CHAPTER 4

BIO-ORACLE: A GLOBAL ENVIRONMENTAL DATASET FOR MARINE SPECIES DISTRIBUTION MODELING

Lennert Tyberghein[1], Heroen Verbruggen[1], Klaas Pauly[1], Charles Troupin[2], Frederic Mineur[3], Olivier De Clerck[1]

[1] Phycology Research Group, Biology Department, Ghent University, Krijgslaan 281 S8, 9000 Ghent, Belgium
[2] GeoHydrodynamics and Environment Research (B5a), Université de Liège, Allée du 6 Août 17, B4000 Liège, Belgium
[3] School of Biological Sciences, Queen's University Belfast, 97 Lisburn Road, Belfast BT9 7BL, UK

Article published as: Tyberghein L, Verbruggen H, Pauly K, Troupin C, Mineur F & De Clerck O (2011) Bio-ORACLE: a global environmental dataset for marine species distribution modeling. *Global Ecology and Biogeography*, DOI: 10.1111/j.1466-8238.2011.00656.x

ABSTRACT

Aim The oceans harbor a great diversity of organisms whose distribution and ecological preferences are often poorly understood. Species distribution modeling (SDM) could improve such knowledge and inform marine ecosystem management and conservation. Although marine environmental data are available from various sources, there are currently no user-friendly, high-resolution, global datasets designed for SDM applications. This study aims to fill this gap by assembling a comprehensive, uniform, high-resolution and readily usable package of global environmental rasters.

Location Global, marine

Methods We compiled global coverage data, e.g. satellite based and *in situ* measured data, representing various aspects of the marine environment relevant for species distributions. Rasters were assembled at a resolution of 5 arcmin (ca. 9.2 km) and a uniform landmask was applied. The utility of the dataset is evaluated by maximum entropy SDM of the invasive seaweed *Codium fragile* subsp. *fragile*.

Results We present Bio-ORACLE (Ocean Rasters for Analysis of CLimate and Environment), a global dataset consisting of 23 geophysical, biotic and climate rasters. This user-friendly data package for marine species distribution modeling is available for download at http://www.bio-oracle.ugent.be. The high predictive power of the distribution model of *Codium fragile* subsp. *fragile* clearly illustrates the potential of the data package for SDM of shallow-water marine organisms.

Main conclusions The availability of this global environmental data package has the potential to stimulate marine SDM. The high predictive success of the presence-only model of a notorious invasive seaweed shows that the information contained in Bio-ORACLE can be informative about marine distributions and permits building highly accurate species distribution models.

INTRODUCTION

During the last two decades, interest in predicting species distributions has grown substantially. Species distribution modeling has become an important tool in ecology, evolution, biogeography and conservation biology (Peterson, 2006; Graham *et al.*, 2004). Common applications include predicting the spread of invasive species (Thuiller *et al.*, 2005), forecasting impacts of climate change (Thomas *et al.*, 2004), inferring spatial patterns of species diversity (Graham *et al.*, 2006) and reconstructing ancestral niches (Martinez-Meyer *et al.*, 2004).

Considering the strong interest in species distribution models (SDM) and their wide application in terrestrial ecosystems, remarkably few studies infer SDM of marine species (Robinson *et al.*, in press). Notable exceptions where robust predictions of geographic distributions of marine fauna and flora were made include studies on fish (Maravelias & Reid, 1997; Wiley *et al.*, 2003; Guinotte *et al.*, 2006), cold-water corals (Davies *et al.*, 2008; Tittensor *et al.*, 2009), jellyfish (Bentlage *et al.*, 2009), crabs (Compton *et al.*, 2010) and seaweeds (Graham *et al.*, 2007; Verbruggen *et al.*, 2009). Although these examples illustrate the utility of SDM for marine studies, several issues have restricted the application of SDM in the marine realm compared to the terrestrial environment. One challenge is that the extensive spatiotemporal variability characterizing the oceans can hinder SDM (Valavanis *et al.*, 2008; Franklin, 2009). A second obstacle is the restricted availability of marine data (Kaschner *et al.*, 2006). SDM algorithms require both high quality species occurrence records and environmental information to infer the macroecological preferences of species (Elith & Leathwick, 2009). Gathering reliable species occurrence records is not straightforward as collecting can be impeded for highly mobile, circumglobal or deep-sea organisms (Kaschner *et al.*, 2006). Furthermore, global marine environmental data, although increasingly available on the internet, are often challenging to use with popular SDM applications. The available data often have coarse spatial resolution and suffer from missing data in coastal regions. Data are frequently provided in different file formats and spatial resolutions, making the assembly of a dataset one of the most time-consuming aspects of marine SDM. WorldClim (Hijmans *et al.*, 2005), a freely available set of global high-resolution climate layers, has served this purpose for the terrestrial SDM community for the past five years but marine species distribution modelers have not had access to a similar pre-packaged dataset

(Robinson *et al.*, in press). Although some environmental datasets exist, such as Aquamaps (Kaschner *et al.*, 2008a), the Hexacoral project environmental data (Fautin & Buddemeier, 2008) and a set of layers related to human impact on marine ecosystems (Halpern *et al.*, 2008), they have not been widely applied in SDM studies.

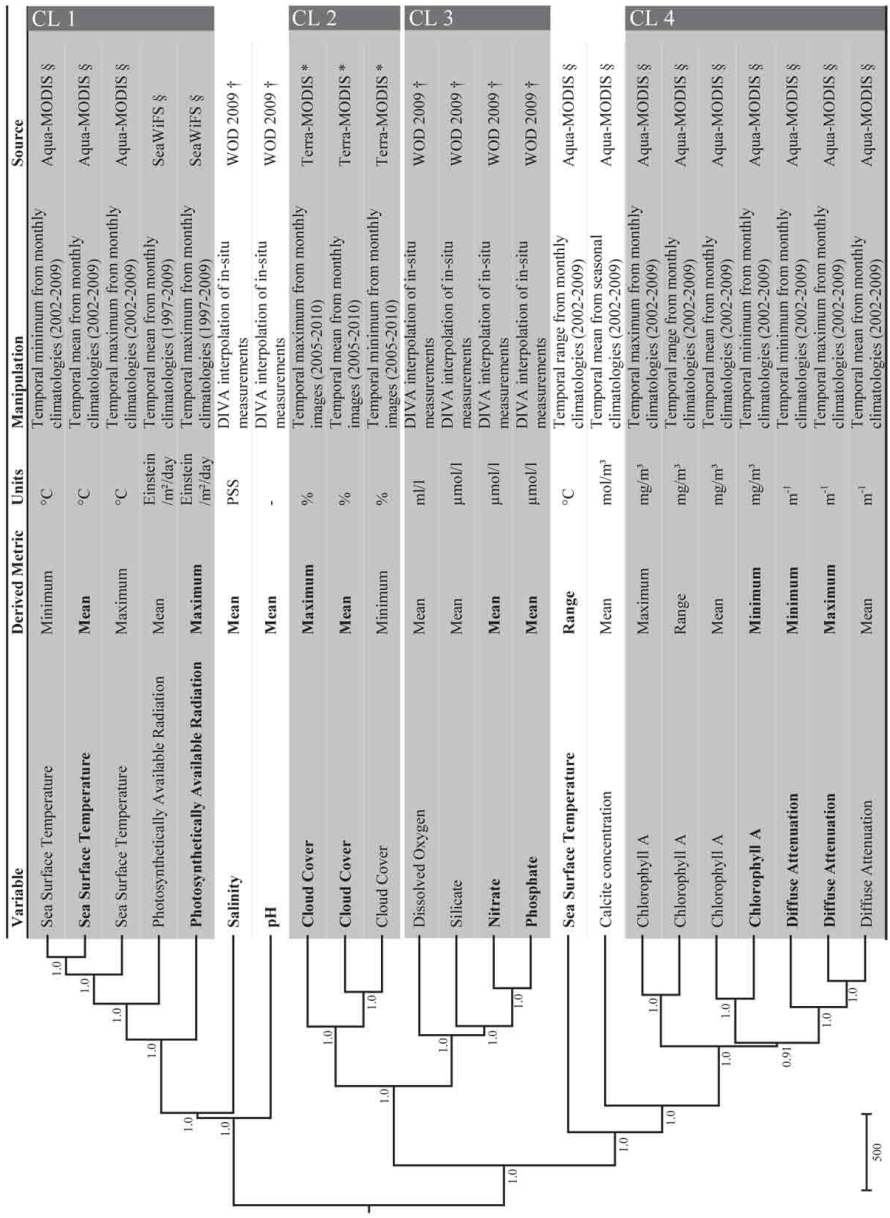
This study aims to facilitate marine species distribution modeling by assembling a comprehensive collection of global environmental rasters and supplying it ready for use in common species distribution modeling software. A broad set of macroecological variables representing environmental dimensions assumed to influence the distribution of marine shallow water organisms are packaged in the Bio-ORACLE database (Ocean Rasters for Analyses of CLimate and Environment). The utility of Bio-ORACLE for marine SDM is evaluated by modeling the distribution of *Codium fragile* subsp. *fragile*, a notorious invasive seaweed.

