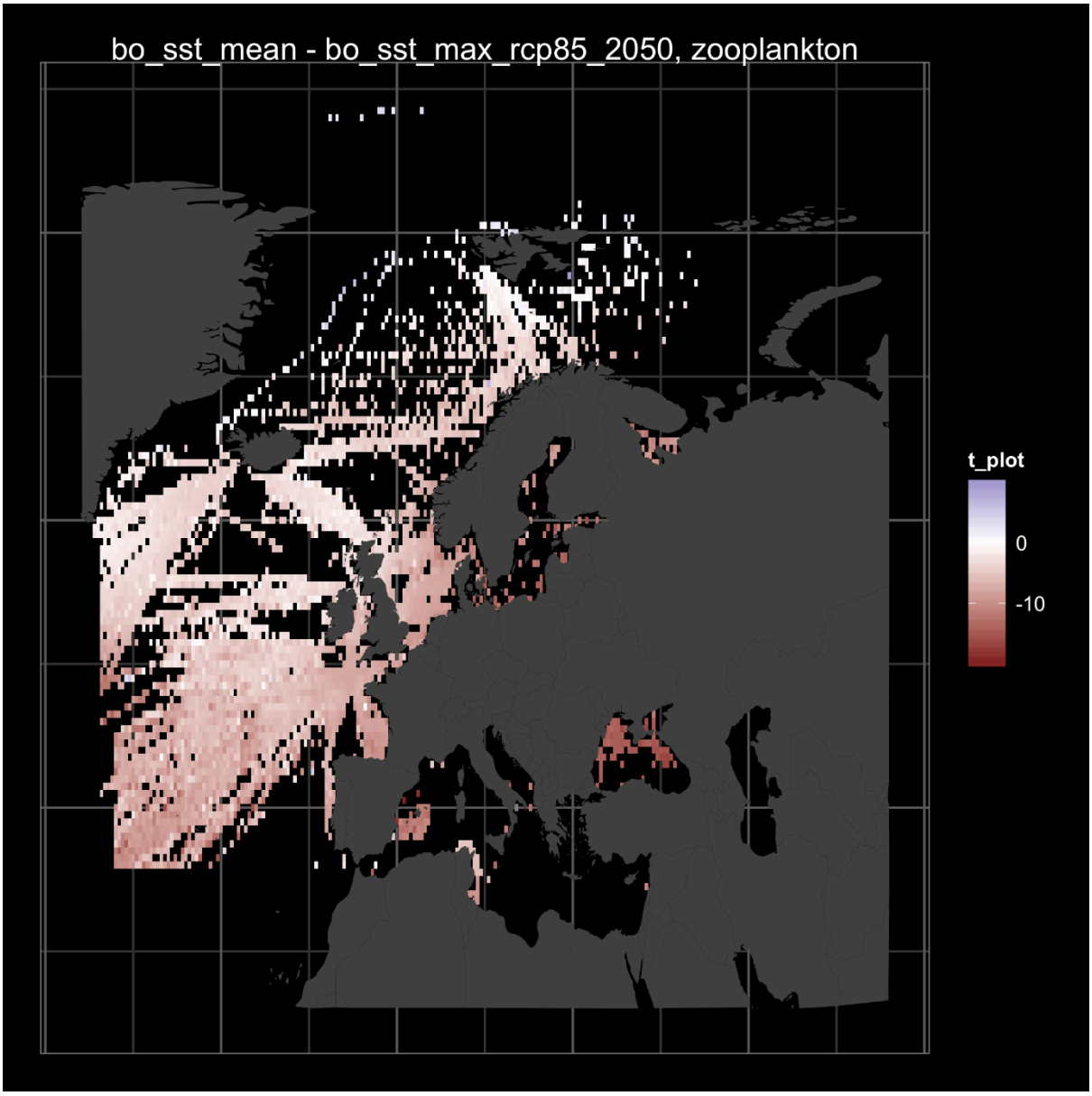


Figure 11. Map showing the difference between mean zooplankton thermal affinity based on mean sea surface temperature (SST) and expected maximum SST in 2050

5 Discussion and outlook

The current portfolio of data products in EMODNET Biology does not yet cover the entirety of the available datasets, but is comprised of examples for each of the essential Ocean Variables. Experience with the building of these data products has shown that major efforts are mainly required in data preparation. From taxonomic homogenization, over methodological bias corrections and detection of false versus real zeroes in datasets, this work requires a large amount of time and is extremely difficult to automate. Within EMODNET Biology a code base and experience have been developed and made publicly available through a dedicated github site. It is hoped that this will help to improve the productivity of this type of work in the future.

For a number of groups at higher trophic level (fish, mammals, birds, turtles) data sets are curated by other parties and transferred to EMODNET at a later stage. Although this causes a delay, it offers great advantages in terms of usability of the datasets for product preparation.

The unequal spreading of observations in the EMODNET data bases is a challenge for the preparation of data products on many groups. There is especially a lack of open data in the Mediterranean and Black Sea. But also in the better covered seas, monitoring and sampling are often very unequal or scarcely distributed in vast areas. For these cases, the use of basic modelling approaches can improve the quality of the interpolated fields. We have gained experience with different modelling approaches. The development of the neural network-based incorporation of environmental variables into the interpolation routine DIVA is especially promising in this respect. It remains consistent with the DIVA approach used throughout EMODNET for the production of gridded products, and at the same time can improve the quality of the interpolations considerably. The case on which the method was tested was especially rewarding, given the well-known importance of salinity in determining animal communities in the Black Sea. It remains to be tested whether this approach would also improve the interpolations for other groups in other areas, but available evidence suggests that this might well be the case. Given the compiled large dataset on macrobenthos in the North Atlantic seas, this is the dataset of choice to test further development of that approach. In particular, this approach may help in overcoming the scale gap in these interpolation maps. While benthos is know to be strongly determined by small-scale variations in physical forcing and sediment characteristics, brute interpolation of data at a coarse scale tends to overlook this aspect and result in rather poor predictions of local abundance.

The trait-based approach has required a major investment of effort in the compilation of life-history and living-mode characteristics of many species. Now that it is completed for benthic species and fish, it will be put to more use than the simple overview maps of functional groups presented in this report. An application has already been prepared by ICES using benthic traits to derive a sensitivity index to bottom trawl fisheries. We are also preparing products estimating the bioturbation potential of macrobenthos, which is of great importance for biogeochemical cycles in the sea. Other applications are in the deciphering of relations between functional characteristics and environmental factors. We envisage great opportunities in this respect, but will not develop these within the framework of EMODNET. Rather, it is hoped that the availability of the datasets will stimulate researchers to develop these possibilities.

Although we have been able to develop a number of products on plankton, we feel that the available data in EMODNET on this functional groups is still underexploited. The derivation of one or a few harmonized datasets, similar to what has been done for macrobenthos, may enhance the activities and lead to better space-covering products. This will be a priority for the future. As plankton datasets tend to have less spatial coverage than benthic datasets, the modelling methodology shown in this phase of EMODNET may prove very useful in the future in order to fill the spatial gaps in the data sets.

In conclusion, this portfolio of products and methodologies shows the progress in working up the large biological database in EMODNET. There is still a large amount of work to do before most relevant data will be worked up in products, but the approaches and methodologies are tested in practice and will be put to use in the coming phase of the project.