The following supplement accompanies the article

Ecological niche modelling of cold-water corals in the Southern Ocean (N Antarctic), present distribution and future projections due to temperature changes

Safa Chaabani, Pablo J. López-González, Pilar Casado-Amezúa, Hendrik Pehlke, Lukas Weber, Irene Martínez-Baraldés, Kerstin Jerosch*

*Corresponding author: kerstin.jerosch@awi.de

Marine Ecology Progress Series 628: 73–93 (2019)

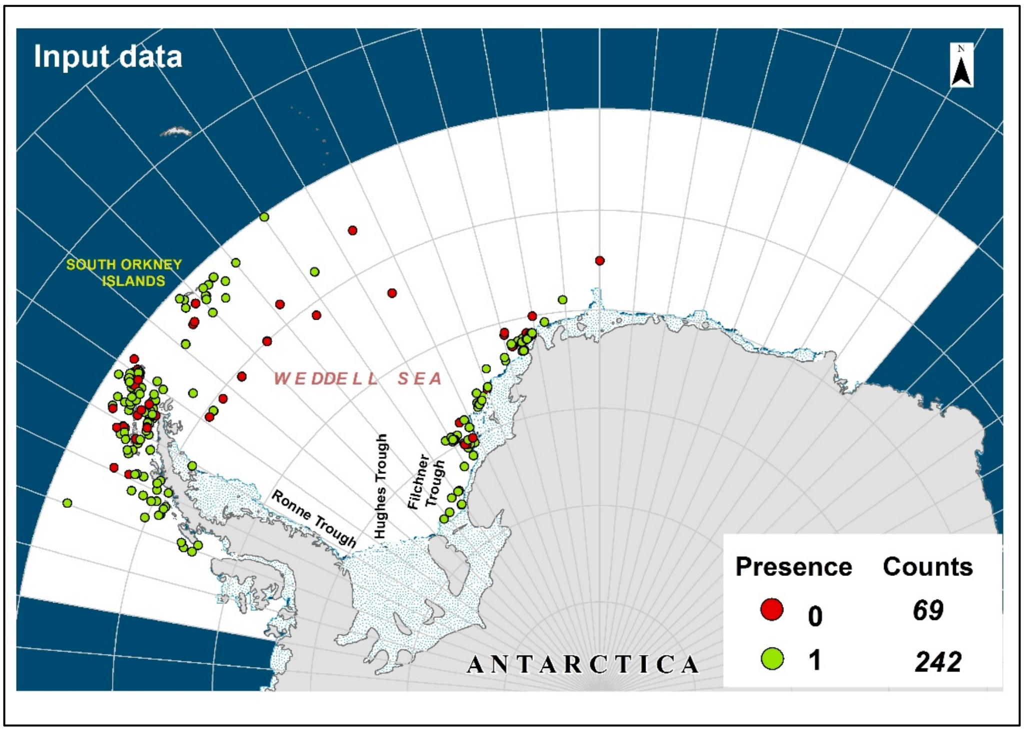

Figure S1: Real presence (green) and absence (red) data used as input to model the distribution of CWCs.

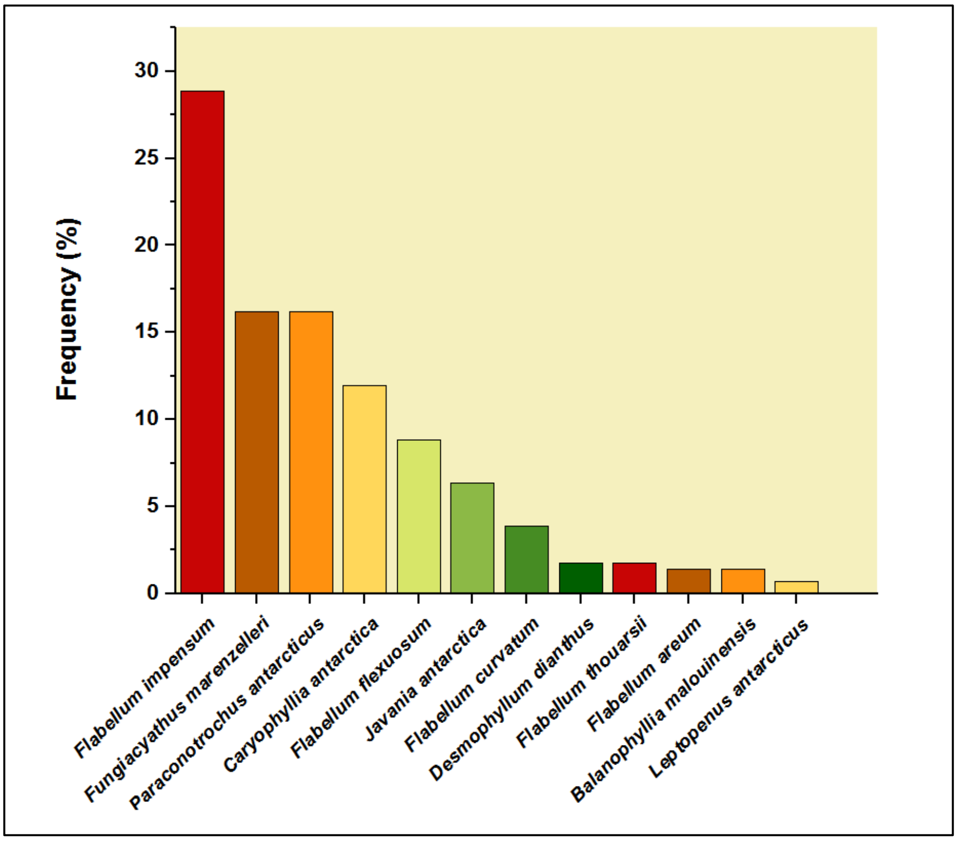

Figure S2: Identified scleractinian species used in the study and their relative frequency of occurrence.