MATERIALS AND METHODS

We compiled preprocessed remotely sensed and in situ measured oceanographic data representing various quantitative environmental predictors of species distributions. In the first place, we looked for proximal variables, i.e. those with a recognized physiological or ecological relevance for marine organisms. Secondly, we included several other variables that may serve as proxies for species' environmental requirements.

REMOTELY SENSED DATA

Remotely sensed data were taken from various ocean observing satellite sensors (figure 1). We acquired monthly level-3 preprocessed satellite data (Aqua-MODIS and SeaWiFS; http://oceancolor.gsfc.nasa.gov/) at a 5 arcminute (ca. 9.2 km) spatial resolution. These geometrically corrected images are two-dimensional arrays with an equidistant cylindrical projection of the globe. Climatological composites, i.e. images summarizing information from the same month across several years, were used to generate three relevant metrics: annual maximum, minimum and mean. For sea surface temperature and chlorophyll A, the annual range (difference between maximum and minimum) was calculated as well. The latter metrics are biologically important as they represent proxies of seasonality and temporal variation in nutrient supply, respectively.



Variable	Derived Metric	Units	Manipulation	Source
Sea Surface Temperature	Minimum	°C	Temporal minimum from monthly climatologies (2002-2009)	Aqua-MODIS §
Sea Surface Temperature	**Mean**	°C	Temporal mean from monthly climatologies (2002-2009)	Aqua-MODIS §
Sea Surface Temperature	Maximum	°C	Temporal maximum from monthly climatologies (2002-2009)	Aqua-MODIS §
Photosynthetically Available Radiation	Mean	Einstein /m²/day	Temporal mean from monthly climatologies (1997-2009)	SeaWiFS §
Photosynthetically Available Radiation	**Maximum**	Einstein /m²/day	Temporal maximum from monthly climatologies (1997-2009)	SeaWiFS §
Salinity	**Mean**	PSS	DIVA interpolation of in-situ measurements	WOD 2009 †
pH	**Mean**	-	DIVA interpolation of in-situ measurements	WOD 2009 †
Cloud Cover	**Maximum**	%	Temporal maximum from monthly images (2005-2010)	Terra-MODIS *
Cloud Cover	**Mean**	%	Temporal mean from monthly images (2005-2010)	Terra-MODIS *
Cloud Cover	Minimum	%	Temporal minimum from monthly images (2005-2010)	Terra-MODIS *
Dissolved Oxygen	Mean	ml/l	DIVA interpolation of in-situ measurements	WOD 2009 †
Silicate	Mean	µmol/l	DIVA interpolation of in-situ measurements	WOD 2009 †
Nitrate	**Mean**	µmol/l	DIVA interpolation of in-situ measurements	WOD 2009 †
Phosphate	**Mean**	µmol/l	DIVA interpolation of in-situ measurements	WOD 2009 †
Sea Surface Temperature	**Range**	°C	Temporal range from monthly climatologies (2002-2009)	Aqua-MODIS §
Calcite concentration	Mean	mol/m³	Temporal mean from seasonal climatologies (2002-2009)	Aqua-MODIS §
Chlorophyll A	Maximum	mg/m³	Temporal maximum from monthly climatologies (2002-2009)	Aqua-MODIS §
Chlorophyll A	Range	mg/m³	Temporal range from monthly climatologies (2002-2009)	Aqua-MODIS §
Chlorophyll A	Mean	mg/m³	Temporal mean from monthly climatologies (2002-2009)	Aqua-MODIS §
Chlorophyll A	**Minimum**	mg/m³	Temporal minimum from monthly climatologies (2002-2009)	Aqua-MODIS §
Diffuse Attenuation	**Minimum**	m^{-1}	Temporal minimum from monthly climatologies (2002-2009)	Aqua-MODIS §
Diffuse Attenuation	**Maximum**	m^{-1}	Temporal maximum from monthly climatologies (2002-2009)	Aqua-MODIS §
Diffuse Attenuation	Mean	m^{-1}	Temporal mean from monthly climatologies (2002-2009)	Aqua-MODIS §

Figure 1. Dendrogram depicting agglomerative clustering of Bio-ORACLE data layers. Numbers at nodes indicate statistical support (bootstrap probabilities). Numbers range from 0 (no support) to 1 (maximal support). The table includes Bio-ORACLE variables, derived metrics, units, manipulation and source. Variables used to build the *Codium fragile* SDM are depicted in boldface. Grey shaded areas represent distinct clusters. Legend: † (Boyer *et al.*, 2009); § (Feldman & Mcclain, 2010); * (Nasa, 2010).

We also included the Terra-MODIS-derived cloud fraction data (http://modis-atmos.gsfc.nasa.gov/) available in monthly composites at a resolution of 6 arcminutes. The monthly data over 10 years (2000-2009) were used to create average monthly composites and then further processed to produce the three standard metrics (annual mean, maximum and minimum). Eventually these layers were resampled to a resolution of 5 arcminutes using bilinear interpolation.

IN SITU MEASURED OCEANOGRAPHIC DATA

In addition to remotely sensed data, spatially interpolated data layers were developed from oceanographic in-situ surface measurements gathered from the World Ocean Database 2009 (WOD09) (Boyer *et al.*, 2009). We rejected all data flagged as erroneous in the WOD09 (Johnson *et al.*, 2009).

Different statistical approaches have been used to generate interpolated environmental surfaces (Daly, 2006). We have used DIVA (Data Interpolating Variational Analysis), a method developed for gridding in situ data using the variational inverse method (Brasseur & Haus, 1991) that has previously been applied to temperature, salinity and phosphate records of the Mediterranean Sea and the North-East Atlantic (Brasseur *et al.*, 1996; Karafistan *et al.*, 2002; Troupin *et al.*, 2010). Compared to other interpolation techniques, DIVA has a number of features that makes it very attractive for our application, including: (i) the ability to work with large amounts of data without intermediate averaging; (ii) the consideration of coastlines and topography; (iii) the generation of coherent error maps useful for identifying regions of uncertainty in the resulting environmental rasters; (iv) the use of a limited amount of parameters estimated in an objective way (Troupin *et al.*, 2010).

Coastlines were extracted from the global GEBCO one arcminute bathymetry (November 2009; http://www.gebco.net/). In order to perform DIVA analyses, two parameters have to be determined: the correlation length L and the signal to noise ratio λ. The software provides tools to estimate these parameters from the data. The correlation length (L) was estimated by fitting the correlation between the data and a theoretical kernel function. As suggested by Troupin *et al.* (2010) and based on sensitivity, the signal-to-noise ratio (λ) was assigned the constant value of one. The variance of the background field was also assigned a value of one to have DIVA generate relative error fields.

