Chapter 4

In search of relevant predictors for marine species distribution modelling using the MarineSPEED benchmark dataset

Samuel Bosch[1,2], Lennert Tyberghein[1], Klaas Deneudt[1], Francisco Hernandez[1] and Olivier De Clerck[2]

[1]*Flanders Marine Institute (VLIZ), InnovOcean site, Wandelaarskaai 7, 8400 Ostend, Belgium*
[2]*Research Group Phycology, Biology Department, Ghent University, Krijgslaan 281/S8, 9000 Ghent, Belgium*

Submitted in March 2017.

Abstract

Aim

Ideally, datasets for species distribution modelling (SDM) contain evenly sampled records covering the entire distribution of the species, confirmed absences and auxiliary ecophysiological data allowing informed decisions on relevant predictors. Unfortunately, these criteria are rarely met for marine organisms for which distributions are too often only scantly characterized and absences generally not recorded. Here, we investigate predictor relevance as a function of modelling algorithms and settings for a global dataset of marine species. Furthermore, we promote the usage of a standardized benchmark dataset (MarineSPEED) for methodological SDM studies.

Location

Global marine.

Methods

We selected well studied and identifiable species from all major marine taxonomic groups. Distribution records were compiled from public sources (e.g. OBIS, GBIF, Reef Life Survey) and linked to environmental data from Bio-ORACLE and MARSPEC. Using this dataset, predictor relevance was analysed under different variations of modelling algorithms, numbers of predictor variables, cross-validation strategies, sampling bias mitigation methods, evaluation methods and ranking methods. SDMs for all combinations of predictors from 8 correlation groups were fitted and ranked, from which the top five predictors were selected as the most relevant.

Results

We collected two million distribution records from 514 species across 18 phyla and made them available with associated environmental data and cross-validation splits through the R package *marinespeed* and at http://marinespeed.org. Mean sea surface temperature and calcite are respectively the most relevant and irrelevant predictors. A less clear pattern was derived from the other predictors. The biggest differences in predictor relevance were induced by varying the number of predictors, the modelling algorithm and the sample selection bias correction.

Main conclusions

While temperature is a relevant predictor of global marine species distributions, considerable variation in predictor relevance is linked to the SDM setup. Future methodological SDM studies should consider the use of a benchmark dataset.

Introduction

Species distributions are increasingly modelled for conservation and ecological purposes. A better understanding of the mechanisms shaping species distributions allows for more accurate predictions of the future distribution of species in a rapidly changing world (Franklin, 2009). Climatological conditions are currently changing at an unprecedented rate and anthropogenic activities displace species out of their native area across the globe resulting in biological invasions (Walther et al., 2009).

A mechanistic link between the abiotic factors and the species distributions is traditionally gleaned from physiological studies subjecting individuals to various environmental conditions and assessing their reaction norms. However, not all species lend themselves equally well to ex situ experiments. Also, the experimental setup may only approximate realistic environmental conditions to a limited degree. Furthermore, such physiological studies typically require prior knowledge on the ecological factors governing the distribution ranges (Kearney & Porter, 2009). Given these difficulties, species distribution modelling (SDM), alternatively known as Ecological Niche Modelling (ENM), offers an attractive alternative (Elith et al., 2010). SDM correlates species occurrences, and optionally absences, with environmental data to create an estimation of the ecological niche and a projection in geographic space of this niche (Austin, 2002). The obvious advantage of correlative SDMs is that they require little knowledge of the mechanistic links between organisms and their environments. Thanks to the availability of an increasing number of online distribution records (e.g. OBIS, GBIF), pre-processed environmental data layers (e.g. Worldclim, Climond, Bio-ORACLE, MARSPEC) and modelling algorithms accessible through various statistical packages, SDM has become a widely applied technique in ecology and conservation biology (Pacifici et al., 2017).

Despite this, studies on general SDM theory and methodology mostly focus on the terrestrial environment (reviewed in Franklin 2009; Elith & Leathwick 2009; Peterson et al. 2011). A minority of papers specifically address distribution modelling methods in the marine environment: presence-only algorithms (Cheung et al., 2008; Ready et al., 2010; Beaugrand et al., 2011), algorithm comparisons (MacLeod et al., 2008; Palialexis et al., 2011; Šiaulys & Bučas, 2012), 3D modelling (Bentlage et al., 2013), rare species (Stirling et al., 2016), joint SDMs (Torres et al., 2008), ensemble modelling (Downie et al., 2013), scale effects (Pittman & Brown, 2011; Nyström Sandman et al., 2013), null models (Merckx et al., 2011), model selection (Verbruggen et al., 2013), pseudo-absence generation (Huang et al., 2011; Coro et al., 2016) and predictor datasets (Tyberghein et al., 2012; Sbrocco & Barber, 2013).

Although the importance of selecting biologically relevant predictors, and its impact on model uncertainty and transferability has been highlighted by several studies (Araújo & Guisan, 2006; Barry & Elith, 2006; Synes & Osborne, 2011; Braunisch et al., 2013; Verbruggen et al., 2013; Petitpierre et al., 2017) to date no comprehensive study on the relevance of the predictors of marine species distributions across taxa has been performed. But, note that Bradie & Leung (2016), in their meta-analysis on variable importance from MaxEnt SDMs, included a limited set of marine species. Bradie & Leung (2016) found that temperature and to a smaller extent bathymetry and salinity contributed the most to marine species distribution models. While the impact of geographic scale, algorithm and pseudo-absence selection on the importance of predictors have been addressed to some degree (VanDerWal et al., 2009; Elith et al., 2010; Nyström Sandman et al., 2013; Bucklin et al., 2015) the impact of these and other aspects of SDM have not been studied on a global scale.

In this study, we created the Marine SPEcies with Environmental Data (MarineSPEED) dataset. This benchmark dataset, containing distribution records belonging to 514 well-studied taxa with a broad taxonomic, climatologic and geographic diversity, is used to investigate marine predictor relevance under an array of modelling parameters and algorithms. With this, we aim to answer two questions: (1) what are the most relevant predictors of marine species distributions and (2) which part of the SDM process impacts the relevance of predictors the most. Additionally, this study aims to promote the usage of benchmark datasets in methodological SDM studies as this allows for reproducible and comparable results.

Methods

Species data

For the marine species benchmark dataset we selected species from an array of taxonomic groups, climatological preferences and distribution patterns. We aimed to include species that are well-studied in terms of their distribution and that often would classify as iconic species. For a species to be considered we required the availability of at least 100 distribution records in public databases.

Species distribution records were collected from the Ocean Biogeographic Information System (OBIS; iobis.org, accessed February 2016), from the Global Biodiversity Information Facility (GBIF; gbif.org, accessed January 2016), the Reef Life Survey (RLS; reeflifesurvey.com, accessed February 2016) and for a few species via personal communications. For downloading the records from OBIS and GBIF the R (R Core Team, 2016) clients *robis* (Provoost et al., 2016) and *rgbif* (Chamberlain et al., 2016a) were used, respectively. A list of data sources is found in Appendix S1 in Supporting information. The distribution records were subsequently filtered until only one record remained in each cell of an equal-area grid with a per cell area of 25 square kilometres. This step eliminates duplicated records from different data sources and limits the number of records from repeated sampling events in the same area. We also removed records located within the land mask of the environmental data. Finally the distributions for all species were visually inspected and cross-checked with available distribution information in order to eliminate erroneous records. The amount of sample selection bias was assessed by visually comparing the spread of the occurrence records with the distribution range of the species and attributing a score ranging from 1 (low bias) to 5 (high bias).

We collected for each species taxonomic and functional group information from the World Register of Marine Species (WoRMS Editorial Board, 2016). The 'functional group' trait divides species into three groups reflecting their habitat: benthos, nekton and plankton (zooplankton and phytoplankton). For species lacking trait data in WoRMS, this information was derived from FishBase (Froese et al., 2017) and SeaLifeBase (Palomares et al., 2017) whereby all seafloor associated species were classified as benthos (i.e. sessile, reef-associated or demersal species), other free swimming species as nekton and drifting species as plankton. Additionally, we identified the latitudinal zones ('polar', 'temperate', 'tropical') for each distribution range. To do this, we checked for the presence of at least five per cent of all occurrence records of a species in each latitudinal zone of the marine ecoregions classification by Spalding et al. (2007). Lastly, species were categorized as oceanic if more than five per cent of their records are located outside the marine ecoregions. Else, species were considered as neritic.

Environmental data

The distribution records in the MarineSPEED dataset were linked to all 71 monthly and annual environmental variables for the current climate available from Bio-ORACLE (Tyberghein et al., 2012) and MARSPEC (Sbrocco & Barber, 2013) with a spatial resolution of 5 arcmin using the R package *sdmpredictors* (Bosch et al., 2016). This environmental data includes variations of sea surface temperature, salinity, bathymetry, nutrients and other predictors of marine species distributions.

Background data

Most presence-only SDM methods use background or pseudo-absence points for building models (Franklin, 2009). In order to facilitate the reproducibility of different studies using MarineSPEED we included a set of 20.000 randomly sampled background points in the benchmark dataset. We also created a second set of target-group background points by randomly sampling 20.000 points from the full set of distribution records. The latter show the same bias as the occurrence records and therefore can be used to mitigate the effect of sample selection bias on presence-only species distribution models (Phillips et al., 2009; Kramer-Schadt et al., 2013; Syfert et al., 2013).

Cross-validation splits

Cross-validation (CV) is a widespread strategy used to perform model selection while avoiding under- and overfitting models (Arlot & Celisse, 2010). We prepared CV folds for the species and background data using three different strategies. As a first strategy we partitioned the data randomly in five folds (random CV). This strategy is easy to perform but has as disadvantage that it results in an overestimated performance of the model because training and validation points selected from nearby locations will be dependent due to the effect of spatial autocorrelation (Bahn & McGill, 2007; Hijmans, 2012; Roberts et al., 2016). As CV only avoids overfitting when training samples are independent from the validation samples this generally leads to the selection of complex models with poor transferability (Arlot & Celisse, 2010; Verbruggen et al., 2013; Petitpierre et al., 2017). The second (disc-based CV) and third (grid-based CV) splitting strategies take into account the spatial nature of the data. The 5-fold disc-based strategy randomly samples a starting point and subsequently selects the nearest one fifth of all distribution records to get the first fold. Then the distribution record furthest away from the starting point is used as a new starting point and the nearest one fifth of the distribution records are included to create the second fold. This process is repeated five times until all records are assigned to a fold. For the 4-fold grid-based strategy records are split into two sets based on their longitude using a random meridian as a dividing line. Then these two halves are separately split in two equal parts using parallels. Additionally, 9-fold grid-based sets were created by using two meridians and parallels for splitting instead of one. By combining the disc- or grid-based CV strategies with the pairwise distance sampling method proposed by Hijmans (2012) to select the pseudo-absence points for the test set spatial sorting bias was eliminated and thus the effect of spatial autocorrelation on the performance evaluation supressed (Bahn & McGill, 2007; Roberts et al., 2016). In order to remove false negatives in the training sets of the spatial cross-validation sets we excluded background points from the training sets that are within 200 km of test occurrences.

Predictor relevance

In order to find out which predictors are most relevant for the set of species in MarineSPEED we ranked distribution models fitted for all combinations of predictors from multiple correlation groups. In addition, we added variation at the different steps of the model creation to assess the variability in predictor relevance under different model setups (Fig. 1).

Following the methodology from Barbet-Massin & Jetz (2014), who identified relevant predictors of bird distributions, distributions were modelled for all combinations of three, four and seven environmental predictors selected from eight correlation groups. After filtering the initial set of 68 predictors down to 19 predictors based on a Pearson product moment correlation coefficient larger than 0.95 we created correlation groups with the R package *sdmpredictors* by grouping all predictors for which some or all of the predictors have an absolute Pearson product moment correlation coefficient larger than 0.7 (Dormann et al., 2013; Barbet-Massin & Jetz, 2014). This resulted in 8 correlation groups of which 6 predictors form a group on their own (shore distance, bathymetry, SST (range), calcite, salinity, pH), 7 predictors belong to the "Chlorophyll a group", grouping chlorophyll a and diffuse attenuation (mean, minimum, maximum and/or range) related variables. The last 6 predictors form the "SST group" with variations of sea surface temperature (SST), photosynthetically active radiation (PAR), phosphate, nitrate and silicate. For a full overview of the different environmental predictors used and the correlation group they belong to we refer to Fig. 2 and to Table S1 in Appendix S3.