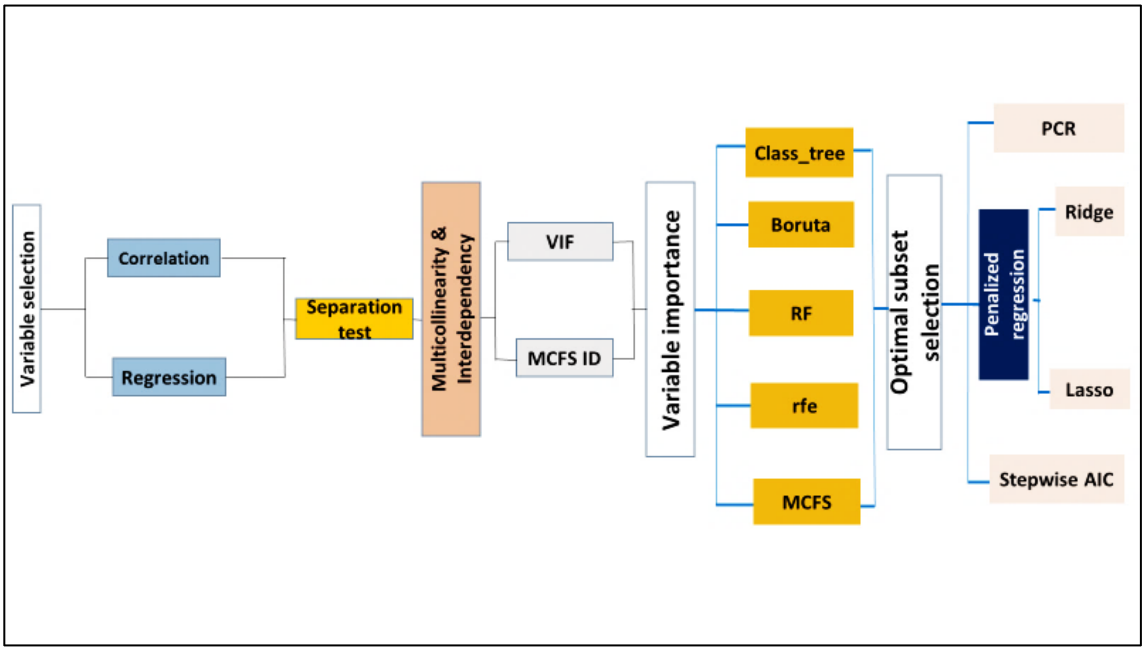

Figure S3: Workflow diagrams of the stepwise environmental variable selection. Correlation and regression analysis were used as a first step to identify the perfectly correlated variables that no additional information is gained by adding all of them. The weak performance of the logistic regression implied some risks of high multicollinearity among the variables, which was verified by the separation test. Multicollinearity and interdependency were investigated using the variance inflation factors (VIF) values and the Monte-Carlo Feature Selection (MCFS) Interdependency Discovery (ID) graph, respectively, to avoid including in the model different variables that have a similar predictive relationship with the response. After identifying groups of collinear and interdependent predictors, the variables were ranked according to their predictive power to identify those which contribute the least in describing the distribution of CWC from each group. This was achieved by comparing 5 of the most cited machine learnings in ecology modelling (classification tree (Class_tree), random forest (RF), Boruta algorithm (Boruta), recursive features elimination (rfe) and the Monte Carlo feature selection (MCFS)). The last step consisted in establishing a threshold for the exclusion of variables by applying 3 optimal subset selection methods (Principal component regression (PCR); Akaike's information criterion (AIC) and penalized regression through Lasso and Ridge techniques.

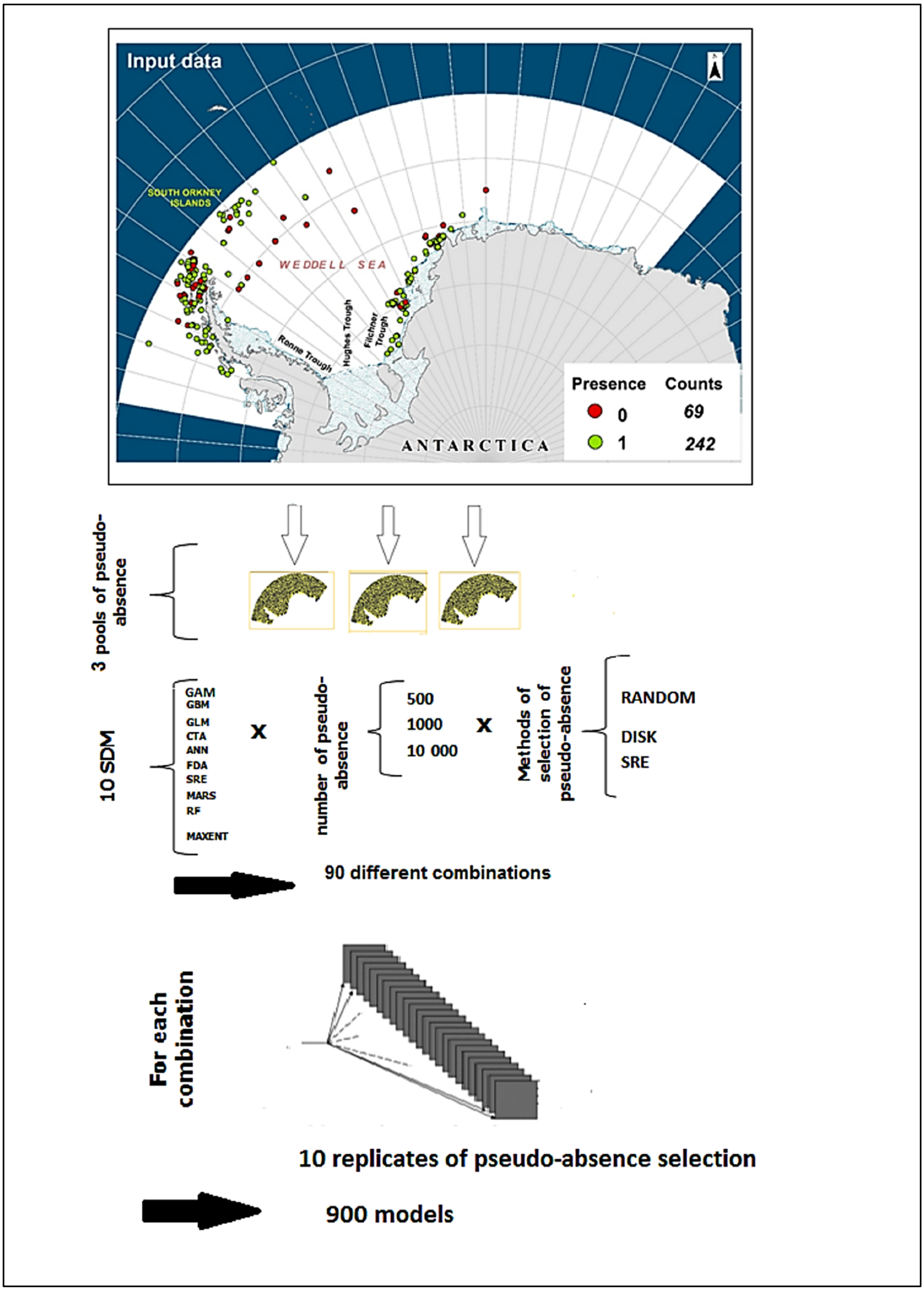

Figure S4: Schematic representation of the pseudo-absence selection procedure. 900 models, resulting from three different numbers of pseudo-absences (500, 1000 and 10 000), three methods to generate them (Random, Disk and SRE) and 10 replicates for each pseudo-absence selection, were compared to derive the optimal modelling input parameters.