As a means of quality control, the outlier detection option was activated and error maps were calculated. Outlier detection is based on a comparison between the data-analysis residual and the expected standard deviation. The removal of data outside the realistic range was not necessary as this part of the quality control was carried out by the World Ocean Data center. The error estimates reflect the confidence in interpolated surfaces depending both on the data coverage and on their quality. Continuity across the 180° meridian was achieved by running two DIVA analyses, one ranging from -180° to 180° and a second starting and stopping at the 0° meridian (i.e. crossing the 180° meridian). The two resulting rasters were weight-smoothed into one final raster, with pixel weights decreasing linearly from the center of the input images to their sides.

PREPROCESSING AND MULTIVARIATE ANALYSIS OF BIO-ORACLE RASTERS

In order to arrive at a ready-for-use data package for SDM applications, a uniform landmask was applied to all data layers. This procedure consisted of correcting discrepancies between environmental data and the coastline, masking data pixels that were on land and calculating values for marine pixels without data by cubic extrapolation. Finally, the Polar Regions, which suffer from missing and imprecise data, were excluded by cropping the Bio-ORACLE rasters to latitudes between 70° N/S.

To better grasp the major environmental dimensions present in Bio-ORACLE, we explored the dataset with multivariate statistics. After the variables had been standardized, they were subjected to hierarchical clustering using 'hclust' of the 'stats' library in R (http://www.r-project.org/). This clustering technique was performed on a matrix of all pixel values in all rasters (7,257,600 pixels × 23 variables). We used the Euclidean distances and the average distance agglomeration method. Confidence in the hierarchical clustering was assessed by multiscale bootstrapping (Shimodaira, 2004) using the R package 'pvclust' (Suzuki & Shimodaira, 2006). We used 100 bootstrap replicates and the default relative sample sizes of bootstrap replications.

Environmental maps for SDM are usually provided in equidistant projections (north-south distances neither stretched nor compressed), but such maps may bias distribution models of species that span a large latitudinal range

(Tittensor *et al.*, 2009). For this reason and for applications like species richness analysis, equal-area cells may be preferred. To accommodate this, we remapped all Bio-ORACLE layers onto a Behrmann equal-area projection using ArcGIS 9.2 (http://www.esri.com). Coordinate transformation formulas are provided in Appendix S3. Both these equal-area and the original equidistant projections are provided to the public.

CASE STUDY: *CODIUM FRAGILE* SUBSP. *FRAGILE*

The ability of Bio-ORACLE to predict the distribution of marine species was evaluated by inferring a presence-only species distribution model of the invasive species *Codium fragile* subsp. *fragile* (hereafter *C. fragile*). This green alga is considered to be native in the NW-Pacific (i.e. Japan and surrounding areas) and has been introduced to Europe, the west and east coasts of North America, Australasia, South Africa and South America (Trowbridge, 1998; Provan *et al.*, 2008). Being a shallow-water marine organism, this species is a suitable candidate to evaluate the utility of Bio-ORACLE.

We collected occurrence records of the species in its native range and the invaded European range. Occurrences for Japan were acquired by georeferencing herbarium collections housed in the TNS (Tsukuba, Japan), SAP (Sapporo, Japan) and GENT (Ghent, Belgium) herbaria. For the European invaded range, occurrences were obtained from the primary literature and collecting activities. The resulting database consisted of 94 records for Japan and 284 for Europe.

Species distribution models were inferred with Maxent version 3.3.2. (Phillips *et al.*, 2006), a machine-learning algorithm for SDM with superior performance among presence-only algorithms (Elith *et al.*, 2006). To avoid modeling issues relating to overparameterization and multicollinearity of environmental layers, we adopted a predictor selection procedure (Guisan & Zimmermann, 2000). Variable reduction was achieved with performance-based forward-stepwise selection. Model performance was measured in terms of the area under the curve (AUC) of the receiver operating characteristic for test data as implemented in Maxent (Phillips *et al.*, 2006). All analyses were replicated five times with random training and test sets (both 50%). The test AUC can be expected to yield a good trade-off between underfitting and overfitting the model. Underfitting is avoided because this naturally leads to low AUC values,

and overfitting is countered by using the test AUC instead of the training AUC as the metric to be maximized. A recent study shows that this approach selects models of appropriate complexity (Warren & Seifert, in press).

The best subset of predictors, producing the highest test AUC value, was used to carry out the final modeling step. Final models were inferred by running the same subsampling procedure (50% training, 50% test data) with ten replicate runs. All analyses used linear and quadratic features (leaving the other Maxent settings at their default values).

RESULTS

Twenty-three global rasters of marine environmental predictors were generated either from remotely sensed or in situ measured oceanographic data. The gridded fields of 5 arcmin spatial resolution (9.2 km) are available for download on the Bio-ORACLE website (http://www.bio-oracle.ugent.be). Elaborate descriptions of the layers, their data sources and quality maps for interpolated layers can be found on the Bio-ORACLE website and in the supplementary material (Appendix S1).

Agglomerative clustering of the data layers resulted in a strongly structured dendrogram with high bootstrap values for nearly all clusters. Four clusters (CL1-4), representing major macroecological axes, were clearly discernable. They represent ocean color bio-optical parameters (CL4), nutrients and dissolved oxygen (CL3), clouds (CL2) and temperature and light resources denoting latitudinal patterns (CL1). Besides these four groups, four singletons were present in the dendrogram. These are the mean salinity, calcite concentration, water pH and the annual range of sea surface temperature, a measure of seasonality.

CASE STUDY: *CODIUM FRAGILE* SUBSP. *FRAGILE*

The variable selection procedure resulted in thirteen sets of environmental layers that obtained the maximum test AUC of 0.989. From these sets, we opted to use the one with the fewest predictor variables for the final SDM, in order to avoid data redundancy and model overparameterization. This set consisted of 12 variables representing all six environmental clusters (variable names in boldface in figure 1).

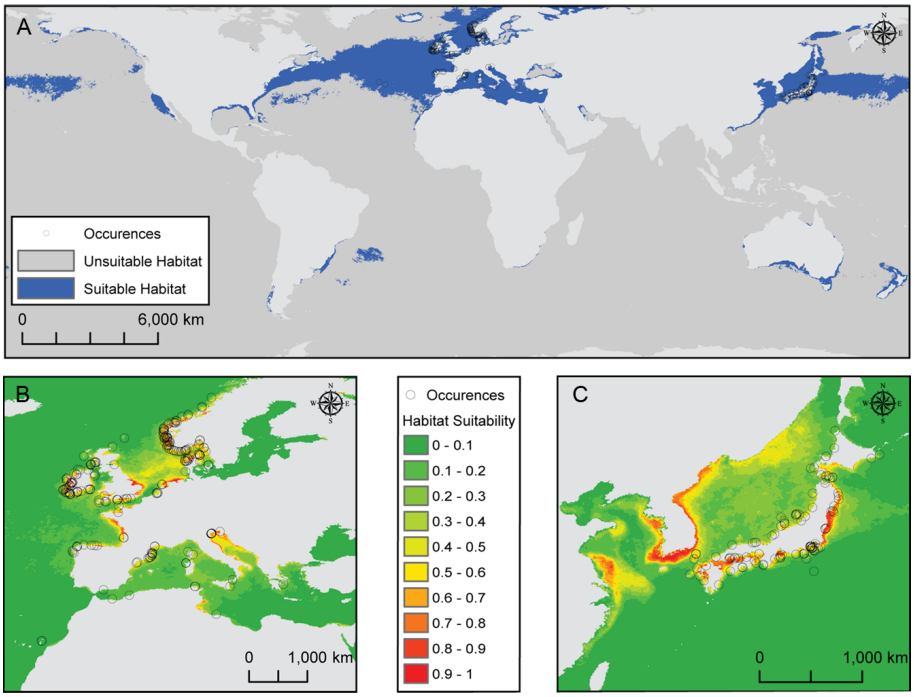


Figure 2. Inferred species distribution model of *Codium fragile* subsp. *fragile* based on occurrence records from the native (Japan) and invaded range (Europe). (a) Binary prediction map applying a minimum training presence threshold (0.019). (b, c) Habitat suitability maps of European and Asian ranges, respectively. Warmer colors represent areas with better predicted conditions. Black circles indicate occurrence records used to build models.