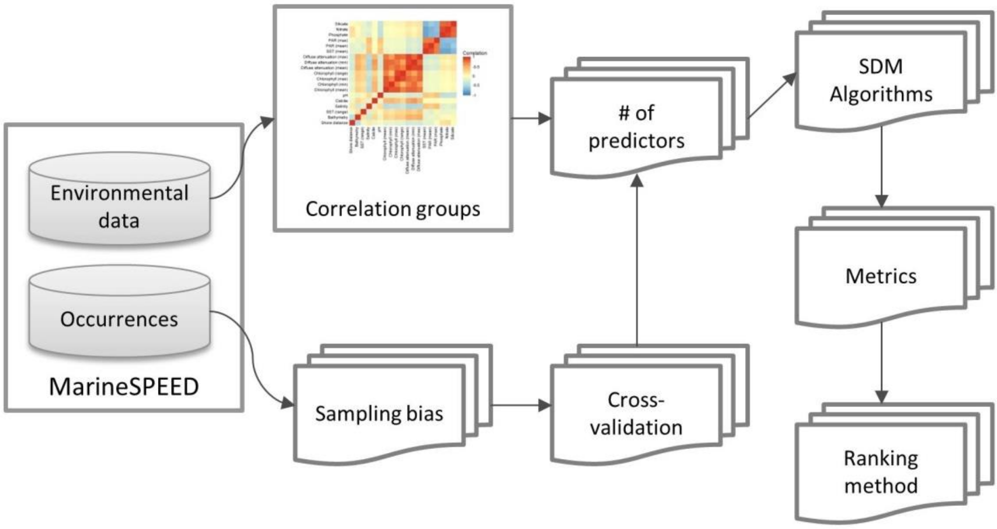


Figure 1. Overview of the predictor selection analysis and the different steps where variations were introduced. Starting from 19 environmental predictors, from Bio-ORACLE and MARSPEC, correlation groups where created. From this all possible predictor combinations were generated for models with three, four and seven predictors. After optional sample selection bias mitigation, occurrence records and background points were split in random or spatial cross-validation folds. SDMs were build using four algorithms (random forests, MaxEnt, generalized linear models and Bioclim) and evaluated using the area under the curve of the receiver operating characteristic (AUC) and the point-biserial correlation (COR). Predictors were ranked based on the performance of the models they were included in.

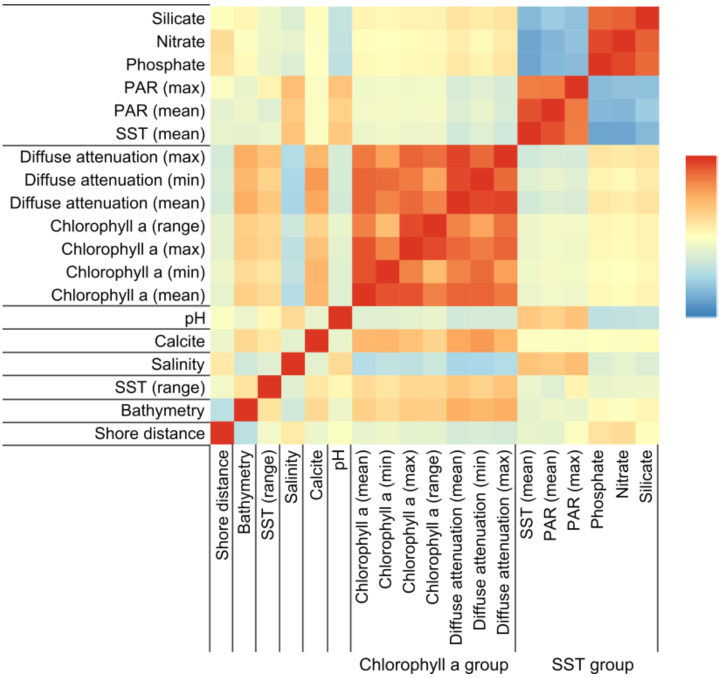


Figure 2. Correlation matrix for all environmental predictors considered for the predictor selection analysis, sorted by correlation group. Note that for creating the correlation groups, predictors are grouped when the absolute correlations between two or more members of a correlation group is more than 0.70. Red indicates a high positive correlation, yellow no correlation and blue a high negative correlation.

SDMs were fitted using four commonly used algorithms: Bioclim (Booth et al., 2014), Generalized Linear Model (GLM), Maximum Entropy modelling (Maxent, Phillips et al. 2004) and Random Forests (RF, Breiman 2001). We used the *dismo* package (Hijmans et al., 2016) in R for fitting Bioclim and MaxEnt models and the R package *randomForest* (Liaw & Wiener, 2002). For all algorithms the default settings were used and GLMs were run with only linear features.

Three variations of sample selection bias correction were performed: 1) no correction, 2) spatial thinning (50 km) with the R package *spThin* (Aiello-Lammens et al., 2015) and a target-group background (Phillips et al., 2009). Performance of the models was evaluated using random as well as spatial disc-based cross-validation. In total six million models were fitted and evaluated using the area under the receiver operating characteristics (ROC) curve (AUC) (Hanley & McNeil, 1982), and the point-biserial correlation (COR) (Zheng & Agresti, 2000; Elith et al., 2006) on the UGent High Performance Cluster.

Predictors where ranked for each model setup, evaluation metric and species combination by ranking the mean or median performance of all models a predictor was used in and by using the Rank Centrality algorithm (Negahban et al., 2017). Rank

Centrality is an iterative algorithm for rank aggregation using pairwise-wise comparisons.

Results

Benchmark data set

The MarineSPEED benchmark dataset is composed of 514 species with an original total of two million distribution records which have been filtered down on a 25 km² grid to nearly nine hundred thousand records. On a species level the median number of filtered distribution records is 506 with a minimum of 52 and a maximum of 45,469. An overview of the information on the species is available in Appendix S2.

A total of 18 different phyla are included in MarineSPEED (Fig. 3), with as most represented phyla: Chordata (245 species), Mollusca (62 species), Echinodermata (38 species), Arthropoda (36 species) and Annelida (32 species). The phylum Chordata is mostly represented by the class Actinopterygii (184 species), and to a lesser extent Elasmobranchii (20 species) and Mammalia (18 species). Marine primary producers, various groups of algae and seagrasses, are represented by 49 species from 5 phyla. When classifying species into functional groups we see that 395 species are associated with the seafloor (benthos), while 87 species are free swimming (nekton) and 32 species are planktonic. While we aimed to select species from different parts of the world a bias towards a few well-researched areas (e.g. the North-Atlantic and Australia) was unavoidable (Fig. 4). Likewise, coastal areas (442 species) are overrepresented compared to open ocean habitats (72 species). On a latitudinal scale, temperate regions are the most represented with 173 species. 91 species only occur in the tropics and 11 species in the polar regions. When considering the sample selection bias criterion we see that 59 species have a very low degree of sample selection bias (value 1), that most species have value 2 (103 species), 3 (156 species) or 4 (178). Only 18 species were assessed as having a very high degree of sample selection bias.

The predefined spatial cross-validation splits all considerably increase the distance between test points and their nearest training point as compared to random splits (Fig. S1 in Appendix S3). Examples of the various cross-validation strategies are visualised for *Didemnum maculosum* Milne Edwards and *Polycarpa aurata* Quoy & Gaimard in Fig. S2 and S3, respectively in Appendix S3.

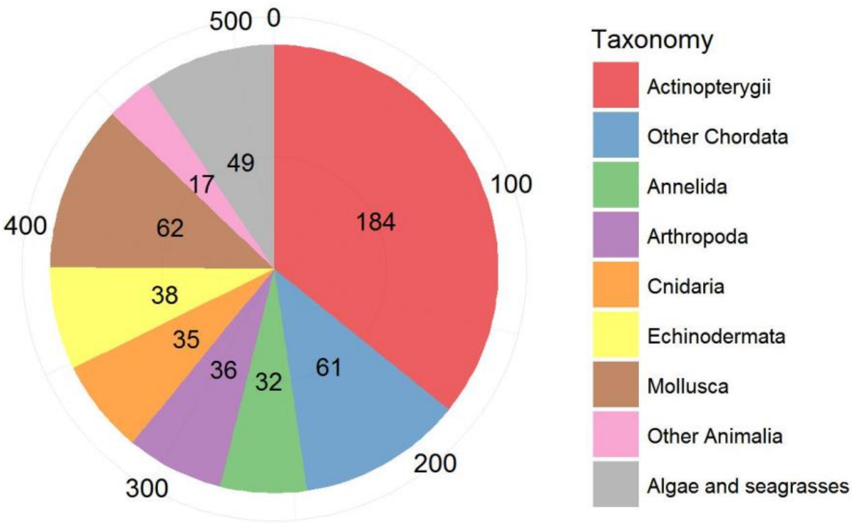


Figure 3. Taxonomic composition of the MarineSPEED dataset on level kingdom, phylum or class. For the kingdom Animalia the most abundant phylum Chordata was split up into the Actinopterygii and other Chordata, the kingdom Plantae was left as one whole and labelled as algae and seagrasses. Numbers represent the number of species in each taxonomic group.

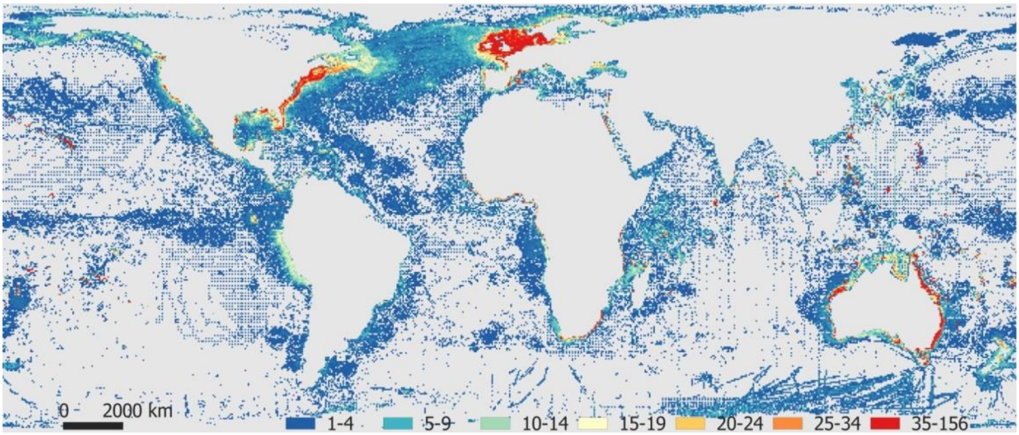


Figure 4. Map of the number of species occurring in each cell of an equal-area grid with a per cell area of 25 km² (Behrmann cylindrical equal-area projection).

Predictor relevance

A first set of analyses exploring the selection of relevant predictors (Fig. 5), highlights the importance of mean sea surface temperature (SST (mean)) as the most relevant predictor of the species distributions in the MarineSPEED benchmark dataset. This result appears robust regardless of modelling algorithms, sample selection bias correction, cross-validation, number of predictors, evaluation metrics and ranking methods. At the other end of the spectrum, calcite is apparently irrelevant as a predictor for most of the species distributions. As for the other predictors, however, there is substantial variation across species and modelling parameters.

Among the different algorithms, GLMs with linear features caused the most variation in the predictor top 5 rankings with a particularly strong effect on SST (mean) with a minimal decrease of 28% in the median percentage of species with SST (mean) in the top 5 ranking (Table 1). Conversely in GLMs bathymetry was selected at least 26% more. The difference between the two evaluation metrics AUC and COR on the other hand was fairly limited with salinity displaying the largest difference. Finally the ranking method showed very small differences between the mean and median ranking algorithm. The rank centrality algorithm consistently ranked the predictors from the “Chlorophyll a group” as less relevant, while increasing the ranking of salinity (+16%) bathymetry (+15%), pH (+13%) and shore distance (+13%).