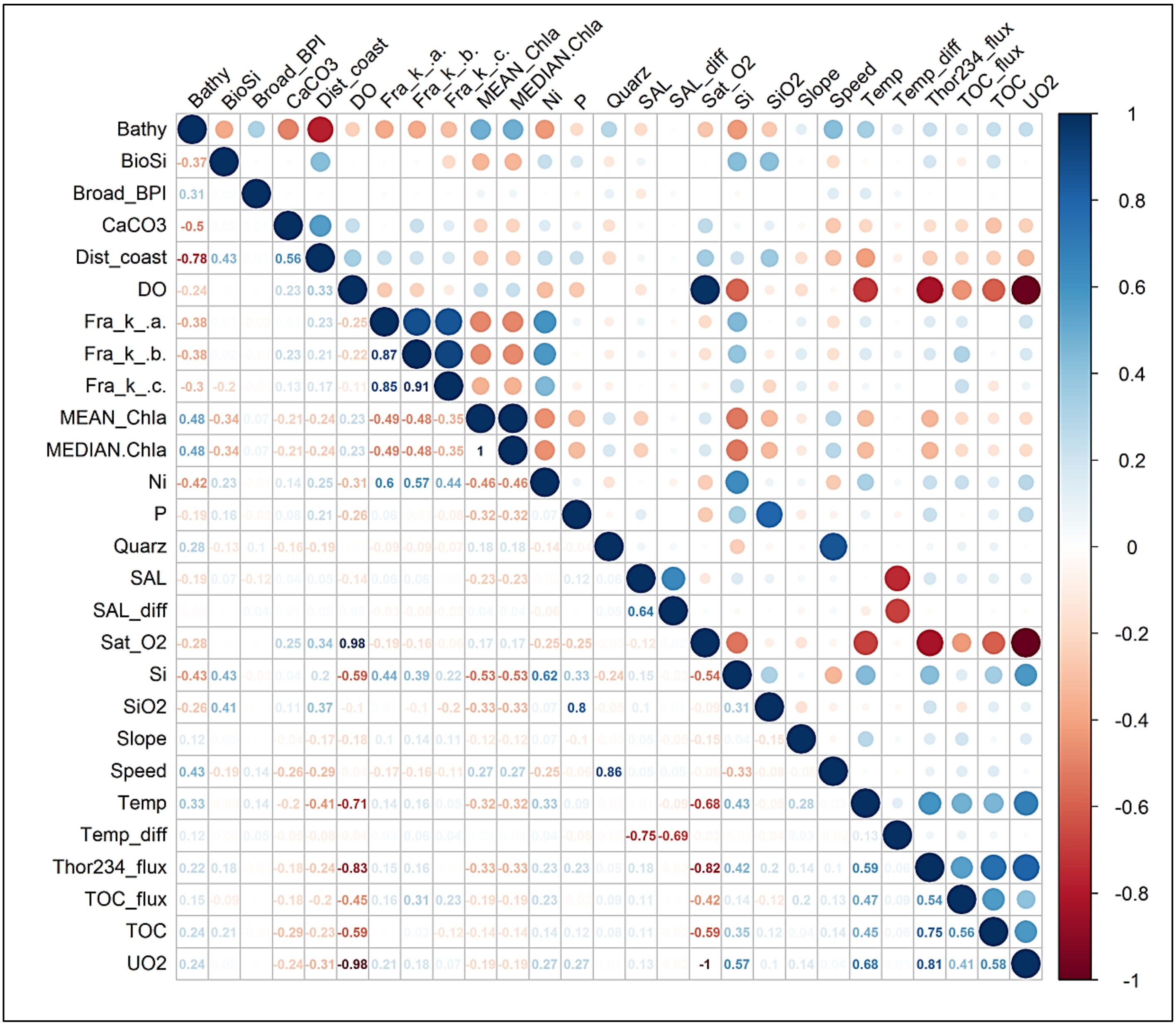

Figure S5: Pearson's correlation matrix comparing paired environmental covariates. Negative correlations are shaded red; positive correlations are shaded blue. Strength of the correlation is indicated by dot size and red or blue color saturation. High correlation between covariates is also indicated by the size of the colored oval delineating each comparison. The correlation coefficient between mean and median chlorophyll a is equal to 1. Thus, only mean chlorophyll a (MEAN_Chla) was kept for further analysis.