The resulting presence-only model for *C. fragile* achieved high classification success (average AUC of ten replicate runs: 0.990). The predicted habitat suitability maps of the model are depicted in figure 2.

The SDM predicted the realized distributions of the species in both the native and European invaded ranges very well. Other predicted areas include South Australia, New Zealand, North America and parts of South America, where the species has also been introduced and is spreading. The annual range of sea surface temperature, annual maximum of diffuse attenuation and mean of phosphate concentration were the most important variables explaining recorded observations (see Appendix S2 for more comprehensive SDM results, including a full list of the variable contributions).

DISCUSSION

Species distribution models play a central role in many fundamental and applied aspects of ecology. Besides improving our understanding of biogeography and dispersal barriers, they allow us to narrow in on species' ecological requirements, predict the effects of species invasions, habitat loss and climate change and can even lead to the discovery of new species (Peterson, 2006). The world's oceans harbor a high biodiversity (Costello *et al.*, 2010), and despite the importance of marine organisms for global biogeochemical cycles and human exploitation, their distribution and specific ecological needs are not nearly as well documented as their terrestrial counterparts. The application of SDM to marine species can also inform us about marine reserve design and conservation, and has the potential to predict how future climate and ocean acidification scenarios will affect the distribution and abundance of keystone species in the biogeochemical cycles.

While these examples clearly illustrate the need of SDM in the marine realm, the number of studies applying these methods to marine taxa is nearly an order of magnitude lower than those applying them to terrestrial taxa (Robinson *et al.*, in press). By introducing Bio-ORACLE, we hope to alleviate the need for user-friendly marine environmental data packages for global SDM applications. The development of a dataset of this type brings up several questions related to data quality, utility of the data for marine SDM, how the data should be used in practical applications, and how Bio-ORACLE compares to other marine environmental datasets. In what follows, we will touch on each of these topics.

DATA QUALITY

Several precautions were taken to produce a dataset of the highest possible quality. These included selection of input data with a high level of a priori quality check, the use of state of the art interpolation techniques and an assessment of uncertainty about the resulting data rasters.

For the interpolated maps based on WOD09 data, quality is spatially variable and depends on local environmental variability, the quality and density of the observations and the interpolation method (Hijmans *et al.*, 2005). The error maps computed by DIVA (see Appendix S1) result from an analysis on the covariance field of the data with respect to the true field. Pixel values in

these error maps represent relative error (ranging from 0 to 1) and give an idea about the level of confidence in the pixel values in the corresponding environmental raster. The overall error is small, and the highest uncertainty, i.e. the highest predicted error, occurred in regions with low data coverage such as high latitude areas(e.g. the northern polar seas, Hudson Bay, Antarctica) and some unsampled areas in the middle of the oceans.

The rasters derived from remotely sensed data only included information with the highest possible quality pixels (resulting from Level-3 quality maps). However, an inevitable source of error results from the irregular temporal sampling of ocean color sensors (MODIS & SeaWiFS) (Gregg & Casey, 2007). Daily data gaps exist due to clouds, thick aerosols, inter-orbit gaps, sun glint and high solar zenith angles (Gregg & Casey, 2007). Binning these data into climatologies makes these gaps disappear but could lead to unpredictable biases. These biases (and resulting uncertainties) are most pronounced at high latitudes. For example, chlorophyll A, photosynthetically available radiation and diffuse attenuation, which are measured at relatively short wavelengths (in the visible spectrum), cannot be accurately measured during the winter season at high latitudes due to high solar zenith angles (Gregg & Casey, 2007). Sea surface temperature data do not suffer from this effect because they are measured in the thermal infrared part of the spectrum (longer wavelengths). A second source of bias affecting the quality of the remotely sensed data rasters follows from intra- and extrapolation of data for pixels with missing data. A simulation experiment (Appendix S1) indicates that extrapolations do lead to errors, but that these are small (generally < 1%) and that extrapolation into coastal pixels was somewhat more error-prone than interpolation of pixels in the open ocean.

The majority of potential errors in the datasets are due to the absence of in situ measurements, bias in remotely sensed data and extrapolation towards coastal areas. All of these problems are more pronounced in high-latitude areas. Therefore, we advise against the use of Bio-ORACLE in latitudes above 70° latitude N/S. We provide data rasters cropped at 70°N – 70°S as well as rasters spanning the entire latitudinal range, the latter only to be used with great vigilance and judicious consideration of results.

UTILITY FOR MARINE SDM

The distribution of marine organisms is controlled by the interplay of a multitude of physical, chemical and biological variables. Bio-ORACLE includes both variables deemed important physiological determinants as well as potentially useful proxies. Each of the clusters and singletons in the dataset (figure 1) represents a distinct aspect of the marine macroecological environment. Temperature is thought to be the most important physical oceanographic variable determining the abundance, the spatial distribution and diversity of marine ectotherms (Schmidt-Nielsen, 1990; Lüning, 1990). Sea surface temperature clusters with photosynthetically available radiation (CL1), an essential and potentially limiting variable for photosynthetic organisms as it provides in their energy needs. Chlorophyll A (CL4) was included because it is a useful proxy for the trophic status of the surface waters (Duan *et al.*, 2007). However, the chlorophyll A metrics do not group with the actual nutrient layers (nitrate, phosphate and silicate) in the cluster analysis and consequently capture another dimension of the marine environment. We presume that this is due to an organic vs. inorganic nutrients dichotomy, chlorophyll A being a proxy of primary production and the WOD09 layers representing inorganic nutrients. Cloud cover variables (CL2) were included because of their potential to indirectly influence marine organisms. Clouds can block the transmission of light and harmful UV radiation and affect intertidal communities and organisms abounding in the ocean surface layer (Karentz & Bosch, 2001; Mangel *et al.*, 2010; Roleda *et al.*, 2005; Dring *et al.*, 1996). Four environmental entities did not cluster in one of the previously mentioned groups: salinity, calcite concentration, pH and sea surface temperature range. Besides temperature, salinity is known to be among the most important factors influencing marine life (Lüning, 1990; Gogina & Zettler, 2010). Ocean acidity (pH) plays a critical role in mediating physiological reactions (Wootton *et al.*, 2008) and numerous important groups of marine organisms have calcium carbonate skeletons that dissolve when pH drops (Doney *et al.*, 2009). The range in sea surface temperature is a measure of temperature seasonality. Our case study clearly shows that this variable can be an important determinant (or proxy) of marine species' distributions as it had the highest variable contribution to the SDM of *Codium fragile* (See Appendix S3). In contrast to the terrestrial environment where seasonal climatic variability increases with latitude (Chown *et al.*, 2004),

the seasonal variability in sea surface temperature is highest at intermediate latitudes (Clarke, 2009).

We aimed to illustrate the utility of Bio-ORACLE for marine SDM by generating a distribution model of *Codium fragile*. This highly invasive species is of economic interest and various aspects of its physiology and bioactive compounds, as well as the ecology and genetic signature of its invasion have been studied (Trowbridge, 1998; Provan *et al.*, 2008). The extremely high AUC values obtained for both training and test data sets show that the Bio-ORACLE rasters capture the macroecological preferences of the species and that, when used correctly (see below), the dataset permits building highly accurate SDM of marine species. From an organismal perspective, the SDM of *Codium fragile* is of interest in that it predicts other areas where the species could thrive. In fact, the species is known to have been introduced and is spreading in some of the predicted regions (Northern America, Australia, New-Zealand and South America) (Provan *et al.*, 2008). An in-depth analysis of the *Codium fragile* models is beyond the scope of this paper and will be presented elsewhere.