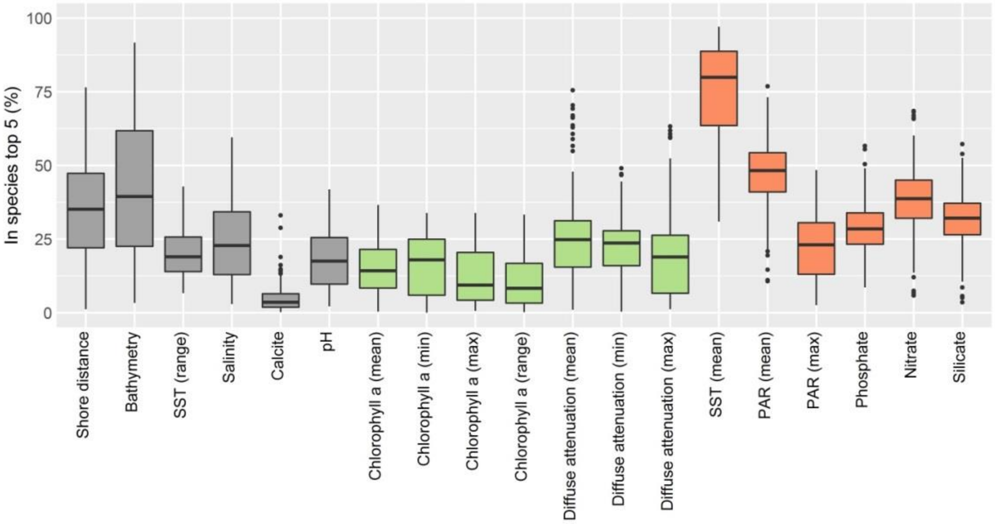


Figure 5. Percentage of species a predictor has a top 5 ranking in the different model setups. In grey are the predictors that form a correlation group on their own, in green the predictors from the “Chlorophyll a group” and in red the predictors from the “SST group”. The results are aggregated from all possible variations. For a detailed view on the different dimensions of the variations we refer to tables 1 to 3, and to the following plots in Appendix S3: modelling algorithms (Fig. S4), evaluation metrics (Fig. S5), ranking methods (Fig. S6), cross-validation strategies (Fig. S7), predictor counts (Fig. S8), sampling bias mitigation methods (Fig. S9), cross-validation folds (Fig. S10) and taxonomic groups (Fig. S11).

When comparing the results of CV splitting strategies, number of predictors, sampling bias mitigation and fold number (Table 2), we can conclude that the number of predictors allowed in the model has the largest effect. Increasing the number of allowed predictors from 3 to 7 causes a decline in the relevance of bathymetry (-31%) and shore distance (-26%) while increasing the relevance of PAR (max) (+17%), diffuse attenuation (max) (+14%) and chlorophyll a (max and range) (+13%). The second largest effect is caused by using a target-group background in order to mitigate the effect of sampling bias on SDMs with a decrease of 25% for bathymetry and 15% for shore distance and an increase of 12% for nitrate. When using the disc-based CV strategy the relevance of SST (mean) and salinity decreased with 19 and 10%, respectively. Using the second fold instead of the first fold, which was only performed for the random CV strategy, only yielded small differences in the top 5 predictors of the species.

While the relevance of most predictors, is similar across taxonomic groups, some predictors exhibit large differences (Table 3). This is especially the case for shore distance, bathymetry and SST (range) with differences between the minimum and maximum of 55, 40 and 33%, respectively. Despite these overall patterns in the median ranking values we see that the spread of the predictor relevance within taxonomic groups is large (Fig. S11).

Table 4 presents the results related to the different traits of the species: functional group, neritic versus oceanic zone, ecoregion and sampling bias. Regarding the functional group some clear trends are visible whereby shore distance, bathymetry and to a lesser extent PAR (mean) are comparatively more relevant predictors for benthic species distributions, less relevant for nekton and least relevant for plankton. For mean and minimum diffuse attenuation we notice an inverse trend with a higher relevance for plankton in comparison to nekton and benthos. With respect to the zone trait we see that shore distance (-21%) and bathymetry (-14%) are less relevant for oceanic species, while phosphate (+15%), nitrate (+13%) and silicate (+15%) are more relevant. The results from the ecoregion trait show clear differences in predictor relevance for multiple predictors. For some predictors such as SST (range), nitrate and phosphate the relevance for temperate species clearly deviates from that for polar and tropical species. The predictor relevance for the different levels of sampling bias are all very similar. For boxplots of the relevance of the predictors for the different variations in model setup, taxonomic groups and traits we refer to Figs. S4 to S15 in Appendix S3.

Table 1. Median percentage of species for which a predictor has a top 5 ranking for the different setup variations that have been calculated for all models. First column shows the results for all models, the next four columns show the results for the different modelling algorithms, the next two columns show the breakdown for the evaluation metrics used. The last three columns show the results for the ranking methods.

			Algorithm				Metric		Ranking method		
Group	Predictor	**All**	Bioclim	GLM	MaxEnt	RF	AUC	COR	Centrality	Mean	Median
	Shore distance	**35**	29	22	39	40	36	34	44	31	27
	Bathymetry	**39**	45	71	36	19	40	37	52	37	33
	SST (range)	**19**	14	24	19	18	18	19	26	16	16
	Salinity	**23**	16	15	25	37	18	26	33	17	16
	Calcite	**4**	4	5	3	3	3	4	6	2	3
	pH	**18**	8	24	14	23	17	18	26	12	13
Chlorophyll a group	Chlorophyll a (mean)	**14**	18	8	14	17	15	13	9	16	18
	Chlorophyll a (min)	**18**	22	4	21	21	17	18	6	22	22
	Chlorophyll a (max)	**9**	15	6	11	15	10	9	5	17	19
	Chlorophyll a (range)	**8**	11	7	9	13	8	9	3	13	15
	Diffuse attenuation (mean)	**25**	21	44	24	24	23	26	10	27	27
	Diffuse attenuation (min)	**24**	22	30	22	21	22	23	9	25	25
	Diffuse attenuation (max)	**19**	12	37	10	16	18	19	7	23	23
SST Group	SST (mean)	**80**	79	51	89	86	79	78	79	79	78
	PAR (mean)	**48**	53	49	48	41	46	49	51	46	46
	PAR (max)	**23**	22	30	20	15	20	24	26	17	22
	Phosphate	**28**	32	23	27	32	29	27	33	26	26
	Nitrate	**39**	41	31	41	44	41	33	41	38	37
	Silicate	**32**	27	29	32	36	32	31	36	29	31

Table 2. Median percentage of species for which a predictor has a top 5 ranking for the different setup variations that have been calculated for a subset of the models. In this table only results from setups that have been done for both options are shown. First column shows the results for all models, followed by the results for the 5-fold random and disc-based spatial cross-validation splitting strategies, the breakdown for the number of predictors used in the models, the impact of sampling bias mitigation techniques and the results for the first and the second fold.

			CV splitting strategy		Predictor count			Sampling bias mitigation			Fold number	
Group	Predictor	**All**	Disc	Random	3	4	7	None	spThin	Target-group	1	2
	Shore distance	**35**	35	30	56	55	30	30	27	12	30	35
	Bathymetry	**39**	42	34	65	62	34	34	33	8	34	37
	SST (range)	**19**	15	21	19	24	21	21	18	18	21	11
	Salinity	**23**	13	23	22	28	23	23	23	28	23	20
	Calcite	**4**	9	3	3	3	3	3	3	2	3	3
	pH	**18**	11	17	17	17	17	17	16	27	17	16
Chlorophyll a group	Chlorophyll a (mean)	**14**	15	18	12	12	18	18	15	22	18	19
	Chlorophyll a (min)	**18**	21	17	18	15	17	17	19	16	17	17
	Chlorophyll a (max)	**9**	16	16	3	4	16	16	17	19	16	14
	Chlorophyll a (range)	**8**	17	15	2	4	15	15	15	14	15	12
	Diffuse attenuation (mean)	**25**	18	26	24	24	26	26	26	28	26	27
	Diffuse attenuation (min)	**24**	24	24	25	21	24	24	25	19	24	24
	Diffuse attenuation (max)	**19**	18	20	6	8	20	20	21	25	20	22
SST Group	SST (mean)	**80**	59	78	85	84	78	78	80	85	78	76
	PAR (mean)	**48**	46	50	37	47	50	50	51	59	50	49
	PAR (max)	**23**	34	25	8	12	25	25	25	25	25	23
	Phosphate	**28**	32	26	28	27	26	26	27	28	26	30
	Nitrate	**39**	36	33	42	38	33	33	34	46	33	44
	Silicate	**32**	36	35	29	29	35	35	32	29	35	29

Table 3. Median percentage of species for which a predictor has a top 5 ranking for the different setup variations that have been calculated for all models and for some taxonomic groups. Within the class Chordata and within the kingdom Animalia taxonomic groups with few species were left out of this comparison.

			Chordata	Other Animalia					Plantae
Group	Predictor	**All**	Actinopterygii	Annelida	Arthropoda	Cnidaria	Echinodermata	Mollusca	Algae and seagrasses
	Shore distance	**35**	42	16	11	66	32	29	44
	Bathymetry	**39**	49	42	33	54	51	31	14
	SST (range)	**19**	14	42	36	9	13	19	14
	Salinity	**23**	18	16	19	11	21	25	31
	Calcite	**4**	2	3	6	3	8	3	4
	pH	**18**	19	6	11	11	13	21	18
Chlorophyll a group	Chlorophyll a (mean)	**14**	11	9	17	9	16	15	18
	Chlorophyll a (min)	**18**	15	16	19	9	16	15	20
	Chlorophyll a (max)	**9**	9	5	11	6	13	10	10
	Chlorophyll a (range)	**8**	8	3	8	6	11	10	8
	Diffuse attenuation (mean)	**25**	17	31	33	11	18	27	35
	Diffuse attenuation (min)	**24**	17	31	25	9	21	23	29
	Diffuse attenuation (max)	**19**	19	9	17	14	21	18	18
SST Group	SST (mean)	**80**	81	81	72	83	71	69	76
	PAR (mean)	**48**	52	41	42	57	45	47	33
	PAR (max)	**23**	18	28	33	9	21	24	18
	Phosphate	**28**	29	19	33	29	26	26	23
	Nitrate	**39**	43	22	36	47	34	35	24
	Silicate	**32**	25	41	36	14	29	35	39

Table 4. Median percentage of species for which a predictor has a top 5 ranking for the different setup variations that have been calculated for all models and traits. For the functional group trait, benthos includes all seafloor associated species, including demersal and reef-associated species; nekton includes all actively swimming pelagic species and plankton are all species unable to swim against a current. The neritic and oceanic zones were defined based on the ecoregion classification by Spalding (2007) whereby species having 5% or more of their distribution records outside of ecoregions are classified as oceanic. Species are a member of an ecoregion when at least 5% of its distribution records are situated in a polar, temperate or tropical ecoregion. Sampling bias was visually assessed from 1 (low bias) to 5 (high bias).