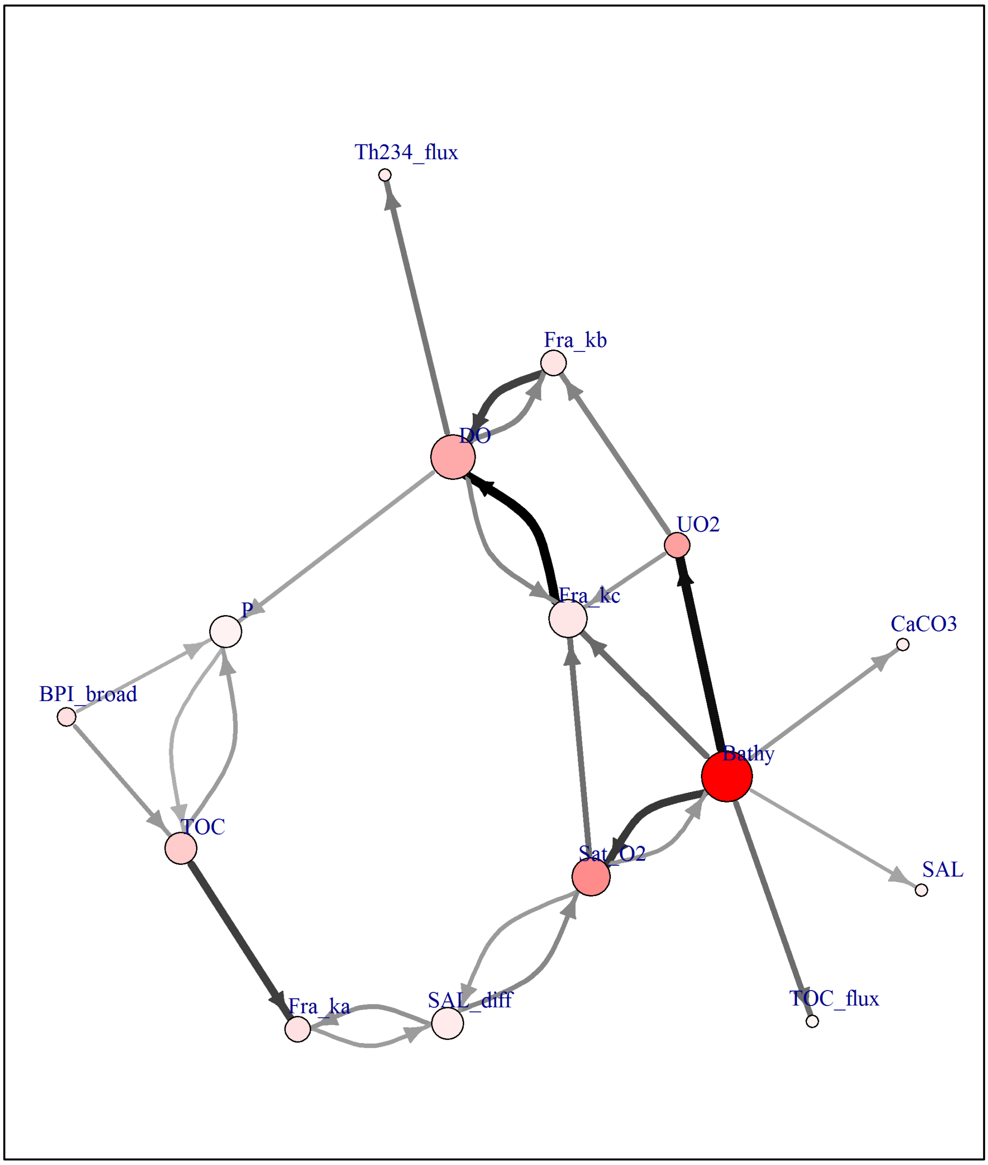

Figure S6: Interdependency Discovery (ID) graph featuring the top 25 ID weights and the 15 most important features identified by the Monte-Carlo Feature selection algorithm (MCFS). The color intensity of a node is proportional to the corresponding feature's relative importance derived from MCFS method. The size of a node is proportional to the number of edges (variables) related to this node. The width and level of darkness of an edge is proportional to the ID weight of this edge. The graph does not show any connection between related features (e.g., Sat_O2, DO and UO2; Fra_k_a, _b and _c). Indeed, such features do not "cooperate" in distinguishing between classes of response variable. This means that they have the same response to the presence-absence of CWC and thus they can have an accumulative effect and weight the distribution model. Based on the previous results and background knowledge, only one of the variables of each group of highly collinear and interdependent features was kept. Hence, both saturated (Sat_O2) and utilized oxygen (UO2) were removed, to keep only the dissolved (DO). All variables representing the diatom Fra_k distribution have been discarded from the analysis to avoid complexity. Moreover, BioSi and SiO2 have been eliminated due to their strong links with diatoms and undefined influence on CWC. Hence, among silicate variables, only Si was kept in the predictors set for the analysis of importance. The Mean chlorophyll a variable (Mean_Chl.a) was eliminated from the set due to the high number of missing values in its raster compared to the other predictors. To summarize, 18 out of 27 variables remained for the importance analysis.

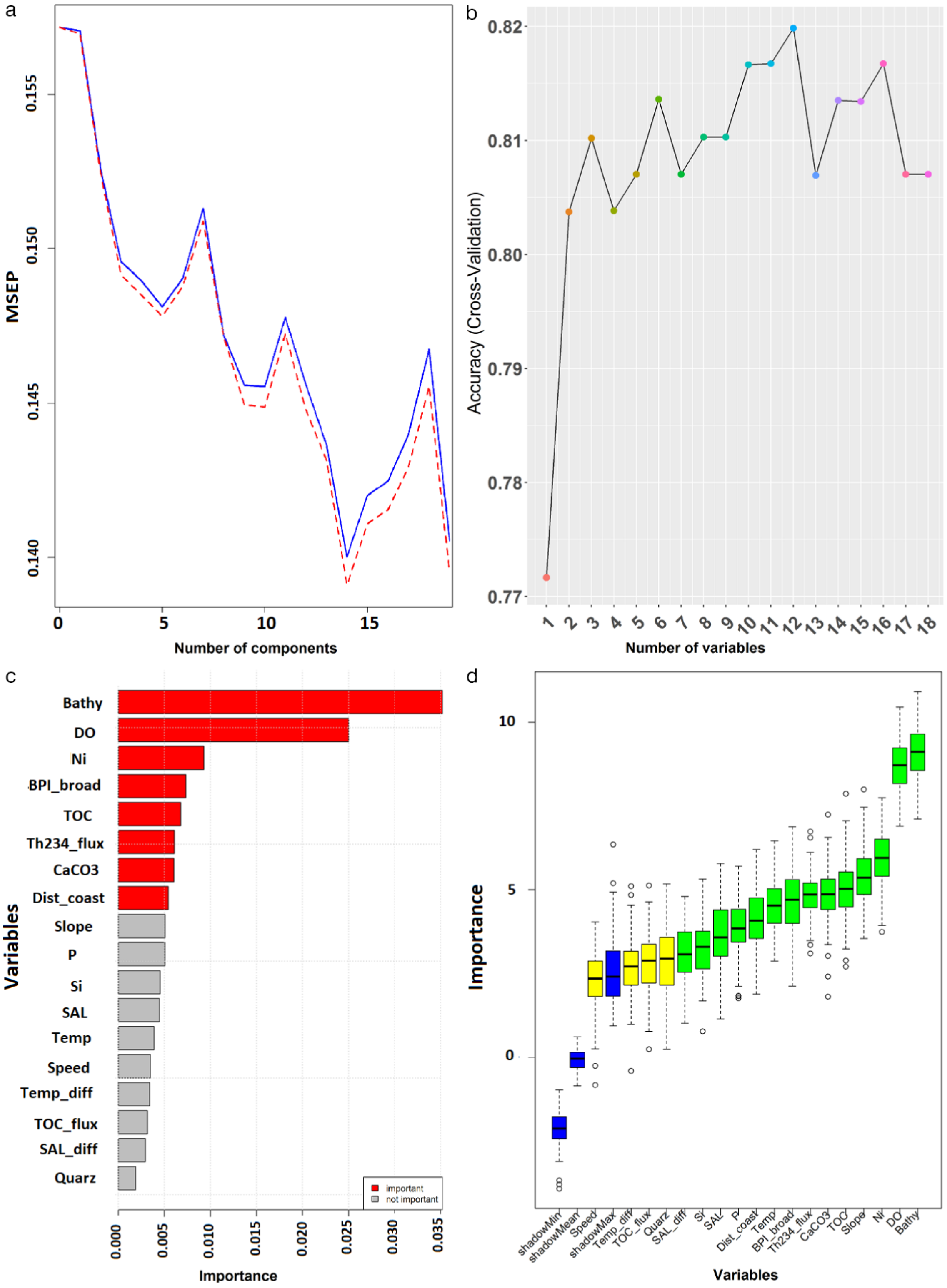

Figure S7: Some results of the optimal subset selection and variable importance analysis supporting the decision of lowering the size of the predictors' set: a) Validation plot of the principal component regression test (PCR). The graph shows the evolution of the mean squared error of prediction (MSEP) with the number of variables (components) in the model. Low cross validation error (MSEP) indicates a better model and thus an optimal number of components. Although 18 components can explain 100% of the variability of the data, the cross validation (CV) error is a little higher than with 14 components, which are sufficient to explain 98.03% of the variance of the presence/absence of CWC. b) Recursive feature elimination (rfe) variable importance. The graph shows the performance profile across different subset sizes of variables. The different subset sizes are evaluated by the accuracy of the relative random forest algorithm. The rfe model scores the highest accuracy (0.82) with 12 variables. But even with a higher or lower number of predictors, the accuracy is still high enough (>0.8) c) Monte Carlo feature selection (MCFS) relative importance (RI). The plot shows the top features set ranked according to their RI. The most important variables are represented in red color and the non-important in gray. d) Variables importance derived from Boruta test based on the mean decrease accuracy. Box plots of all the attributes plus minimum, average and max shadow scores are represented (blue). Variables having boxplot in green are considered important and the yellow color indicates they are tentative. Tentative Attributes refer to the variables which importance score is so close to their best shadow attributes that Boruta is unable to decide within the default number of random forest runs where to classify them. Boruta considered 14 variables (in green) as enough to describe the distribution of CWC.