USING BIO-ORACLE FOR MARINE SDM

It is evident that ecological preferences differ between species and not all variables are useful in predicting a species' distribution. Choosing the right predictor variables for a particular species of interest is considered to be one of the most crucial steps in the SDM procedure (Guisan & Zimmermann, 2000). Variables could be chosen based on the knowledge that they are ecologically meaningful for the target species (Guisan & Zimmermann, 2000; Austin, 2002) and/or have good explanatory power (Araujo & Guisan, 2004). The latter aspect has been given much attention in regression techniques using presence-absence data, where several methods for predictor selection are available (e.g. stepwise selection with cross-validation, ridge regression, lasso) (Guisan & Thuiller, 2005). Unfortunately, predictor selection has been getting much less attention in recent presence-only modeling approaches.

For our case study, a performance-based forward stepwise variable selection procedure resulted in the selection of 12 out of the 23 variables in Bio-ORACLE. The importance of predictor selection is confirmed by the fact that a model built with all 23 layers resulted in considerably lower predictive power, most likely as a consequence of predictor collinearity and model

overparameterization. This illustrates that datasets like Bio-ORACLE and WorldClim should not be used blindly but that SDM requires meticulous species-specific variable selection, preferably based on a combination of physiological knowledge and variable selection approaches. In this context, the development of information criterion-based model selection (e.g. Akaike and Bayesian information criteria) for use in presence-only SDM applications would be useful (Warren & Seifert, in press).

COMPARISON TO OTHER MARINE ENVIRONMENTAL DATASETS

Bio-ORACLE was developed to be a ready-to-use global environmental dataset for shallow-water marine species distribution modeling. Other marine datasets for SDM do exist but the uniformity and user-friendliness of Bio-ORACLE is unique. Table 1 lists strengths and weaknesses of Bio-ORACLE compared to other marine environmental datasets. Noteworthy examples of marine preprocessed datasets that contain environmental data potentially informative for SDM are AquaMaps (Kaschner *et al.*, 2008a) and HexaCoral (Fautin & Buddemeier, 2008). AquaMaps is an approach to generate predictions of the natural occurrence of marine species based on their environmental tolerances (Kaschner *et al.*, 2008b). The AquaMaps datasets represent long-term averages of temporally varying environmental variables (Ready *et al.*, 2010). The HexaCoral datasets were developed to enable environmental classification (typology) and understand spatial and temporal patterns in biogeochemistry and biogeography. Both AquaMaps and HexaCoral can be downloaded at a spatial resolution of 30 arcminutes.

Common SDM applications require a set of uniformly constructed environmental layers. Bio-ORACLE provides data with a consistent landmask across all layers. We also made the data available in the ascii raster grid format used by many popular SDM algorithms (e.g. GARP, Maxent). It has been common practice in marine environmental modeling to use data with a spatial resolution of 30 arcminutes (Guinotte *et al.*, 2006). Bio-ORACLE has a considerably higher resolution of 5 arcminutes (ca. 9.2 km). Our choice for this resolution is a trade-off between the desire for sufficient resolution in near-shore environments, manageability of the rasters on current desktop computers and avoiding unreasonable interpolations.

Table 1. Comparison between freely available marine environmental datasets

	WOD2009 [e]	OCEAN COLOR [f]	HEXACORAL [g]	AQUAMAPS [h]	HALPERN [i]	Bio-ORACLE
Resolution	30-60 arcminutes (~ 55–110 km)	2.5-5 arcminutes (~ 4–9 km)	30 arcminutes (~ 55 km)	30 arcminutes (~ 55 km)	0.5 arcminutes (~ 1 km)	5 arcminutes (~ 9 km)
Uniform landmask [a]	Yes	No	Yes	No	No	Yes
Uniform geographic range	Yes	Yes	No	Yes	No	Yes
Dataset suitable for fine scale coastal studies	No	Yes	No	No	Yes	Yes
Multiple depth levels [b]	Yes	No	Yes	Yes	No	No
Uniform file format	Yes	Yes	Yes	Yes	Yes	Yes
Uniform file format suitable for common SDM applications [c]	No	No	No	No	Yes [d]	Yes
Equal-area grids available	No	No	No	No	Yes	Yes

[a] Uniformity of landmask across all data layers in package, [b] Layers provided at different subsurface depths, [c] ascii raster grid format, [d] Suitability for SDM hampered by large file sizes.

References: [e] WOD2009 (Boyer et al., 2009), [f] Ocean Color (Feldman & Mcclain, 2010), [g] Biogeoinformatics of the Hexacorals (Fautin & Buddemeier, 2008), [h] AquaMaps (Kaschner et al., 2008b), [i] Global Mapping of Human Impacts to Marine Ecosystems (Halpern et al., 2008)

Furthermore, this resolution makes the dataset suitable for addressing questions about distributions at a global scale while still allowing model predictions at a resolution fine enough for most management purposes. In this context, it is important to note that the variables included in the Bio-ORACLE dataset are situated at the macroecological level, and when interpreting models at a fine spatial resolution, certain aspects of the organisms' microhabitat preferences (e.g. presence of suitable substrate for benthic species) become important to consider besides the macroecological niche dimensions.

Even though the comparison in Table 1 shows that Bio-ORACLE is currently among the best datasets for marine SDM, we also want to emphasize the utility of the other datasets. Our evaluation was focused on SDM applications and Bio-ORACLE was specifically designed for this purpose whereas others were not. Nonetheless, these datasets can complement Bio-ORACLE in various ways. For example, they contain some environmental dimensions that are not included in our database and variables at multiple depth levels.

CONCLUSIONS AND PERSPECTIVES

Species distribution models have gained importance in various biological disciplines in recent years. Remarkably, they are less commonly used in studies of marine species than of terrestrial taxa. The present study was carried out to develop a marine counterpart of the WorldClim database, which is widely used for terrestrial SDM. Bio-ORACLE is a dataset consisting of 23 environmental rasters for marine species distribution modeling at a global scale. We hope that the availability of this set of environmental rasters will bring marine SDM on par with terrestrial studies. Our species distribution model of the invasive seaweed *Codium fragile* clearly shows that the rasters contain information relevant to the distribution of marine species and permits developing very accurate species distribution models.

We consider the present version of Bio-ORACLE an important first step in the development of a more complete set of environmental data rasters. Progress can obviously still be made, for example by including depth-related variables, various other physical parameters and layers representing important limiting nutrients in the marine environment. The present dataset will hopefully

provide our colleagues and us with the necessary groundwork to move this objective forward.

ACKNOWLEDGEMENTS

We are grateful to three anonymous referees and the associate editor for their constructive criticisms and suggestions. We thank Satoshi Shimada (SAP herbarium) and Taiju Kitayama (TNS herbarium) for providing *C. fragile* subsp. *fragile* records. LT is funded by the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT). HV is a postdoctoral fellow of the Research Foundation – Flanders. Part of this work was carried out using the Stevin Supercomputer Infrastructure at Ghent University.

SUPPLEMENTARY MATERIAL

SUPPORTING INFORMATION S1

Appendix 1: Extended materials and methods used to build Bio-ORACLE dataset

General Approach

A set of 23 macroecological variables representing environmental dimensions assumed to influence the distribution of marine shallow water organisms was assembled. Our data collecting approach consisted of the compilation of preprocessed remotely sensed and in situ measured oceanographic data. We limited our analyses to variables relevant at the macroecological level and with global coverage. This supplementary material aims to give an overview and more detailed information of the Bio-ORACLE variables.

Our data processing pipeline consisted of the following steps:

- Data acquisition (WOD09 and Ocean Color Web)
- In case of WOD09 data, raster generation by interpolation with DIVA
- Conversion of raster to a convenient format: ESRI ascii
- Calculation of metrics (e.g. rasters with mean, minimum and maximum values)
- Application of uniform landmask
- Clipping rasters to 70°N – 70°S

Recommendations & Availability

The authors advise against the use of the Bio-ORACLE variables at latitudes higher than 70° N/S to reduce possible biases and potential errors in the data at high latitudes (see text for explanation).

Bio-ORACLE is released under the GNU General Public License and is available for download on http://www.bio-oracle.ugent.be.

