Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ORIGINAL ARTICLE

Ensemble learning in ecological niche modeling of the *Siderastrea* (Cnidaria:Scleractinia) Atlantic complex: an information theory approach

Henrique J. Schipanski MSc[1] | **Lucas B. Perroni**[2] | **Caio F. Sguario**[3] | **Marcos S. Barbeitos PhD**[1]

[1]Departamento de Zoologia, Universidade Federal do Paraná, Curitiba, PR, Brazil

[2]Departamento de Física, Universidade Federal do Paraná, Curitiba, PR, Brazil

[3]Departamento de Informática, Universidade Federal do Paraná, Curitiba, PR, Brazil

Correspondence
Marcos S. Barbeitos PhD, Departamento de Zoologia, Universidade Federal do Paraná, Curitiba, PR, 81531980, Brazil.
email:msbarbeitos@gmail.com

Funding Information
This research was supported by the Conselho Nacional de Pesquisa (CNPq), PhD Scholarship Number: 140363/2023-8; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), MSc Scholarship Number: 1810146

Abstract

Ecological niche modeling (ENM) is an invaluable tool in biodiversity science and many algorithms were developed for or incorporated into this area over the years. Back in 2007, Araújo and New 2007 outlined a framework to integrate different sources of variation into robust model projections. Almost 20 years later, surveys of real world applications show that standards are still lacking, particularly when it comes to combining different modeling strategies into a single set of predictions. In this paper we introduce a novel information theory approach to ensemble learning using the *Siderastrea* Atlantic Complex (SAC) as a working example. We employed nine models (BioClim, Gower, and Mahalanobis distances, GAM, GLM, MARS, ANN, MaxEnt, and SVM) to test for cross-area transferability of two species in SAC. We used Shannon's entropy to evaluate model overfitting to the data. Our method reveals that overfitting is not spatially uniform, thus, we developed a weighting strategy that reduces the influence of overfitted models on ensemble predictions on a cell-by-cell basis. Overfitting is obscured by traditional measures, such as AUC and TSS, that provided inflated estimation of real model performances across projection areas. We also introduce reliability maps, analogous to the Maps of Biogeographical Ingnorance (MoBIS), combining occurrence probability and ensemble efficiency on each cell. Our approach uncovered the Yucatán peninsula as an adequate habitat for the endemic Brazilian species *Siderastrea stellata* and, conversely, the Northeastern Brazilian coast as a suitable area for the Caribbean species *S. siderea*, supporting previously published records. It also uncovered a stretch of Western African coast as highly suitable habitat for the genus, a result that was recently corroborated by citizen-science reports.

KEYWORDS
areas of accessibility, confusion matrix, coral, cross-validation, extrapolation, interpolation, machine learning, "No Free Lunch" Theorem, reef, sensitivity, Shannon's entropy, species distribution modeling, specificity, testing, training, transferability

1 | INTRODUCTION

Ecological niche modeling (ENM) has broad applications in agriculture, public health, biodiversity, and conservation (Soley-Guardia et al. 2024). The goal of ENM is to construct a function that maps the input (environmental) data to its output pair,

usually a continuous target variable expressing the adequacy of any given locality to the existence of a species (Peterson et al. 2011). ENM algorithms are grouped in three main categories (Thuiller 2024): (i) environmental "envelopes", such as BioClim (Booth et al. 2014) and Mahalanobis distances (Farber and Kadmon 2003); (ii) statistical models, including generalized linear models (GLM - Guisan et al. 2002) and multivariate adaptive regression splines (MARS - Friedman 1991); and (iii) machine learning (ML) models, e.g. Maximum Entropy (MaxEnt - Phillips and Dudik 2008), and Support Vector Machines (SVM - Drake et al. 2006). In spite of such division, all correlative ENM algorithms are in fact ML methods because they learn the fundamental niche dimensions automatically from "labeled" data, with no need for targeted human programming (Sarker 2021). The labeled data used to guide the model during function construction are the occurrence points where the species was actually recorded, and (pseudo) absence points, if necessary. Hence ENM is a regression task that falls into the ML class known as supervised learning (Sarker 2021).