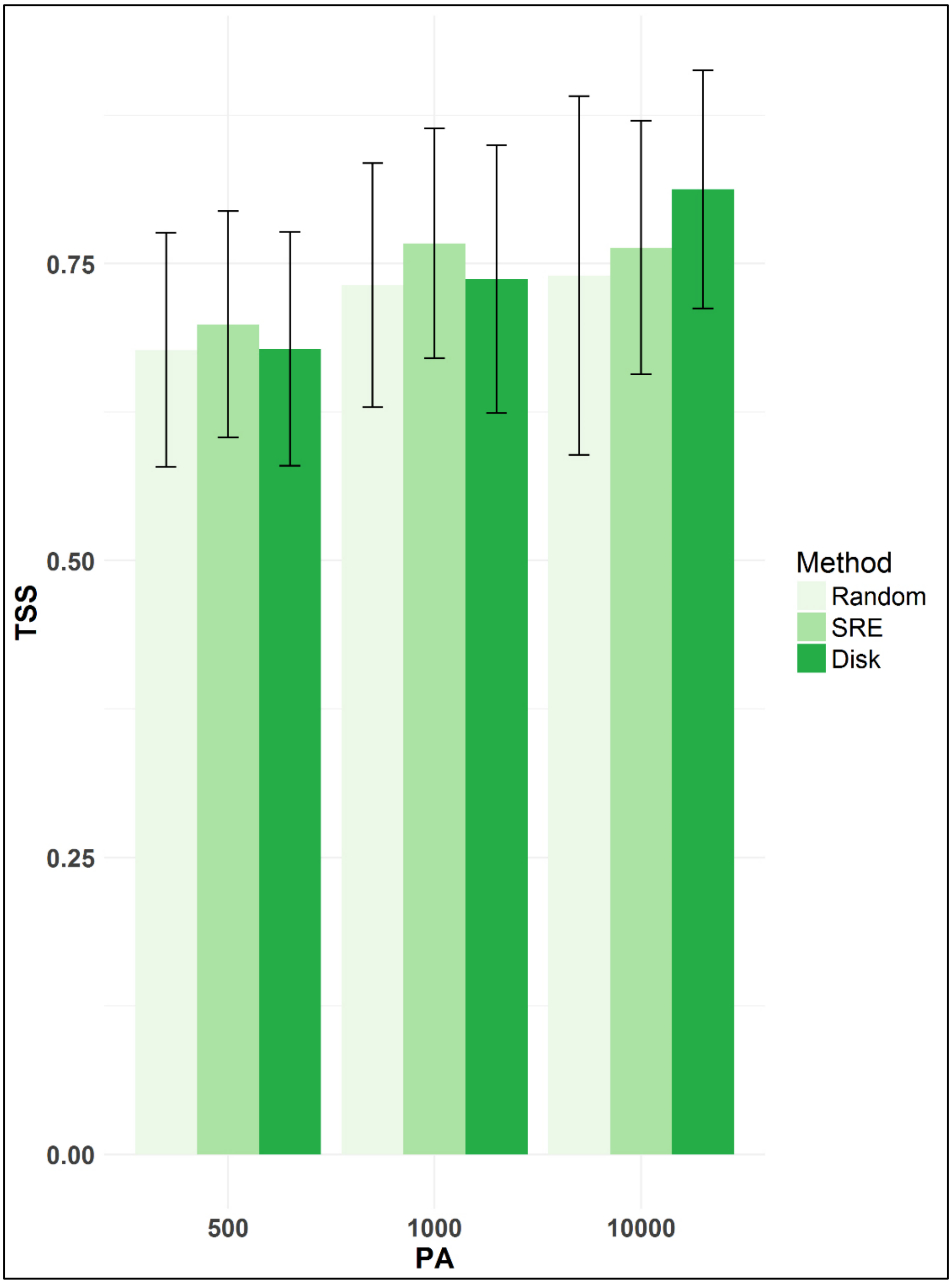

Figure S8: Response of the model accuracy, assessed via the mean TSS ±SD (y-axis), to the method (color-fill) and the number of generated pseudo-absences (x-axis).

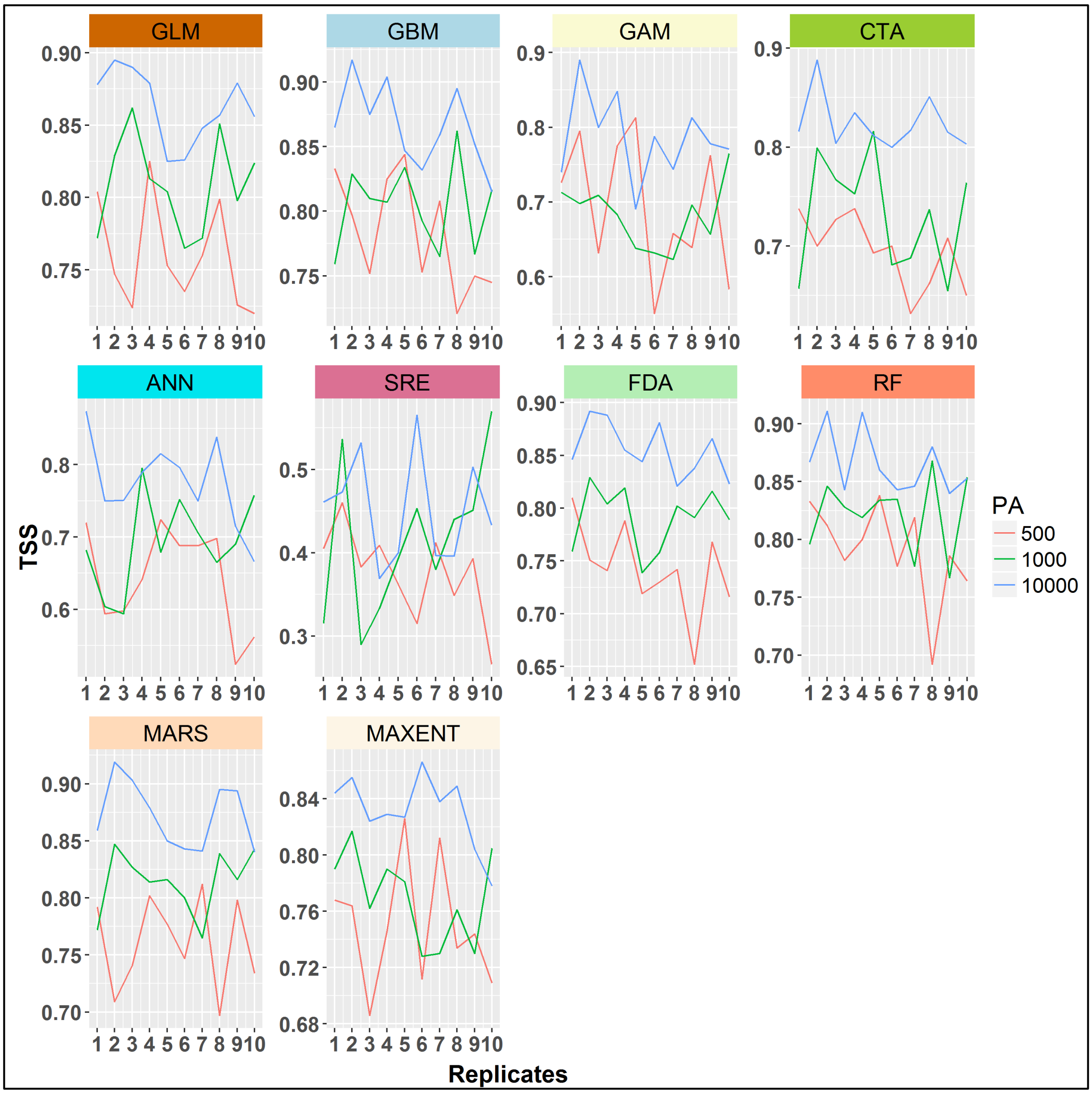

Figure S9: Individual responses (mean TSS) of the SDM to the number of selected pseudo-absences (colors of the curves; red=500/ green=1000 and blue=10 000) and the number of replicates (from 1 to 10 in the x-axis) used to generate them.