Data layers

Remotely sensed data

Calcite concentration (mol/m³)	
Original Spatial Resolution	5 arcmin (9.2 km)
Sensor	Aqua-MODIS
Data	Seasonal climatologies
Temporal Range	2002 - 2009
Brief description	Calcite concentration indicates the concentration of calcite (CaCO3) in oceans.
Manipulation	Derivation of metric: Mean
Source	Reference: (Feldman & McClain 2010) URL: http://oceancolor.gsfc.nasa.gov/

Chlorophyll A concentration (mg/m³)	
Original Spatial Resolution	5 arcmin (9.2 km)
Sensor	Aqua-MODIS
Data	Monthly climatologies
Temporal Range	2002 - 2009
Brief description	Chlorophyll A concentration indicates the concentration of photosynthetic pigment chlorophyll A (the most common "green" chlorophyll) in oceans. Please note that in shallow water these values may reflect any kind of autotrophic biomass.
Manipulation	Derivation of metrics: Mean, Minimum, Maximum, Range
Source	Reference: (Feldman & McClain 2010) URL: http://oceancolor.gsfc.nasa.gov/

Cloud fraction (%)	
Original Spatial Resolution	6 arcmin (11 km)
Sensor	Terra-MODIS
Data	Monthly images
Temporal Range	2005 - 2010
Brief description	Cloud fraction indicates how much of the earth is covered by clouds.
Manipulation	Derivation of metrics: Mean, Minimum, Maximum Bilinear interpolation (10 km → 9.2 km)
Source	Reference: (NASA 2010) URL: http://neo.sci.gsfc.nasa.gov/Search.html

Diffuse attenuation coefficient at 490 nm (m^{-1})	
Original Spatial Resolution	5 arcmin (9.2 km)
Sensor	Aqua-MODIS
Data	Monthly climatologies
Temporal Range	2002 - 2009
Brief description	The diffuse attenuation coefficient is an indicator of water clarity. It expresses how deeply visible light in the blue to the green region of the spectrum penetrates in to the water column.
Manipulation	Derivation of metrics: Mean, Minimum, Maximum
Source	Reference: (Feldman & McClain 2010) URL: http://oceancolor.gsfc.nasa.gov/

Photosynthetically Available Radiation (Einstein/m^2/day)	
Original Spatial Resolution	5 arcmin (9.2 km)
Sensor	SeaWiFS
Data	Monthly climatologies
Temporal Range	1997 - 2009
Brief description	Photosynthetically Available Radiation (PAR) indicates the quantum energy flux from the Sun (in the spectral range 400-700 nm) reaching the ocean surface.
Manipulation	Derivation of metrics: Mean, Maximum Minimum PAR was considered, but excluded due to the high level of artifacts in original data.
Source	Reference: (Feldman & McClain 2010) URL: http://oceancolor.gsfc.nasa.gov/

Sea Surface Temperature (°C)	
Original Spatial Resolution	5 arcmin (9.2 km)
Sensor	Aqua-MODIS
Data	Monthly climatologies
Temporal Range	2002 - 2009
Brief description	Sea surface temperature is the temperature of the water at the ocean surface. This parameter indicates the temperature of the topmost meter of the ocean water column.
Manipulation	Derivation of metrics: Mean, Minimum, Maximum, Range
Source	Reference: (Feldman & McClain 2010) URL: http://oceancolor.gsfc.nasa.gov/

In situ measured oceanographic data

Dissolved oxygen (ml/l)	
Database	World Ocean Database 2009
Data	Standard Level Data: Ocean Station Data (OSD); High-resolution Conductivity-Temperature-Depth (CTD) (Surface)
Temporal Range	1898 - 2009
Number of data points	540582
Brief description	Dissolved oxygen concentration [O_2]
Manipulation	DIVA interpolation
Source	Reference: (Boyer *et al.* 2009) URL: http://www.nodc.noaa.gov/

Nitrate (µmol/l)	
Database	World Ocean Database 2009
Data	Standard Level Data: OSD (Surface)
Temporal Range	1928 - 2008
Number of data points	189530
Brief description	This layer contains both [NO_3] and [NO_3+NO_2] data. By this we mean chemically reactive dissolved inorganic nitrate and nitrate or nitrite. (It is important to note that data reported as [NO_3] in the WOD09 should be used with caution because it is difficult to verify that the [NO_3] (nitrate) data are [NO_3+NO_2] or [NO_3]. (Boyer *et al.* 2009))
Manipulation	DIVA interpolation
Source	Reference: (Boyer *et al.* 2009) URL: http://www.nodc.noaa.gov/

pH (unitless)	
Database	World Ocean Database 2009
Data	Standard Level Data: OSD (Surface)
Temporal Range	1910 - 2007
Number of data points	117833
Brief description	Measure of acidity in the ocean.
Manipulation	DIVA interpolation
Source	Reference: (Boyer *et al.* 2009) URL: http://www.nodc.noaa.gov/

Phosphate (µmol/l)	
Database	World Ocean Database 2009
Data	Standard Level Data: OSD (Surface)
Temporal Range	1922 - 1986
Number of data points	226816
Brief description	Reactive ortho-phosphate concentration [HPO_4^{-2}] in the ocean.
Manipulation	DIVA interpolation
Source	Reference: (Boyer *et al.* 2009) URL: http://www.nodc.noaa.gov/

Salinity (PSS)	
Database	World Ocean Database 2009
Data	Standard Level Data: CTD (Surface)
Temporal Range	1961 - 2009
Number of data points	532377
Brief description	Salinity indicates the dissolved salt content in the ocean.
Manipulation	DIVA interpolation
Source	Reference: (Boyer *et al.* 2009) URL: http://www.nodc.noaa.gov/

Silicate (µmol/l)	
Database	World Ocean Database 2009
Data	Standard Level Data: OSD & CTD (Surface)
Temporal Range	1930 - 2008
Number of data points	234417
Brief description	This variable indicates the concentration of silicate or ortho-silicic acid [$Si(OH)_4$] in the ocean.
Manipulation	DIVA interpolation
Source	Reference: (Boyer *et al.* 2009) URL: http://www.nodc.noaa.gov/

References

Boyer T.P., Antonov J.I., Baranova O.K., Garcia H.E., Johnson D.R., Locarnini R.A., Mishonov A.V., O'Brien T.D., Seidov D., Smolyar I.V. & Zweng M.M. (2009). *World Ocean Database 2009*. U.S. Gov. Printing Office, Washington D.C.

Feldman G.C. & McClain C.R. (2010). Ocean Color Web. URL http://oceancolor.gsfc.nasa.gov/

NASA (2010). NASA Earth observations (NEO). URL http://neo.sci.gsfc.nasa.gov/Search.html