			Functional group			Zone		Ecoregion			Sampling bias				
Group	Predictor	**All**	Benthos	Nekton	Plankton	Neritic	Oceanic	Polar	Temperate	Tropical	1	2	3	4	5
	Shore distance	**35**	39	24	13	38	17	13	25	49	29	28	42	35	28
	Bathymetry	**39**	44	26	13	40	26	39	25	60	22	30	47	44	22
	SST (range)	**19**	17	22	28	19	19	13	28	5	19	28	15	18	17
	Salinity	**23**	20	24	22	22	18	26	27	14	31	28	17	17	28
	Calcite	**4**	3	2	3	3	1	0	3	2	5	4	3	3	6
	pH	**18**	17	17	6	19	7	4	18	14	26	22	13	13	14
Chlorophyll a group	Chlorophyll (mean)	**14**	12	16	19	13	17	17	17	10	14	14	13	13	17
	Chlorophyll (min)	**18**	16	19	17	17	17	13	20	11	20	20	14	15	22
	Chlorophyll (max)	**9**	8	11	9	9	10	4	10	8	12	9	8	10	11
	Chlorophyll (range)	**8**	8	8	9	8	10	4	8	8	10	8	7	8	11
	Diffuse attenuation (mean)	**25**	22	30	44	24	25	30	33	12	24	28	21	24	28
	Diffuse attenuation (min)	**24**	21	27	34	23	24	22	31	10	24	24	21	22	22
	Diffuse attenuation (max)	**19**	18	16	16	19	15	13	16	20	17	15	19	19	22
SST Group	SST (mean)	**80**	79	77	78	79	77	74	74	86	63	74	85	81	72
	PAR (mean)	**48**	49	45	34	48	44	26	42	56	42	46	53	46	42
	PAR (max)	**23**	21	29	19	22	19	17	25	14	25	26	19	20	22
	Phosphate	**28**	27	25	34	25	40	48	21	34	25	22	29	29	33
	Nitrate	**39**	39	33	31	36	49	57	27	49	30	28	41	43	28
	Silicate	**32**	28	39	41	28	43	52	38	16	37	36	23	31	33

Data access

While distribution maps for all species can be consulted and all data is downloadable in an R Shiny interface (Chang et al., 2016) at <http://marinespeed.org>, we opted to also create the *marinespeed* R package allowing for an easy usage of the data (Table 4). The first step, after installation from CRAN and loading the library, is to run the function ‘list_species’ which returns the scientific names and WoRMS identifiers for all species. Additional information on the taxonomy, sampling bias estimate and latitudinal zones can be viewed using the ‘species_info’ function. In order to run a function for all species either the ‘lapply_species’ or the ‘lapply_species_kfold’ function can be used. Alternatively, if you only need data for specific species, the ‘get_occurrences’ and ‘get_fold_data’ methods can be used. On top of this other lower level functions for loading background data and creating cross-validation splits are also available.

Table 4. Overview of the most important functions in the *marinespeed* R package. Lower level functions for accessing occurrences, background data and creating cross-validation folds are also available.

Function	Description
list_species	Get the list of scientific names and WoRMS identifiers for all species.
species_info	Additional species information.
lapply_species	Execute a function for all distribution records for multiple species.
lapply_kfold_species	Execute a function for one or more pre-made CV folds for multiple species.

Discussion

Species distribution modelling is widely used to identify areas that are ecologically suitable for the presence of species under past, current and future climates. Most studies concentrate, however, on terrestrial environments, while marine species distribution modelling kicked off comparatively late (Robinson et al., 2011). A direct consequence of the relative scarcity of marine SDM studies is that most of the methodological progress in SDM is biased towards terrestrial studies, despite marine environments being significantly different with respect to the ecological factors that control distributions and their spatio-temporal variation. These differences raise questions with respect to the environmental predictor relevance and the effects of model algorithms and settings on predictor relevance. By fitting presence-only SDMs for all combinations of predictors from different correlation groups, we assessed the predictor relevance and the variation therein for marine species distributions. To this end, we created a benchmark dataset (MarineSPEED) which bundles marine species distributions of 514 taxa and associated environmental variables.

Relevant predictors

SST (mean) is the most relevant predictor of global marine species distributions, regardless of model algorithms and parameter settings. This result confirms the importance of temperature for species distributions identified in the meta-analysis by Bradie & Leung (2016) and its importance for the distribution of birds (Barbet-Massin & Jetz, 2014). Moreover the importance of SST as a predictor in marine ecology was previously confirmed for marine species richness (Tittensor et al., 2010) and biogeographic structure of marine benthic fauna (Belanger et al., 2012). While bathymetry and shore distance also appear to be very relevant, there is considerable variance in the results, which might be because they are distal environmental predictors (Austin, 2002). In contrast to previous results (Nyström Sandman et al., 2013; Bradie & Leung, 2016) bathymetry was not the most important predictor, which can be explained by the global scale of our study. The importance of bathymetry has been shown to decrease with increased geographical scale (Nyström Sandman et al., 2013). Moreover the relevance of bathymetry is strongly linked to the species taxonomy (see Table 3 and 4 and Fig. S11-S14). At the other end of the spectrum, calcite is rarely selected as a meaningful predictor. The irrelevance of calcite is consistent with the fact that only one study in the meta-analysis by Bradie & Leung (2016) used calcite as a predictor. The remaining predictors are on average less part of the best scoring models, reflecting an overall reduced relevance toward predicting species distributions.

Despite this general trend the variance in predictor relevance is relatively high across model algorithms and settings.

The high variance when using different modelling algorithms is consistent with the results by Bucklin et al. (2015) who also demonstrated a significant interaction between predictor set and modelling algorithm. Especially predictor selection under GLM deviates from the other algorithms. Linear GLM-based models do not capture the relevance of SST (mean) very well. The lower relevance of SST in GLM models indicates that the global distribution of marine species is inadequately modelled by a linear relationship. Potentially, this effect can be mitigated by including polynomial features, an option which was not explored in the current analyses. In MaxEnt, with automatic selection of feature complexity and therefore yielding complex models, the relevance of SST (mean) is consistently high and displaying hardly any variation.

We expect that decreasing the complexity of the features fitted by MaxEnt will result in models more similar to GLM-based models. As for the other three algorithms, predictor selection seems to be largely consistent, echoing results of Barbet-Massin & Jetz (2014).

We also compared the predictor relevance under two different evaluation measures, AUC and COR, respectively. Although AUC, as an absolute measure for model performance, has been criticized earlier (Lobo et al., 2010) its use is warranted here as we only compared relative AUC values and only modelled in a fixed geographical extent. Both AUC, which measures the ability to discern presences from background data, and COR, which provides a measure for the calibration of the model showed very similar predictor rankings. This similarity is indicative for the generalizability of the results across model evaluation metrics.

Likewise, for most predictors the ranking method used did not affect the predictor relevance. The rank centrality method consistently gave a lower ranking to all predictors from the "Chlorophyll a group". As ranking from pairwise comparisons is an active research field, a future study comparing the rank centrality algorithm with other recent ranking methods such as spectral ranking (Fogel et al., 2016), sync rank (Cucuringu, 2016) and Microsoft's TrueSkill method (Herbrich et al., 2006) could lead to additional insights on the impact of the ranking algorithm on the predictor relevance.

The impact of cross-validation strategies was assessed by using spatial disc-based and random sampling of training and testing sets. Using a spatial instead of a random data splitting strategy in combination with the removal of spatial sorting bias resulted in a lower relevance of SST (mean). This can be attributed to two different factors: (1) extrapolation and (2) scale effects. Firstly, the spatial data splits sometimes causes a restriction in the predictor space, which leads to extrapolation (Roberts et al., 2016). With SST being in general the most relevant, extrapolation outside of its range will lead to low evaluation scores and therefore a lower ranking. On the other hand, due to the pairwise selection of test pseudo-absences at a similar distance to the test points as the distance between the test points and their nearest training point, the mean distance to evaluation background points decreases causing a scale effect. These results confirm that SST is especially relevant on a global scale but less so on a smaller scale (Nyström Sandman et al., 2013).

Restricting the number of predictors included in a model directly influences the relevance of the predictors. For most marine species the relevance of bathymetry and shore distance diminishes when more predictors are included in the model.

These predictors are only distally related to the suitability of an environment for species distributions and therefore the potential choice of more proximate predictors will result in their lower relevance in predictor-rich models. Inversely predictors from the "Chlorophyll a group" are selected more, suggesting that if combined with some of the predictors from the other correlation groups they provide a better explanation of the species distribution then bathymetry and shore distance do.

Unlike the effect of spatial thinning, using a target-group background resulted in large differences in predictor relevance. As most of the species occurrence records are located along the coast, the target-group background, which is a subsample of it, is expected to have the same bias resulting in a lower relevance of shore distance and bathymetry. These results confirm the importance of background selection on SDMs (Chefaoui & Lobo 2008; Phillips et al. 2009; VanDerWal et al. 2009; Barbet-Massin et al. 2012; Acevedo et al. 2012; Smith 2013; Senay et al. 2013). It is therefore recommended to investigate the impact of alternative pseudo-absence selection methods in future studies. Note that in general it is advised to create a species specific target-group with occurrence records from the same sampling campaign(s) and/or from similar species, reflecting the sampling bias of the species modelled (Phillips et al., 2009).

In this study we explored the impact of several parameter settings on predictor selection, however the potential analyses are by no means exhaustive. For example the regularisation parameter and the complexity of the features in MaxEnt, the number of trees fitted in random forests and the usage of polynomial features in GLM were kept constant or were not explored. It is likely that applying species-specific tuning of the algorithms will not only impact model performance but also affect the predictor selection (Anderson & Gonzalez, 2011; Merow et al., 2014).

From a species perspective we noted that the taxonomy and the traits of a species have an influence on the relevance of predictors. The overarching pattern of predictor relevance holds up across traits, but some marked differences in predictor relevance were found for shore distance and bathymetry and to a lesser extent for diffuse attenuation, phosphate, nitrate and silicate. To some extent these differences are intuitive. For example, subdividing the taxa between oceanic and neritic species results in a higher relevance of shore distance for neritic species. Likewise, SST range is less relevant for tropical and polar species, because low and high latitudes typically exhibit very little annual sea surface temperature fluctuations compared to mid-latitudes. Despite some pronounced differences across traits, trends for inorganic nutrients (nitrate, phosphate, silicate) are less easily explained.

Benchmark dataset

Inspired by the widespread use of benchmark datasets in machine learning and other computational fields we set out to create MarineSPEED. Although a series of papers was published using the same set of 226 terrestrial species (e.g. Elith et al. 2006; Guisan et al. 2007; Phillips et al. 2009; Hijmans 2012) most studies discussing new methods related to SDM use a small set of different species. Moreover while the resulting algorithm and methods are regularly made available through ready to use R packages or desktop programs, the species distribution records used in these studies often are not. With the release of MarineSPEED and its associated R package researchers can download all occurrences, background records and cross-validation data sets.

The marine character of the dataset is ideally suited for the study of methodological issues and parameterizations for distribution modelling of non-terrestrial species. This is necessary as the marine environment poses its own challenges for SDM (Kaschner et al., 2006; MacLeod et al., 2008; Dambach & Rödder, 2011; Robinson et al., 2011; Bentlage et al., 2013). Species distribution records from public databases contain a combination of opportunistic records and systematic sampling campaigns. They show large biases in amount and location of occurrences where the coastal areas are often more intensely sampled than offshore areas. The lower detectability of marine species in combination with the wide extent of the marine environment leads to false absences and a general lack of distribution records in comparison to the real world range extent of marine species. MacLeod et al. (2008) found that in contrast to the terrestrial environment, presence-absence methods don't perform better than presence-only methods in the marine environment. Although absences are rarely reported for marine species and not included in MarineSPEED, this study could be confirmed by using estimated absence data for species included in systematic surveys in OBIS (Coro et al. 2016).

Applications

Combining the marinespeed R package with one of the numerous SDM packages like *BIOMOD2*, *dismo*, *sdm* or *zoon*, other machine learning packages like *caret*, *gbm*, *randomForest* or *xgboost* and the general R ecosystem allows for numerous applications.

While several papers have compared the performance of SDM algorithms (e.g. Elith et al. 2006; Tsoar et al. 2007; Meynard & Quinn 2007; Liu et al. 2011; Lorena et al. 2011), new SDM modelling algorithms are regularly released (e.g. MaxLike (Royle et al., 2012), Plateau (Brewer et al., 2016), GRaF (Golding & Purse, 2016)). Consistent usage of MarineSPEED to explore the performance of modelling algorithms would allow for a direct comparison of the strengths and weaknesses of them. On top of this, SDM algorithms benefit from species-specific parameter settings (Anderson & Gonzalez, 2011; Merow et al., 2013; Shcheglovitova & Anderson, 2013) but useful ranges for the different parameters are unknown for these newer modelling algorithms.

Over the years, numerous studies have been published on methods for correcting sample selection bias (e.g. Dudík et al. 2005; Phillips et al. 2009; Boria et al. 2014; Varela et al. 2014; Barnes et al. 2014; Fernández & Nakamura 2015; Aiello-Lammens et al. 2015; Ranc et al. 2016) and selecting pseudo-absence records (e.g. Wisz & Guisan 2009; Lobo & Tognelli 2011; Barbet-Massin et al. 2012; Acevedo et al. 2012; Senay et al. 2013; Assis et al. 2015). Comparing these techniques with MarineSPEED can result in guidelines for sampling bias mitigation and pseudo-absence selection in the marine environment.