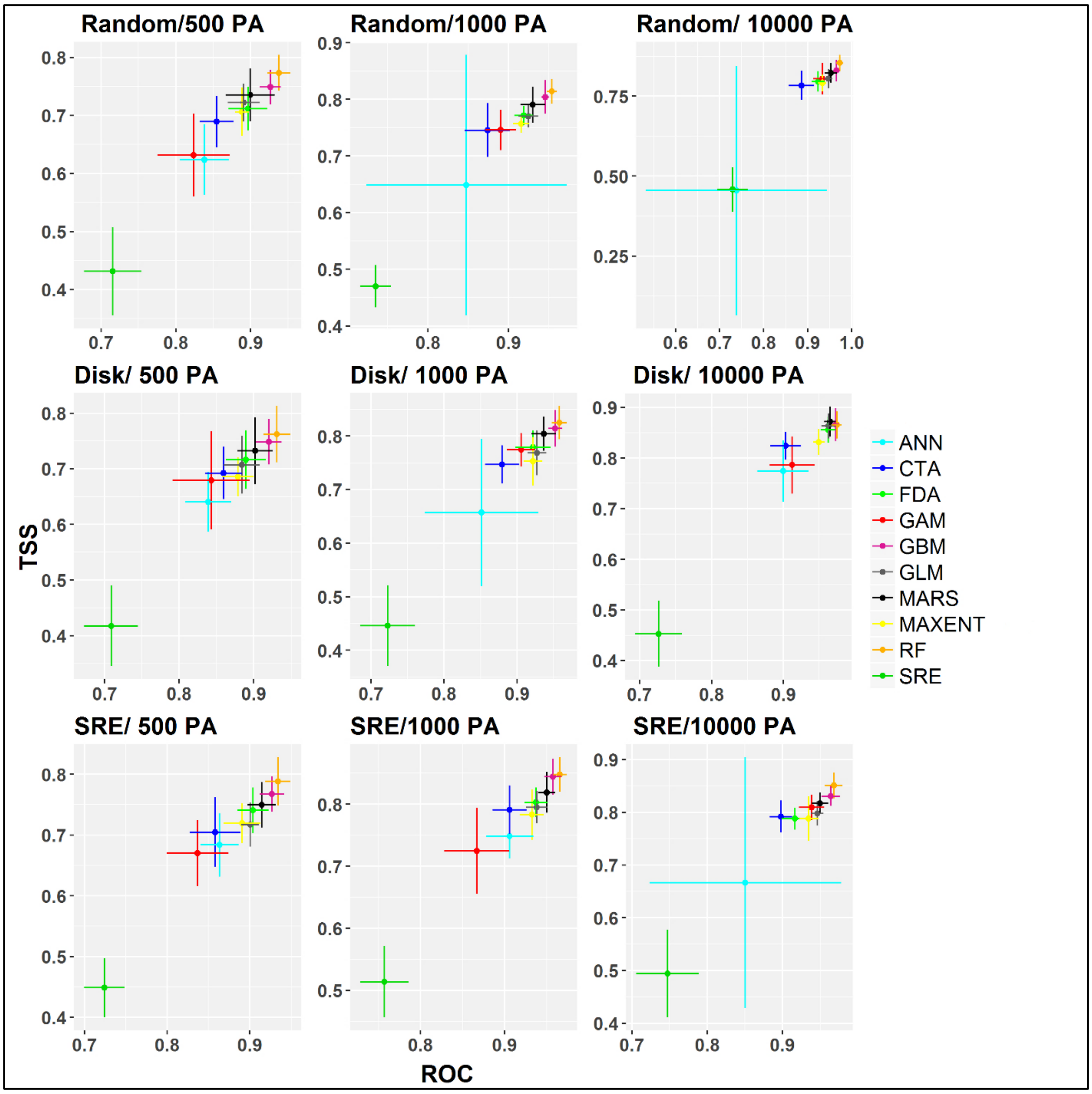

Figure S10: Evaluation scores (ROC and TSS in x and y axis respectively) of the SDM for the different combinations of number of pseudo-absences (PA) and methods to generate them. Points represent the average score and lines the standard deviation of the evaluation scores across each model’s runs.