Correlation Matrix

	calcite_mean	chlo_max	chlo_mean	chlo_min	chlo_range	cloudfr_max	cloudfr_mean	cloudfr_min	da_max	da_mean	da_min	dissox	nitrate_mean	par_max	par_mean	ph_mean	phos_mean	sal_mean_an	silica_mean	sst_max	sst_mean	sst_min	sst_range
calcite_mean	1.00	0.40	0.46	0.48	0.27	0.00	-0.04	-0.06	0.46	0.55	0.61	0.07	-0.04	-0.05	-0.04	-0.16	-0.02	-0.24	0.01	-0.03	-0.05	-0.07	0.14
chlo_max		1.00	0.92	0.74	0.92	0.06	0.01	-0.02	0.85	0.83	0.75	0.17	0.00	-0.10	-0.09	-0.22	0.02	-0.47	0.07	-0.11	-0.15	-0.18	0.25
chlo_mean			1.00	0.91	0.72	0.04	0.00	-0.03	0.77	0.84	0.83	0.17	0.00	-0.10	-0.09	-0.25	0.01	-0.51	0.07	-0.11	-0.15	-0.17	0.20
chlo_min				1.00	0.42	0.02	-0.02	-0.04	0.60	0.72	0.81	0.13	-0.03	-0.09	-0.07	-0.25	-0.01	-0.44	0.03	-0.09	-0.11	-0.13	0.14
chlo_range					1.00	0.07	0.02	-0.01	0.79	0.71	0.55	0.15	0.02	-0.08	-0.07	-0.16	0.03	-0.37	0.08	-0.09	-0.13	-0.16	0.25
cloudfr_max						1.00	0.92	0.79	0.22	0.17	0.08	0.56	0.58	-0.59	-0.65	-0.30	0.59	-0.24	0.49	-0.58	-0.59	-0.59	0.01
cloudfr_mean							1.00	0.95	0.13	0.08	0.00	0.57	0.65	-0.58	-0.71	-0.28	0.65	-0.16	0.48	-0.62	-0.62	-0.62	-0.01
cloudfr_min								1.00	0.06	0.01	-0.06	0.52	0.62	-0.52	-0.66	-0.21	0.61	-0.10	0.40	-0.58	-0.58	-0.57	-0.04
da_max									1.00	0.96	0.82	0.37	0.09	-0.26	-0.27	-0.29	0.12	-0.53	0.16	-0.29	-0.35	-0.38	0.35
da_mean										1.00	0.94	0.38	0.07	-0.28	-0.26	-0.32	0.10	-0.59	0.17	-0.29	-0.34	-0.37	0.30
da_min											1.00	0.32	0.00	-0.25	-0.20	-0.31	0.04	-0.59	0.12	-0.24	-0.27	-0.30	0.23
dissox												1.00	0.68	-0.81	-0.88	-0.41	0.70	-0.49	0.64	-0.97	-0.98	-0.97	-0.05
nitrate_mean													1.00	-0.57	-0.67	-0.41	0.95	-0.08	0.86	-0.77	-0.74	-0.71	-0.23
par_max														1.00	0.80	0.39	-0.60	0.41	-0.58	0.79	0.78	0.75	0.18
par_mean															1.00	0.34	-0.68	0.36	-0.58	0.89	0.91	0.92	-0.07
ph_mean																1.00	-0.44	0.30	-0.41	0.42	0.41	0.39	0.11
phos_mean																	1.00	-0.15	0.83	-0.79	-0.76	-0.73	-0.24
sal_mean_an																		1.00	-0.18	0.38	0.42	0.44	-0.17
silica_mean																			1.00	-0.70	-0.67	-0.64	-0.24
sst_max																				1.00	0.99	0.96	0.17
sst_mean																					1.00	0.99	0.01
sst_min																						1.00	-0.10
sst_range																							1.00

Simulation study: Cubic inter/extrapolation

In order to produce rasters with a uniform landmask across the complete dataset two operations were necessary:

- Values that fell on land needed to be masked;
- Missing values needed to be estimated.

The latter operation was done using a cubic inter/extrapolation algorithm. We chose this algorithm because of its stable and smooth characteristics. To assess the quality of the interpolation we performed a simulation study on the remotely sensed data*.
The evaluation consisted of:

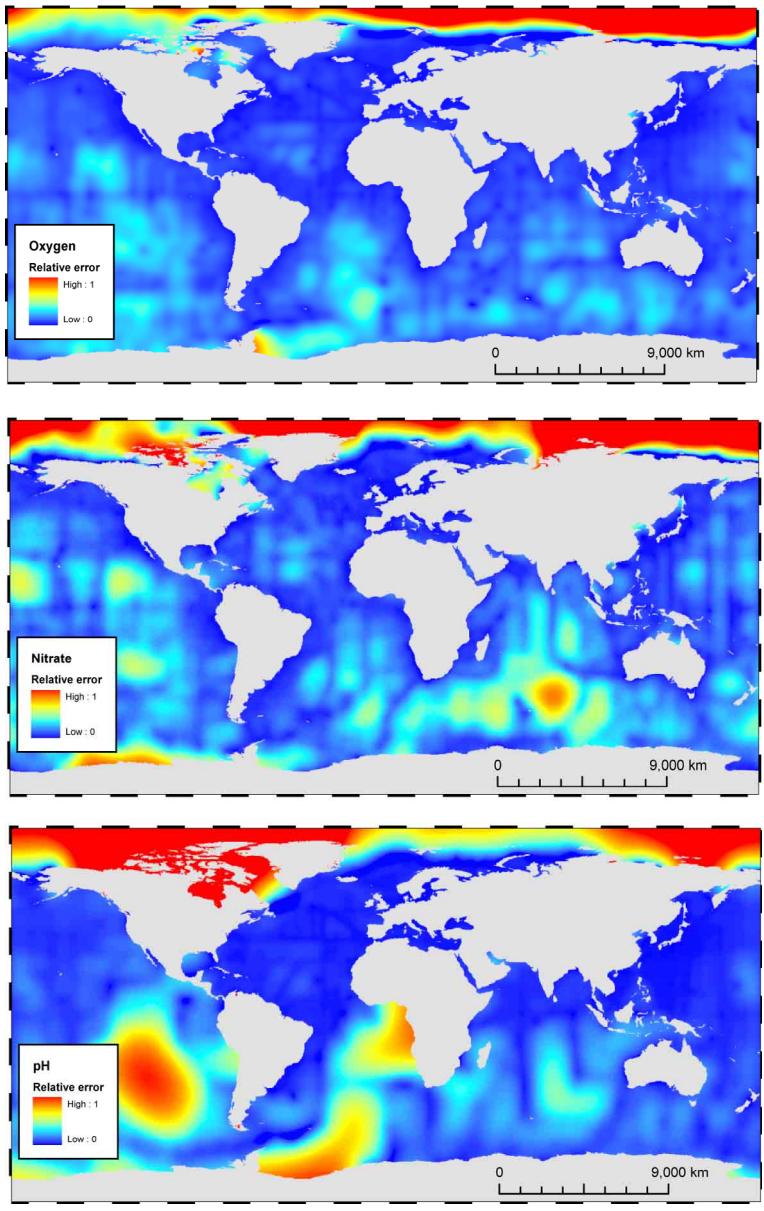
- Random selection of a monthly climatology of each remotely sensed variable;
- Random selection of 100 'open sea' pixels (i.e. pixels that are completely surrounded by other sea pixels) from the respective climatology;
- Random selection of 100 'coastal' pixels (i.e. pixels directly adjacent to one or more no-data value) from the respective climatology;
- Removal of these 200 data points from the respective climatology;
- Application of the cubic inter/extrapolation algorithm;
- Extraction of data from both the original and the inter/extrapolated layer;
- Calculation of the average difference between both layers;
- Evaluation of that difference in relation to the total variable range.

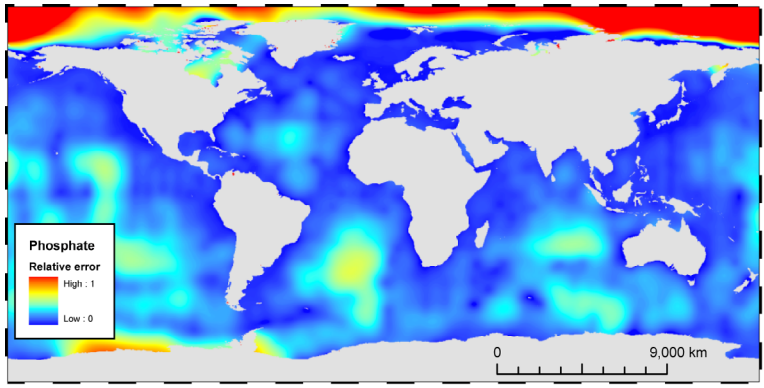
The results are shown in the table below:

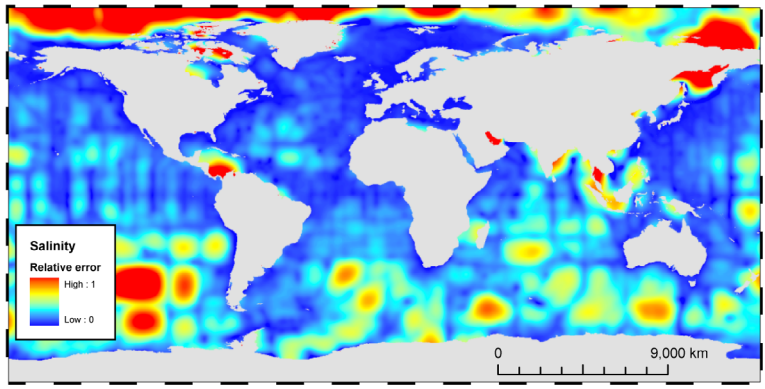
Variable	Coast		Open Sea	
	Δ(avg(Interpol – Real))	% of total variable range	Δ(avg(Interpol – Real))	% of total variable range
Sea Surface Temperature	0.22675 °C	0.57	0.04161 °C	0.11
Chlorophyll A Concentration	0.38901 mg/m^3	0.60	0.00831 mg/m^3	0.01
Diffuse Attenuation	0.00922 m^{-1}	0.80	0.00088 m^{-1}	0.08
Photosynthetically Available Radiation	0.49984 Einstein/m^2/day	0.73	0.18093 Einstein/m^2/day	0.27
Calcite Concentration	0.00086 mol/m^3	1.54	0.000134 mol/m^3	0.24

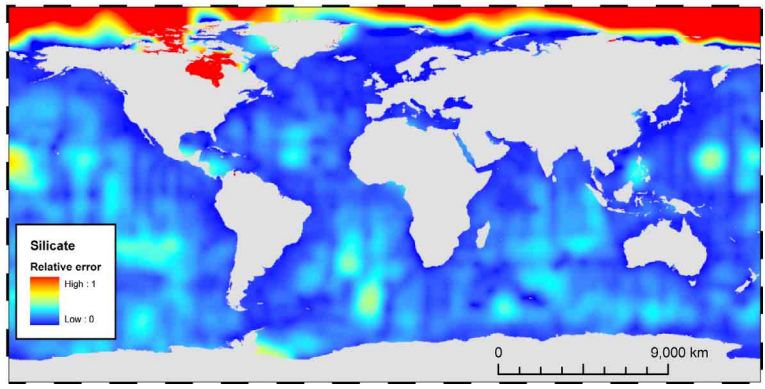
* Cloud cover data was not taken into account as no pixels needed to be inter/extrapolated.

Supporting figure S1. Error maps computed by DIVA.




Phosphate
Relative error
High : 1
Low : 0
0
9,000 km


Salinity
Relative error
High : 1
Low : 0
0
9,000 km


Silicate
Relative error
High : 1
Low : 0
0
9,000 km

SUPPORTING INFORMATION S2

Appendix 2: Extended results: Species Distribution Modeling: *C. fragile* subsp. *fragile*

This appendix summarizes the results of 10 replicate Maxent runs (see text for explanation). Figures are developed by Maxent version 3.3.2. (Phillips *et al.* 2006).

Analysis of omission/commission

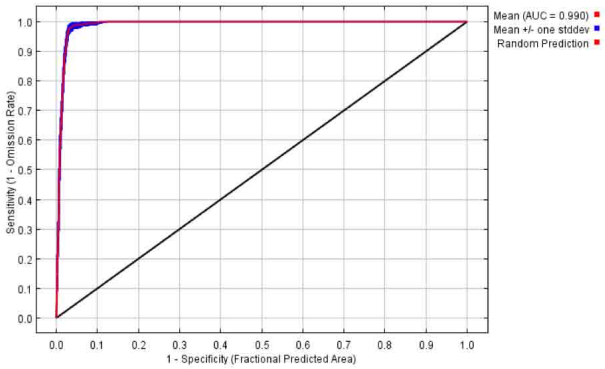


Figure S2.1 The following figure shows the receiver operating characteristic (ROC) curve, averaged over the replicate runs. The average test AUC for the replicate runs is 0.990, and the standard deviation is 0.001.

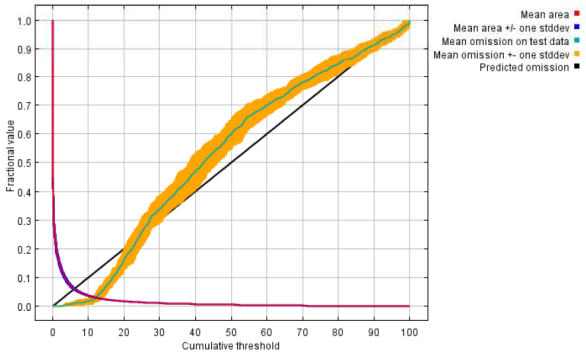


Figure S2.2 This figure shows the test omission rate and predicted area as a function of the cumulative threshold, averaged over the replicate runs.

Analysis of variable contributions

Table S2.1 This table represents an estimate of relative contributions of the environmental variables to the Maxent model. Values shown are averages over replicate runs.

Variable	Percent contribution
Sea surface temperature (RANGE)	36.3
Diffuse attenuation (MAX)	15.6
Phosphate concentration	12.5
Sea surface temperature (MEAN)	10.8
Salinity (MEAN)	8.8
PAR (MAX)	7
Cloud cover (MAX)	3.9
Nitrate concentration (MEAN)	1.7
Cloud cover (MEAN)	1.5
Diffuse attenuation (MIN)	1.1
Chlorophyll A concentration (MIN)	0.7
pH (MEAN)	0

Response curves

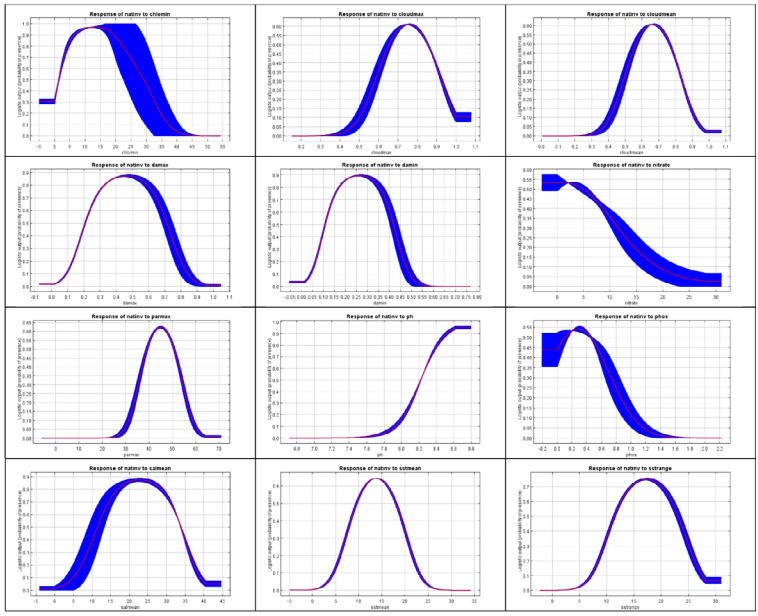


Figure S2.3 These curves represent models created using only the corresponding variable. They reflect the dependence of predicted suitability both on the selected variable and on dependencies induced by correlations between the selected variable and other variables.

Reference

Phillips S.J., Anderson R.P. & Schapire R.E. (2006). Maximum entropy modeling of species geographic distributions. *Ecol. Model.*, 190, 231-259.

SUPPORTING INFORMATION S3

Appendix 3: Transformation methods 'Behrmann Projection'

There exist several ways to transform geographical coordinates (lat/lon WGS84) to a Behrmann equal-area projection.

1. Using ArcMap 9.2 (http://www.esri.com)

- Import pointfile as shapefile
- Transformation: ArcToolbox → Data Management Tools → Projections and Transformations → Feature → Project

2. Using transformation formulae

   The Behrmann projection is a cylindrical equal-area projection with fixed standard parallels at ±30°. The transformation equations are:

$$x = R\,\lambda\cos\varphi_0$$

$$y = \frac{R}{\sin\varphi_0}\sin\varphi$$

   where λ is the longitude, φ is the latitude and φ_0 is the standard latitude, all expressed in radians. R represents the earth radius (6378000m).

   These formulas yield approximate coordinates because differences in the earth's radius are not taken into account.