Next to the availability of marine species with environmental data and traits we expect that the *marinespeed* R package, with its implementation of cross-validation methods, to be a useful tool for SDM. Installation instructions, data downloads and species information can be found at <http://marinespeed.org/>.

Acknowledgements

The research was carried out with financial support from the ERANET INVASIVES project (EU FP7 SEAS-ERA/INVASIVES SD/ER/010) and financial, data & infrastructure support provided by VLIZ as part of the Flemish contribution to the LifeWatch ESFRI. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation and the Flemish Government – department EWI. We thank An Verfaillie, Taban Mestdag and Sara Martinez for contributing to the creation of the dataset. We further thank Bart Vanhoorne and Leen Vandepitte from WoRMS and Pieter Provoost from OBIS for providing support.

Data accessibility

The benchmark data can be downloaded from <http://marinespeed.org/>. The release version of the R package is on CRAN and the latest development version can be found at <https://github.com/lifewatch/marinespeed>.

Supporting information

Appendix S1

List of OBIS and GBIF datasets used for compiling MarineSPEED. Available at: http://www.phycology.ugent.be/research/marinespeed/MS_AppendixS1.docx.

Appendix S2

List of species included in MarineSPEED with their taxonomy, sampling bias, ecoregions and SST statistics. Available at: http://www.phycology.ugent.be/research/marinespeed/MS_AppendixS2.xlsx.

Appendix S3

Setup

Table S1. Overview of the different predictors used in the predictor selection analysis. The first column is the layer code used by the *sdmpredictors* R package to identify a predictor, the second column is the dataset the predictor was found in, the description column gives a short description of the predictor and the correlation groups column gives an indication of the correlation group a predictor belongs to.

Layer code	Dataset	Description	Correlation group
BO_chlomax	Bio-ORACLE	Chlorophyll a (maximum)	Chlorophyll a group
BO_chlomean	Bio-ORACLE	Chlorophyll a (mean)	Chlorophyll a group
BO_chlomin	Bio-ORACLE	Chlorophyll a (minimum)	Chlorophyll a group
BO_chlorange	Bio-ORACLE	Chlorophyll a (range)	Chlorophyll a group
BO_damax	Bio-ORACLE	Diffuse attenuation coefficient at 490 nm (maximum)	Chlorophyll a group
BO_damean	Bio-ORACLE	Diffuse attenuation coefficient at 490 nm (mean)	Chlorophyll a group
BO_damin	Bio-ORACLE	Diffuse attenuation coefficient at 490 nm (minimum)	Chlorophyll a group
BO_nitrate	Bio-ORACLE	Nitrate	SST group
BO_parmax	Bio-ORACLE	Photosynthetically available radiation (maximum)	SST group
BO_parmean	Bio-ORACLE	Photosynthetically available radiation (mean)	SST group
BO_phosphate	Bio-ORACLE	Phosphate	SST group
BO_silicate	Bio-ORACLE	Silicate	SST group
BO_sstmean	Bio-ORACLE	Sea surface temperature (mean)	SST group
BO_calcite	Bio-ORACLE	Calcite	Calcite
BO_ph	Bio-ORACLE	pH	pH
BO_salinity	Bio-ORACLE	Salinity	Salinity
BO_sstrange	Bio-ORACLE	Sea surface temperature (range)	SST range
MS_bathy_5m	MARSPEC	Bathymetry	Bathymetry group
MS_biogeo05_dist_shore_5m	MARSPEC	Distance to the shoreline	Shore distance

Table S2. Overview of all different setups for which models have been fitted for all combinations of predictors. Models for all species where build for all combinations of 3, 4 or 7 predictors using the random or disc-based splitting strategy to create the cross-validation (CV) data and the first or second fold from the 5-fold random cross-validation dataset. The last variation in setups is whether any and which sample selection bias correction method is used. For each predictor count we get a different total number of predictor combinations resulting in the calculation of a different number of models as models where fitted for all 514 species using 4 different SDM algorithms (bioclim, GLM, MaxEnt and random forests).

Predictor count	CV splitting strategy	Fold number	Sampling bias mitigation	Number of combinations	Number of models
3	Random	1	None	467	960,152
4	Random	1	None	905	1860,680
7	Disc-based	1	None	265	544,840
7	Random	1	None	265	544,840
7	Random	2	None	265	544,840
7	Random	1	spThin	265	544,840
7	Random	1	Targetgroup	265	544,840

Cross-validation splits

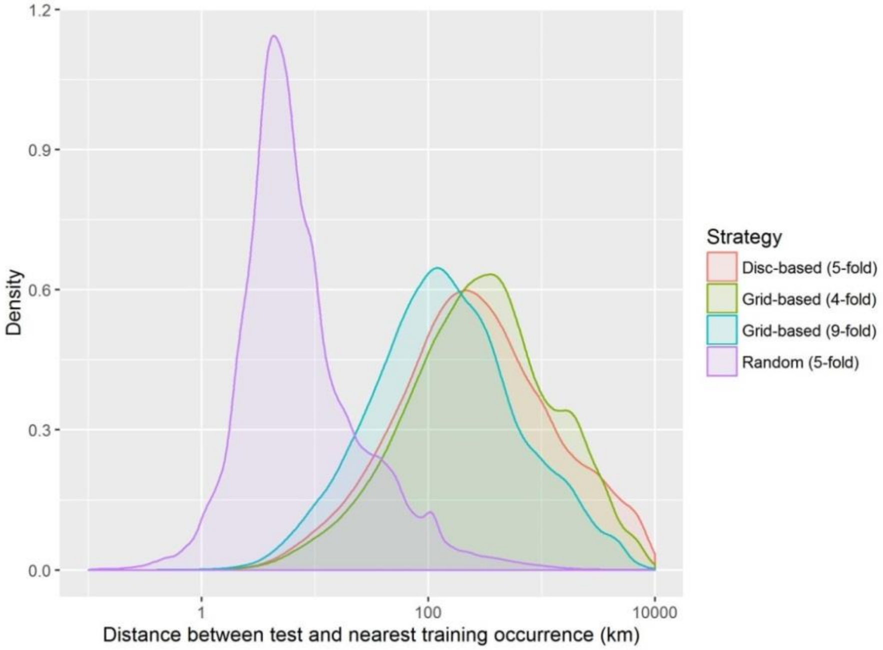
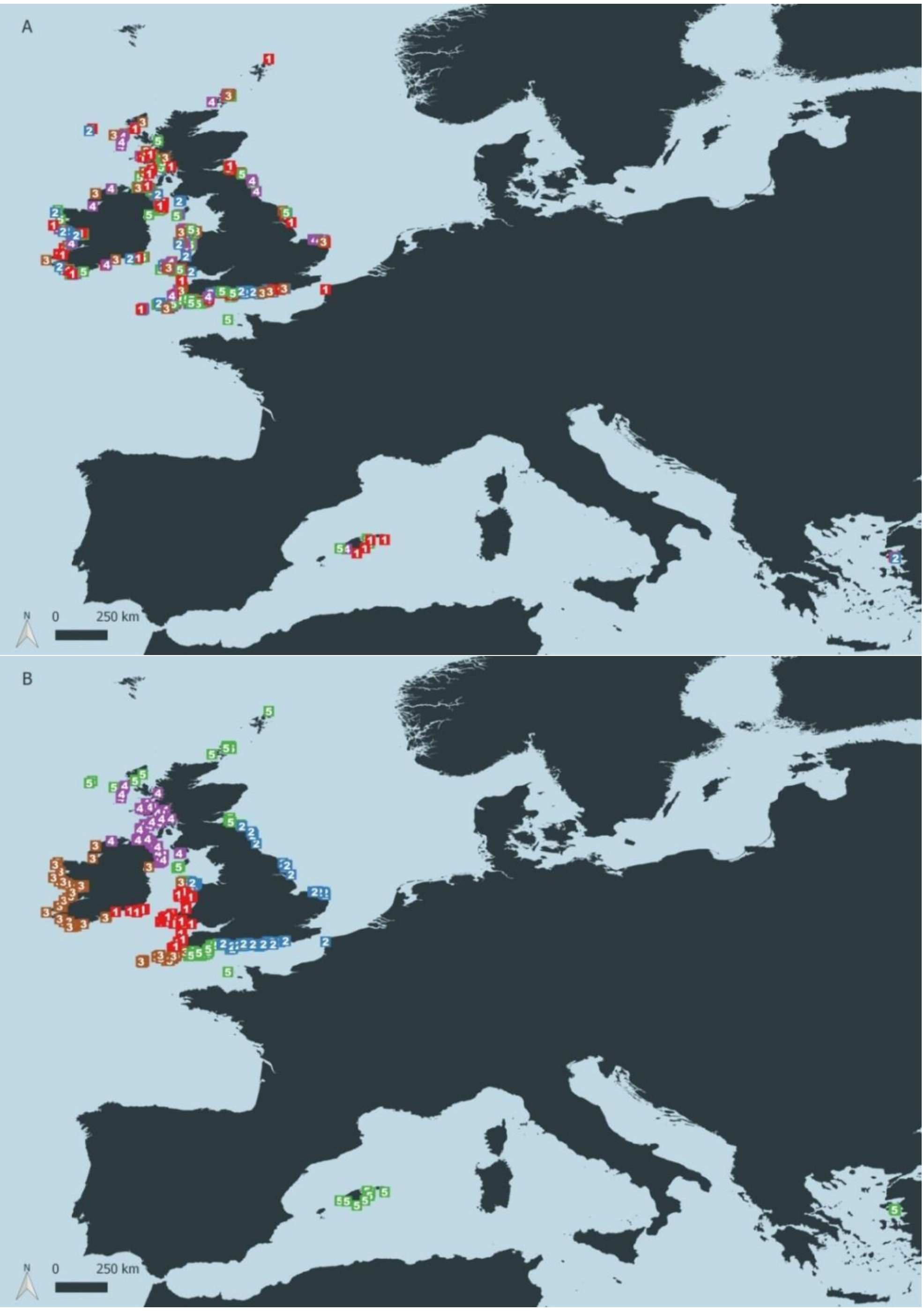


Figure S1. Density plot for the distance, on a log scale, between each test point and the nearest training occurrence for all folds of the four cross-validation splitting strategies with the 5-fold disc-based strategy in orange, the 4-fold grid-based strategy in green, the 9-fold grid-based strategy in blue and the 5-fold random strategy in purple.