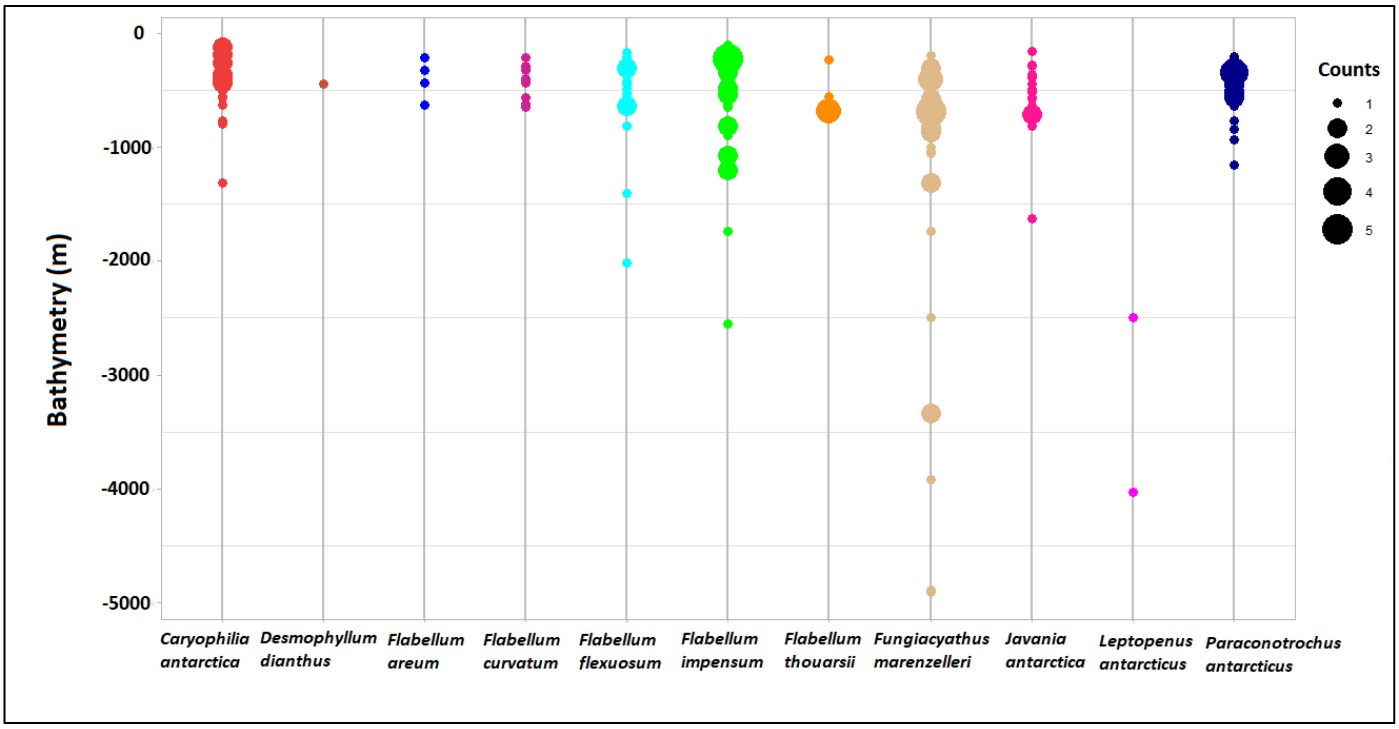

Figure S11: Count-plot of the species distribution according to the different bathymetry ranges. The y-axis represents the bathymetry values and the x-axis the different species. The size of the colored circles is related to the number of counts. Please note that data were not available for one species (*Balanophyllia malouinensis*).

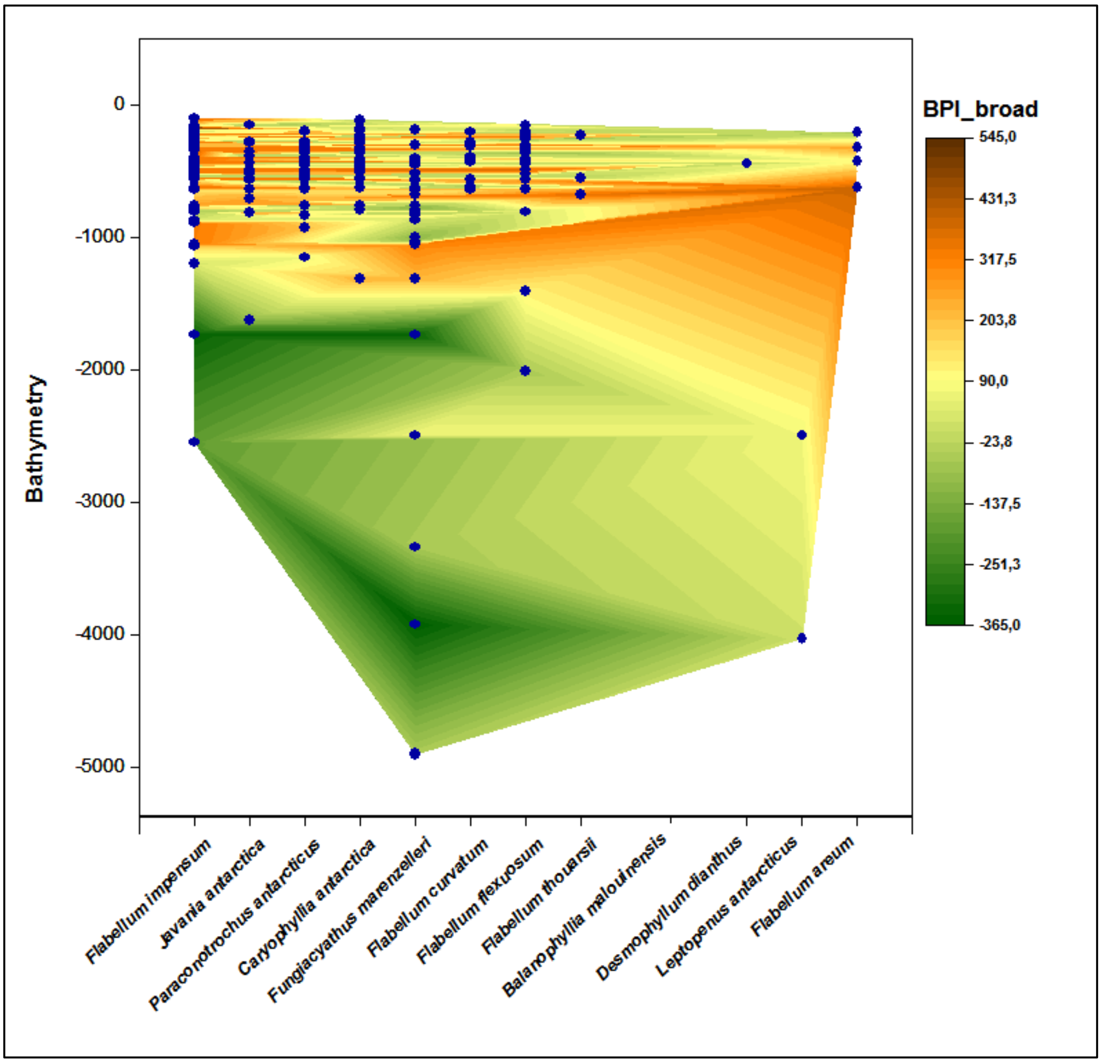

Figure S12: Observed presences of CWC (blue points) in the different bathymetry ranges (y-axis) and the corresponding BPI values (color scale). Data were not available for one species (*Balanophyllia malouinensis*).

Table S1: Summary of the logistic regression model (GLM) fitted to the full set of variables. The statistically significant relations (at α= 0.05) with the response variable are written in red and marked with an asterisk (*). Only two (Bathy and Ni) of the twenty-seven predictors were significantly related to the CWC (at α= 0.05). Some variables (eg. Fra_k (a, b and c); UO2; Speed) had large or low estimates and high standard errors which can be a sign of a separation problem in the dataset and hence multicollinearity. The separation issue is frequently occurring when binary response data are analyzed by logistic regression models. If not dealt with appropriately, it may lead to biased conclusions with regards to the relevance of a particular variable to the presence or absence of CWC.

Predictors	*Estimate*	*Std.Error*	*z.value*	*Pr(>|z|)*
(Intercept)	-770.39	617.31	-1.25	0.21
Bathy	0.002	0.0008	2.12	**0.03 ***
BioSi	0.15	0.4	0.37	0.71
BPI_broad	-0.004	0.003	-1.29	0.2
CaCO3	-0.07	0.14	-0.51	0.61
DO	-2.27	12.13	-0.19	0.85
Dist_coast	5.22 e-06	7.87 e-06	0.66	0.51
Fra_k.a	-13.04	21.06	-0.62	0.54
Fra_k.b	13.51	24.39	0.55	0.58
Fra_k.c	-0.27	15.25	-0.02	0.99
Mean_Chla	0.07	0.22	0.33	0.74
Ni	-4.36	2.56	-1.71	**0.09 ***
P	42.18	28.92	1.46	0.14
Quarz	-0.008	0.008	-1.04	0.3
SAL	5.05	6.72	0.75	0.45
SAL_diff	-0.88	16.27	-0.05	0.96
SiO2	0.23	1.39	0.17	0.87
Si	-0.02	0.11	-0.22	0.83
Slope	0.11	0.11	1.00	0.31
Speed	42.56	41.54	1.02	0.31
Temp	-0.4	0.81	-0.49	0.63
Temp_diff	2.12	1.96	1.08	0.28
TOC	-11.61	11.13	-1.04	0.3
TOC_flux	-3.52	4.97	-0.71	0.48
UO2	86.19	71.62	1.20	0.23
Th234_flux	0.007	0.006	1.27	0.20
Sat_O2	6.38	5.95	1.07	0.28

Table S2: Variables inflation factor (VIF). A VIF for a single explanatory variable is obtained using the r-squared value of the regression of that variable against all other explanatory variables. A value higher than 10 is commonly considered as very high. The analysis exhibit the groups of variables that have a high degree of multicollinearity: bathymetry and its derivatives such as slope, BPI and distance to coast; Silicate group (BioSi, SiO2 and Si); Oxygen variables (DO, UO2 and SAT_O2) and the diatoms group (Fra_k.a, Fra_k.b and Fra_k.c).

PREDICTORS	VIF
Bathy	7222.80
BioSi	5468.26
BPI_broad	244.70
CaCO3	120.78
DO	116.13
Dist_coast	61.32
Fra_ka	45.83
Fra_kb	39.60
Fra_kc	39.11
Mean_Chl.a.	25.86
Ni	19.47
P	17.00
Quarz	15.80
SAL	15.28
SAL_diff	15.25
SiO2	13.80
Si	13.59
Slope	9.41
Speed	6.68
Temp	5.99
Temp_diff	5.29
TOC	3.23
TOC_flux	2.88
UO2	2.40
Th234_flux	1.66
Sat_O2	1.59

Table S3: Summary of the variable importance estimated by the different selection methods and the corresponding rank computed for each explanatory variable and each method. The classification tree (Class-Tree) method calculated importance values only for 14 variables which were considered in building the tree and assigned zeros to the features deemed unimportant. The ranks corresponding to the most important variables are written in red while the least important in blue. The applied features selection methods did not assign similar ranks to the variables as a consequence of the different approaches used by each method to estimate importance. However, the lowest variable importance values were assigned to Temp_diff, TOC_flux, Sal_diff and Speed by the majority of the algorithms and were considered for elimination.

var	*MCFS*	*Boruta*	*rfe*	*RF*	*Class_ Tree*	*rank MCFS*	*rank Boruta*	*rank rfe*	*rank RF*	*Rank Class_tree*
Bathy	0.0352	9.13	14.34	0.04543	17.5	1	1	1	1	2
DO	0.02501	8.7	13.33	0.04236	10.71	2	2	2	2	6
Ni	0.00928	5.92	10.6	0.02458	12.6	3	3	3	3	4
BPI_broad	0.00735	4.63	8.66	0.02281	18.11	4	8	6	4	1
TOC	0.00676	4.99	8.82	0.02205	7.97	5	5	5	5	8
Th234_flux	0.00608	4.85	8.42	0.02145	11.46	6	7	7	6	5
Slope	0.00507	5.39	8.85	0.01946	0	9	4	4	7	-
SAL	0.00445	3.66	7.33	0.01785	3.84	12	12	12	8	12
Temp	0.00388	4.48	8	0.01748	0	13	9	9	9	-
CaCO3	0.00604	4.89	8.27	0.01656	8.95	7	6	8	10	7
Dist_coast	0.00542	4.1	6.64	0.01456	12.95	8	10	14	11	3
Quarz	0.00187	2.89	6.82	0.01351	0	18	15	13	12	-
P	0.00503	3.87	7.8	0.01284	4.9	10	11	10	13	9
Si	0.00452	3.25	7.48	0.01184	4.14	11	13	11	14	10
SAL_diff	0.00291	3.1	6.23	0.01155	3.69	17	14	15	15	13
Temp_diff	0.0034	2.69	6.16	0.01028	4.02	15	17	16	16	11
TOC_flux	0.00313	2.79	5.8	0.00788	2.76	16	16	17	17	14
Speed	0.00346	2.29	5	0.00754	0	14	18	18	18	-

Table S4: AIC stepwise variable selection results. The stepwise-selected model is returned along with the steps taken in the backward search. The minus sign means that the variable was eliminated. The deviance measures the goodness of fit of the model (the lower the better). The residual deviance "Resid. Dev" column refers to a constant minus twice the maximized log likelihood. Df corresponds to the degree of freedom and the residual degrees of freedom (Resid.Df) indicates the total Df minus the Df of the model. The first row in the table indicates the AIC value of the initial model including the full set of the 18 variables. Both Temp_diff and Sal_diff appeared among the eliminated attributes by AIC stepwise selection, while TOC_flux and Speed were deemed important. Excluding Temp_diff, Si, SAL_diff or BPI_broad can reduce the AIC by 2 units, while the decrease resulting from the stepwise elimination of the rest of variables (SAL,Temp, Dist_coast & CaCO3) is not important (=< 1 unit). These results support the elimination of Temp_diff and SAL_diff.

Final Model:
CWC ~ Bathy + DO + Ni + P + Quarz + Slope + Speed + TOC + TOC_flux + Th234_flux

Step	*Df*	*Deviance*	*Resid.Df*	*Resid.Dev*	*AIC*
	-	-	292	248.15	286.15
- Temp_diff	1	4.34e-05	293	248.15	284.15
- Si	1	0.03	294	248.18	282.18
- SAL_diff	1	0.09	295	248.27	280.27
- BPI_broad	1	0.07	296	248.34	278.34
- SAL	1	0.86	297	249.19	277.19
- Temp	1	1.39	298	250.59	276.59
- Dist_coast	1	1.72	299	252.30	276.30
- CaCO3	1	0.77	300	253.07	275.07

Table S5: Regularized GLM: Ridge & Lasso coefficients. Only those predictors that have non-zero or very close to 0 coefficients are significant for the response variable. Both SAL_diff and Temp_diff were assigned very low values (especially by Lasso regression), which means that these variables have minor contribution in the model. Lasso reduced the complexity of the fitting function by forcing the coefficients of Si and BPI_broad to 0. Based on these results, the stepwise AIC and the variable importance analysis, Temp_diff and SAL_diff, BPI_broad and TOC _flux were eliminated from the variables set.

Predictors	***Ridge***	***Lasso***
(Intercept)	1.4477	1.7255
Bathy	0.4025	0.6995
BPI_broad	0.1208	0
CaCO3	-0.2806	-0.5607
Dist_coast	0.1859	0.4033
DO	-0.5814	-0.6968
Ni	-0.7	-2.1503
P	0.0313	0.6375
Quarz	-0.2154	-0.621
SAL	-0.0801	-0.185
SAL_diff	-0.103	0.0105
Si	-0.1839	0
Slope	0.2236	0.39
Speed	0.3577	0.8173
Temp	0.1133	0.3728
Temp_diff	-0.0305	0.0054
Th234_flux	0.3867	1.6878
TOC	0.0106	-0.4731
TOC_flux	-0.0596	-0.3869