A
N
0
250 km
B
N
0
250 km

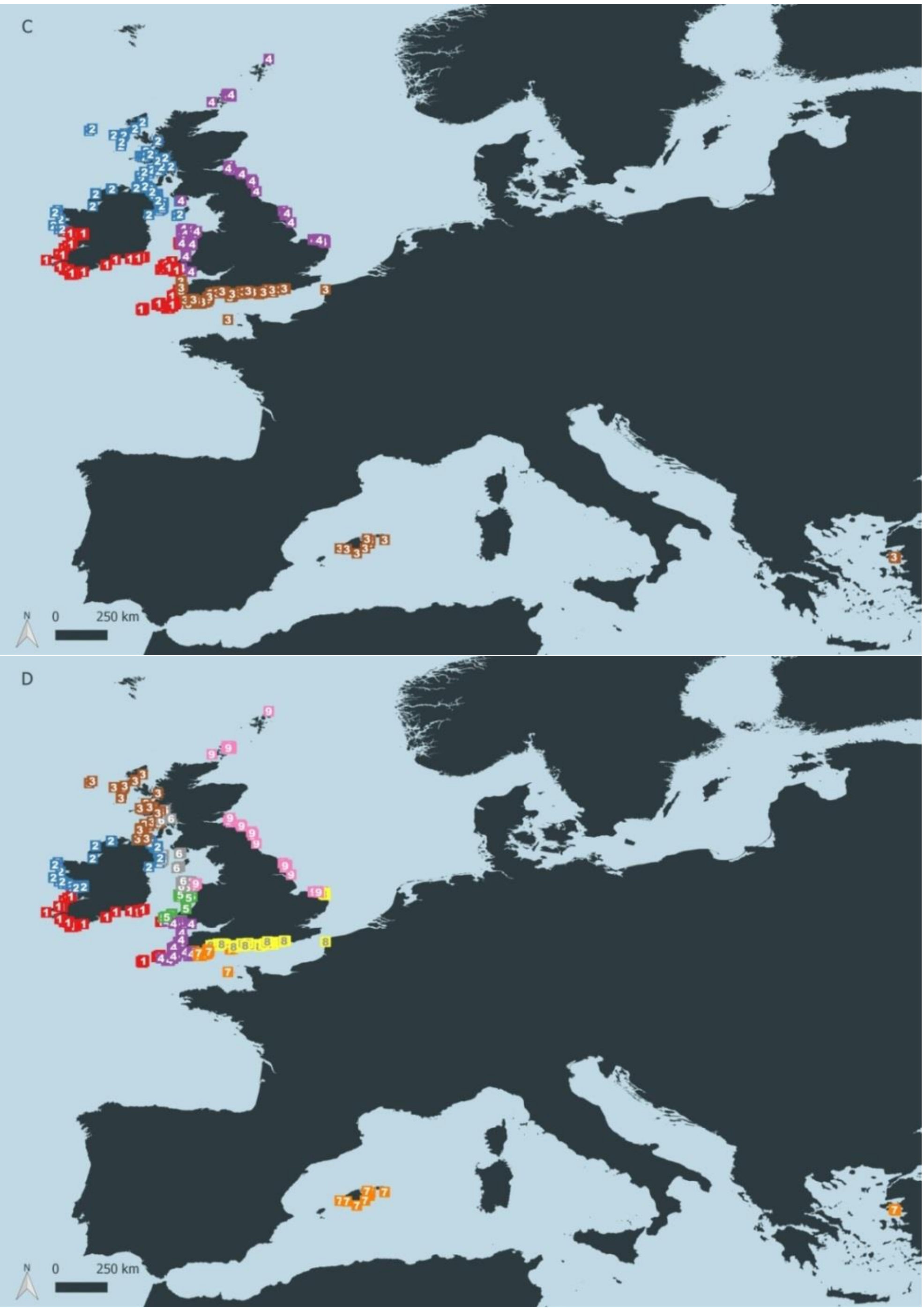
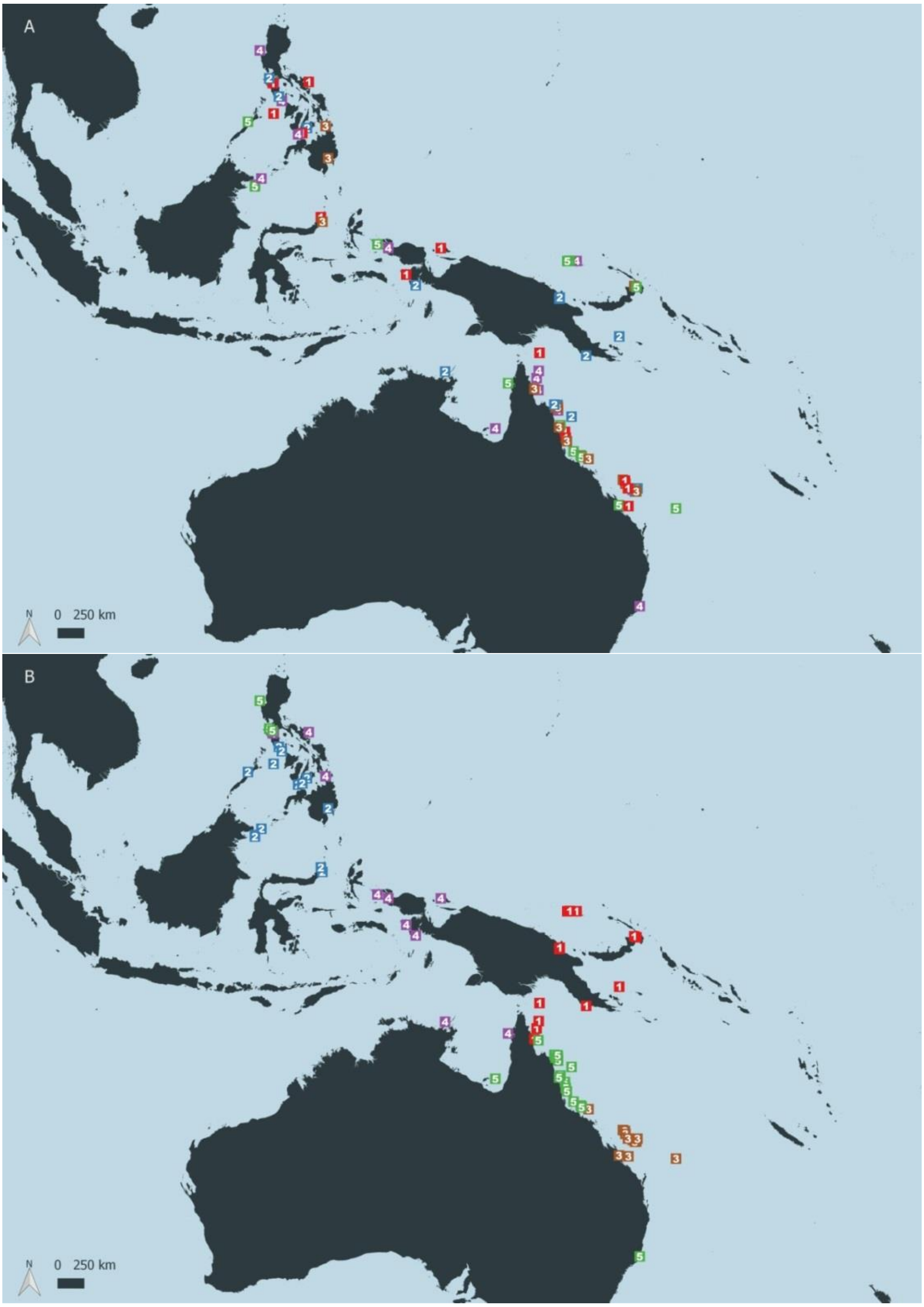


Figure S2. The cross-validation (CV) splits for the species *Didemnum maculosum* Milne Edwards using different methods: 5-fold random (A), 5-fold disc-based (B), 4-fold grid-based (C) and 9-fold grid-based (D). The different folds are numbered and coloured in the map (red=1, blue=2, brown=3, purple=4, green=5, grey=6, orange=7, yellow=8 and pink=9).


A
N
0 250 km
B
N
0 250 km

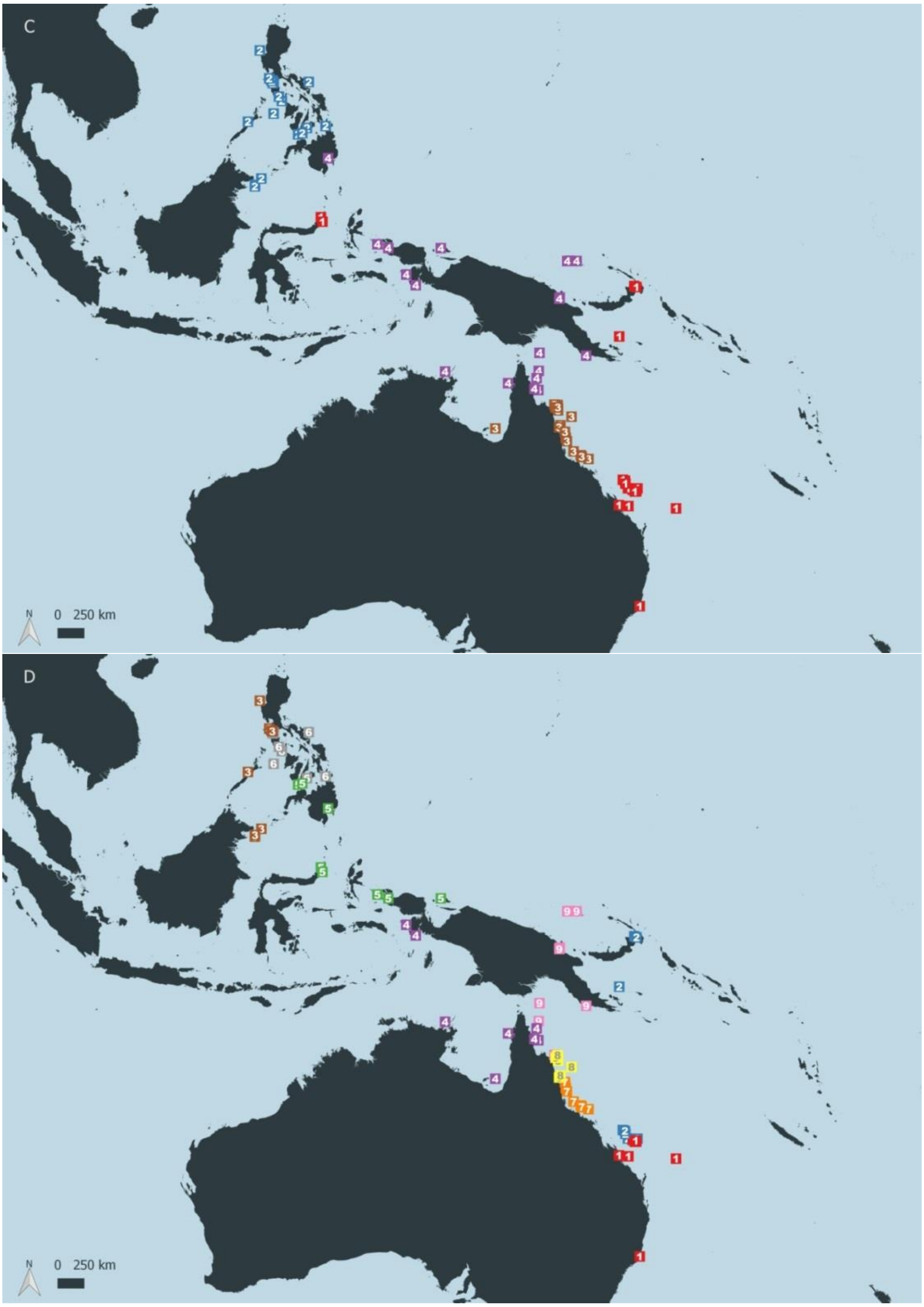


Figure S3. The cross-validation (CV) splits for the species *Polycarpa aurata* Quoy & Gaimard using different methods: 5-fold random (A), 5-fold disc-based (B), 4-fold grid-based (C) and 9-fold grid-based (D). The different folds are numbered and coloured in the map (red=1, blue=2, brown=3, purple=4, green=5, grey=6, orange=7, yellow=8 and pink=9).

Predictor relevance boxplots

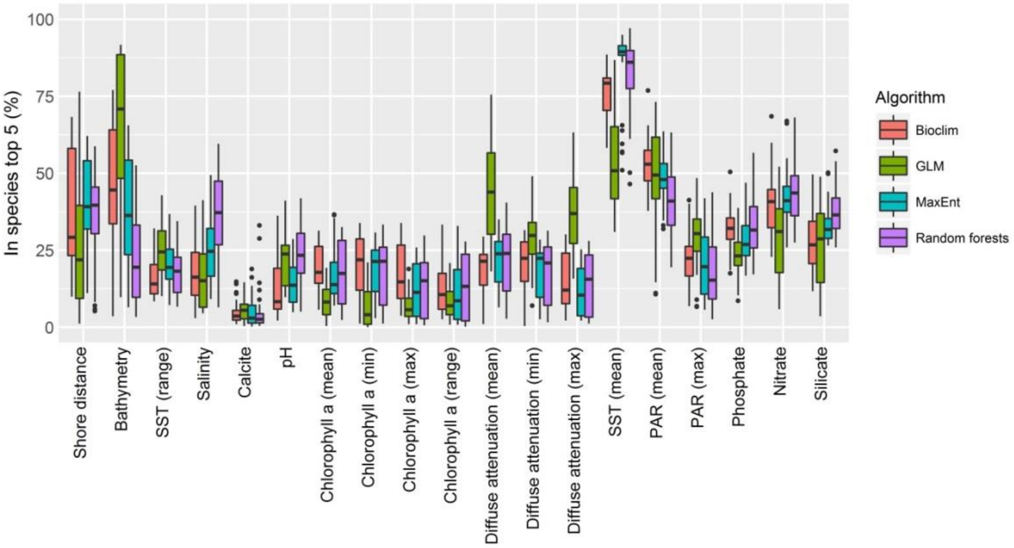


Figure S4. Percentage of species a predictor has a top 5 ranking in the different model setups for the different algorithms: bioclim (red), GLM (green), MaxEnt (blue) and random forests (purple).

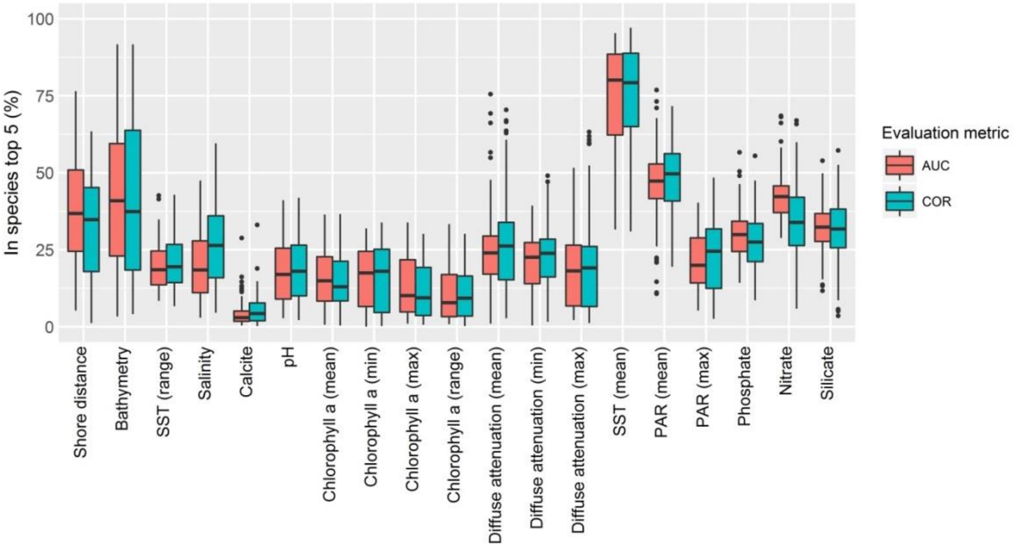


Figure S5. Percentage of species a predictor has a top 5 ranking in the different model setups for the two evaluation metrics: area under the receiver operating characteristic curve (AUC, red) and the point-biserial correlation (COR, blue).

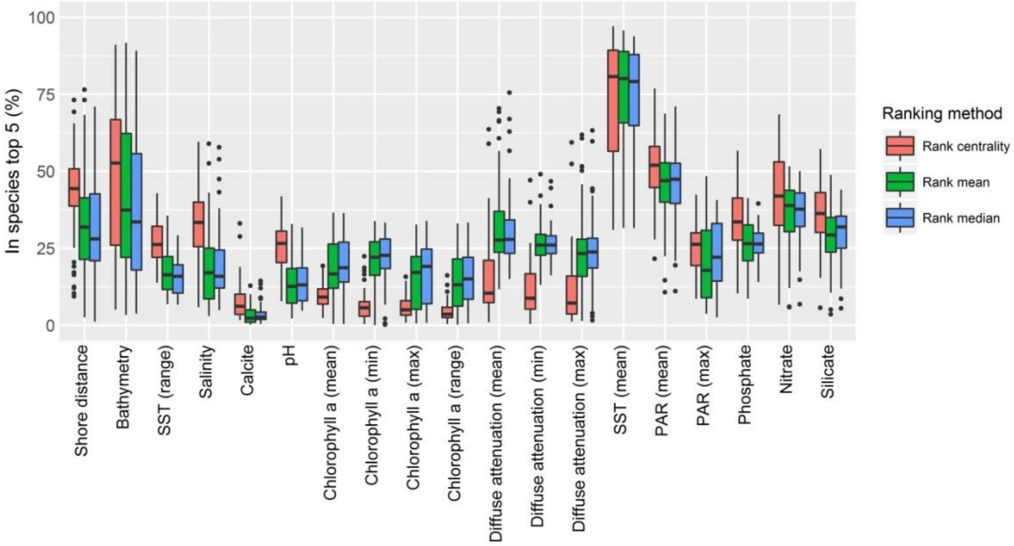


Figure S6. Percentage of species a predictor has a top 5 ranking in the different model setups for the three ranking methods: rank centrality (red), rank mean (green) and rank median (blue).

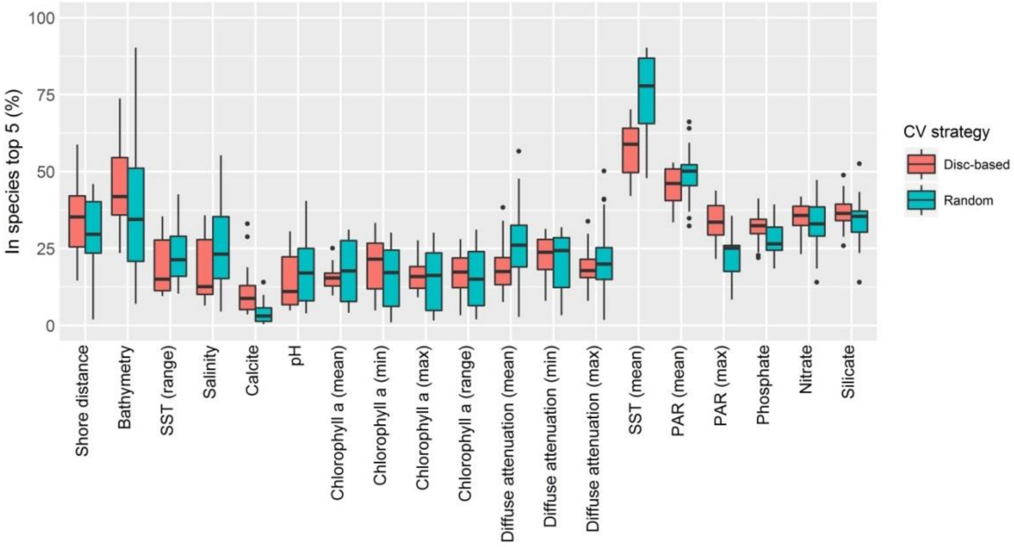


Figure S7. Percentage of species a predictor has a top 5 ranking in the different model setups for the two cross-validation (CV) strategies: disc-based CV (red) and random CV (blue). Note that only results for model setups that were run for both CV strategies are shown here.

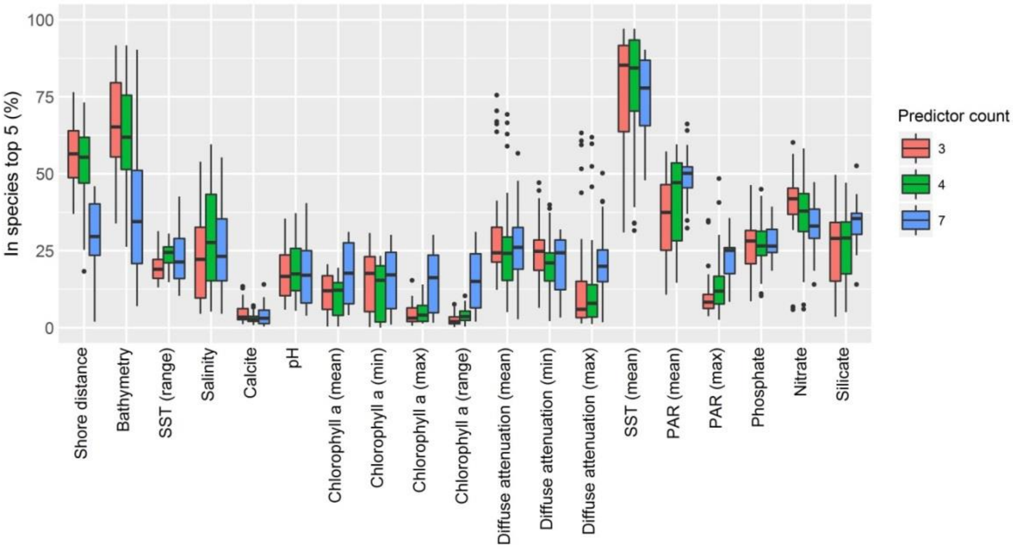


Figure S8. Percentage of species a predictor has a top 5 ranking in the different model setups for the different number of predictor counts: 3 (red), 4 (green), 7 (blue). Note that only results for model setups that were run for all three predictor counts are shown here.

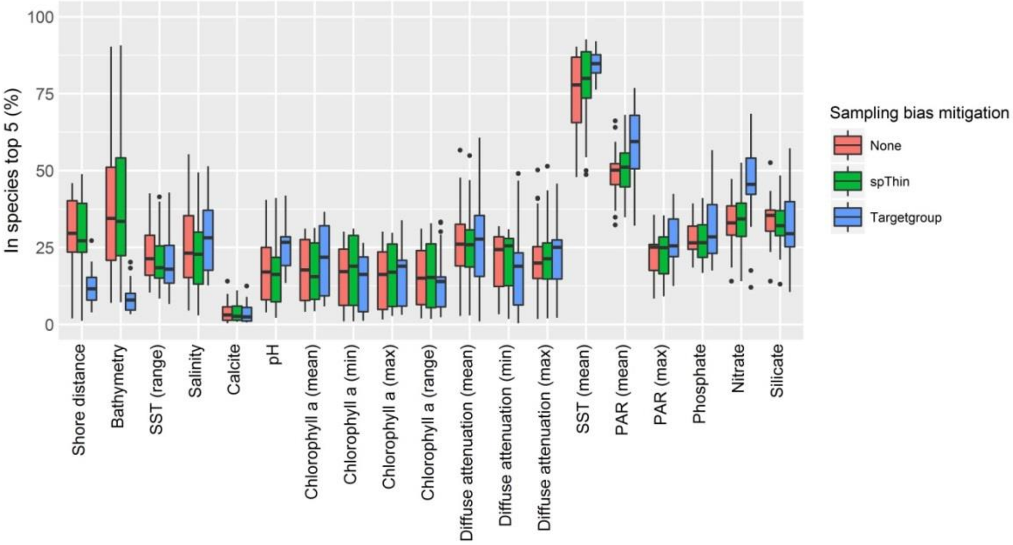


Figure S9. Percentage of species a predictor has a top 5 ranking in the different model setups for the different sampling bias mitigation methods: nothing (red), spatial thinning (spThin, green) and targetgroup background (blue). Note that only results for model setups that were run for all sampling bias mitigation methods are shown here.

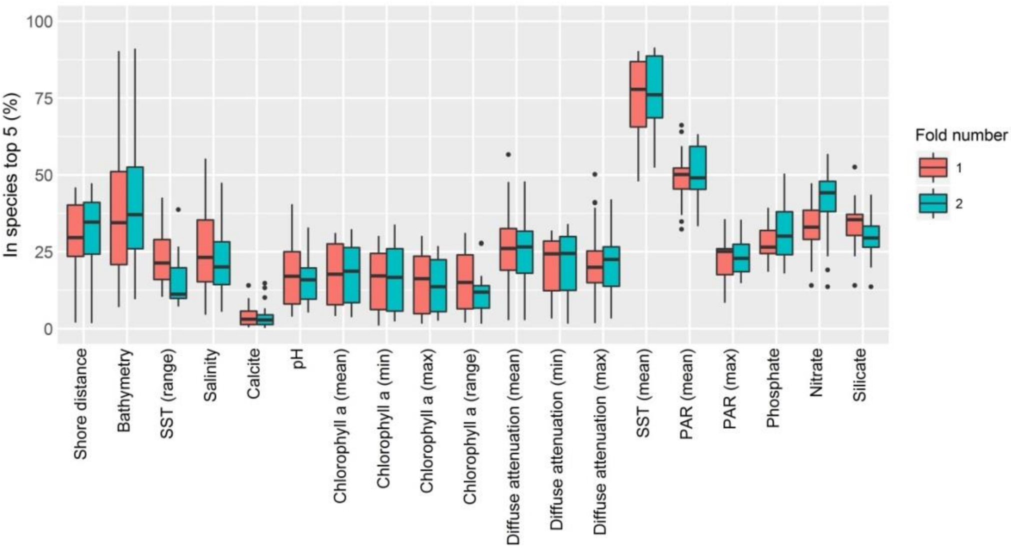


Figure S10. Percentage of species a predictor has a top 5 ranking in the different model setups for the two explored folds: 1 (red) and 2 (blue). Note that only results for model setups that were run for both folds are shown here.

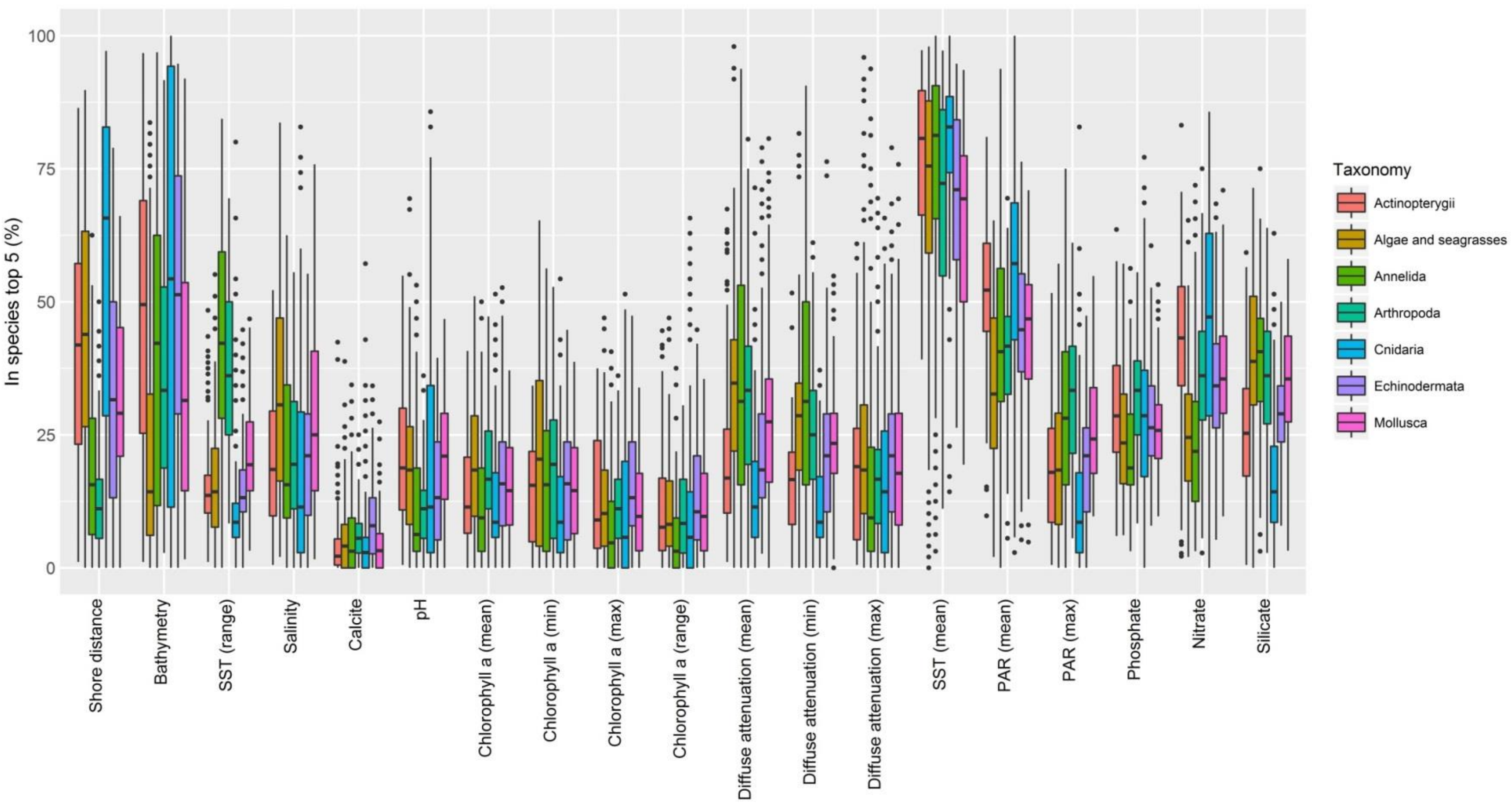


Figure S11. Percentage of species a predictor has a top 5 ranking in the different model setups for a selection of common taxonomic groups: Actinopterygii (red), algae and seagrasses (brown), Annelida (green), Arthropoda (cyan), Cnidaria (blue), Echinodermata (purple) and Mollusca (pink).

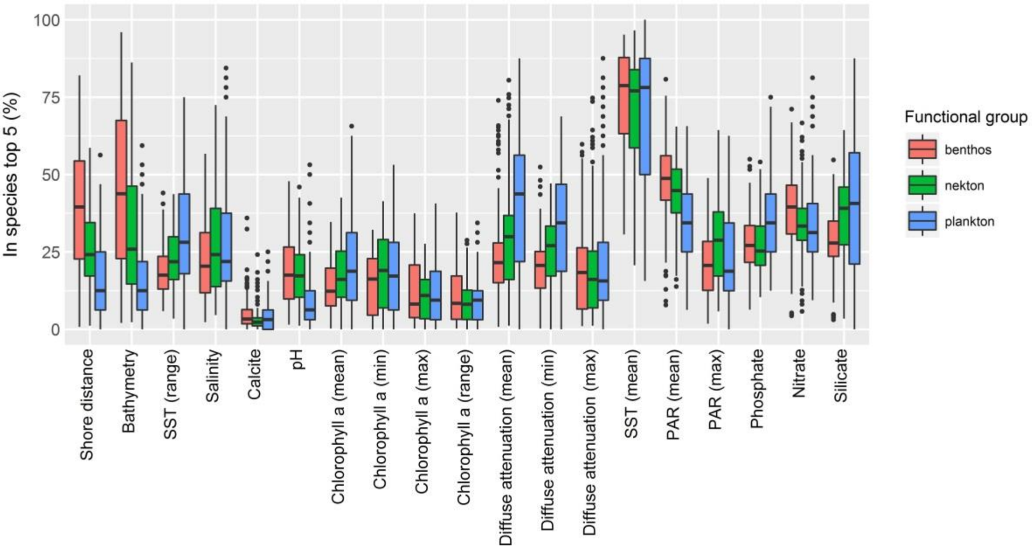


Figure S12. Percentage of species a predictor has a top 5 ranking in the different model setups for the different functional groups: benthos (red), nekton (green) and plankton (blue).

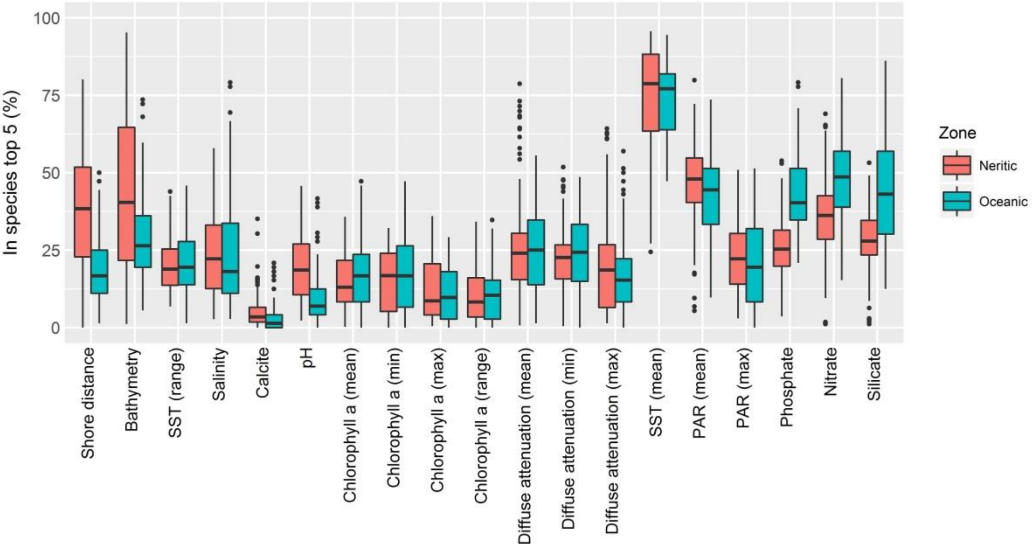


Figure S13. Percentage of species a predictor has a top 5 ranking in the different model setups for the different zones: neritic (red) and oceanic (blue).

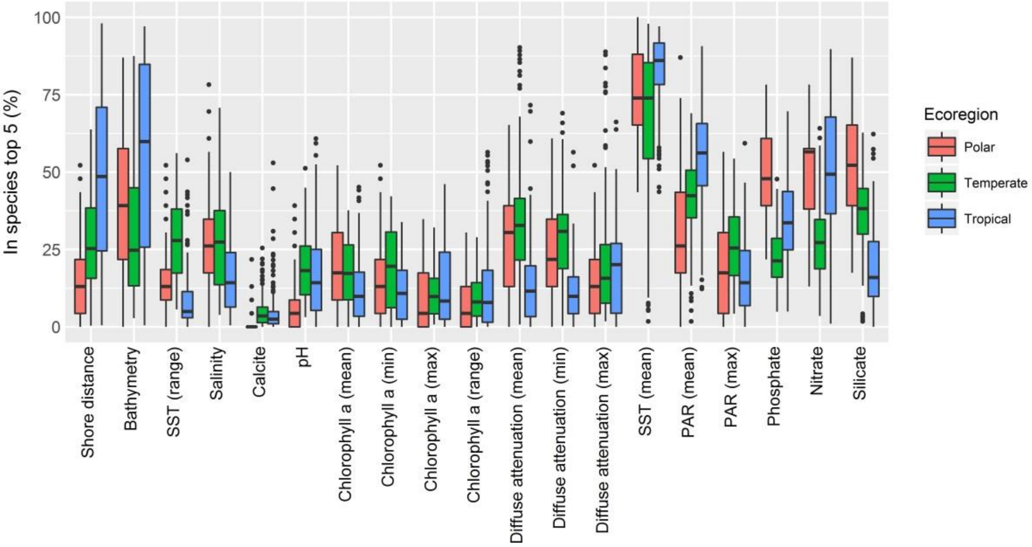


Figure S14. Percentage of species a predictor has a top 5 ranking in the different model setups for the different ecoregions: polar (red), temperate (green) and tropical (blue).

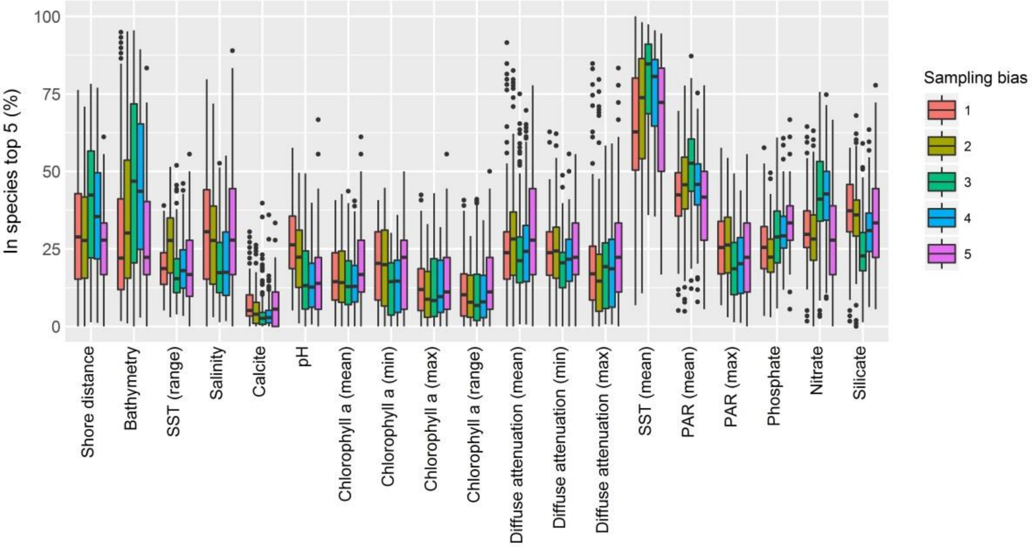


Figure S15. Percentage of species a predictor has a top 5 ranking in the different model setups for the different levels of sampling bias: 1 (low bias, red), 2 (brown), 3 (green), 4 (blue) and 5 (high bias, purple).