In their seminal paper Araújo and New (2007) stated that forecast ensembles have clear advantages over single model predictions because they may enable more robust decisions in face of uncertainty. Using their definition, an ensemble is made up by a set of initial conditions (IC), different model classes (MC = ENM algorithms) and their parameters (MP), and boundary conditions (BC, e.g. a IPCC Emission Scenarios) onto which predictions will be projected. The authors favored a statistical mechanics perspective in which each instance defined in this multidimensional space represents one possible state of the system. Taken together, these instances make up the ensemble forecast. In his original formulation, Gibbs (1902) proposed that the probabilities describing the states of a system imperfectly known should be estimated from completely characterized random instances of this system. Araújo and New (2007) see this probabilistic characterization of species distributions as the "end-game" of ensemble forecasting.

The framework intended by those authors requires some degree of MP sampling but, in reality, researchers mostly employ default settings when running their analysis (Hao et al. 2019). Parameter space exploration is normally conducted during model tuning, an essential step in ENM (Cerasoli et al. 2022, Velazco et al. 2022, Soley-Guardia et al. 2024). Although there are specialized packages (e.g. *BIOMOD2* - Thuiller et al. 2009, *ENMeval* - Muscarella et al. 2014, *flexsdm* - Velazco et al. 2022), their use is still incipient because tuning is laborious and time-consuming (Muscarella et al. 2014): in their meta-analysis Feng et al. (2019) found this step in 45% of the 163 surveyed papers. Variations in initial conditions (IC) are introduced by stochastic model training and testing through some random sub-sampling procedure. This is intended to estimate how much ENM results are conditioned on the particularities of the chosen dataset. Model robustness to variations in IC may be quantified by system performance measures such as the Area Under the Receiver Operating Characteristic curve (AUC - Fielding and Bell 1997), True Skill (TSS - Allouche et al. 2006), or Kappa Statistics (Pearson et al. 2006). Model evaluation is a widespread practice, reported in 90% of the cases examined by Feng et al. (2019). On the other hand, most studies still resort to a single model class (MC): Qiao et al. (2015) found that 80% of 245 empirical studies employed only one algorithm, mostly MaxEnt (Phillips and Dudik 2008). Melo-Merino et al.

(2020) focused on ENM applied to marine organisms and found that only 74 out of 328 (~22%) surveyed publications made use of more than one model class. In contrast, other areas that employ ML, where ensemble learning is very much the norm (Zeng et al. 2014, Ahmed et al. 2015, Bolón-Canedo and Alonso-Betanzos 2019, Mienye and Sun 2022, Mohammed and Kora 2023), regression/classification tasks may resort to 100+ algorithms(e.g. Fernández-Delgado et al. 2014). We recovered just 25 ENM algorithms from the literature (Table S1), but their free availability suggests that ENM is behind ML standards. Ensembles have higher prediction accuracy, decrease overfitting, overcome algorithmic biases (Abrahms et al. 2019, Austin et al. 2019, Mienye and Sun 2022, Sunny et al. 2025), and are better at dealing with imbalanced training sets (Barandela et al. 2003), high-dimensional, and/or noisy data (Rokach 2009, 2010, Mohammed and Kora 2023). Therefore, they provide better approximation to complex ecological functions (Coro et al. 2024), potentially revealing hidden patterns that a single algorithm may miss (Mahajan et al. 2023).

Almost 20 years after Araújo and New's (op. cit.), surveys of real world applications show that standards are still lacking, particularly when it comes to combining different modeling strategies into a single set of predictions. In this study, we revisit and extend that framework by introducing an information theory approach to ensemble learning. Instead of using a single generalization from one model class (MC) to generate predictions, as it is commonly done in ENM, our approach allows for the integration of predictions over IC and MCs, as intended by those authors. We extend their framework by using Shannon's entropy (H) Shannon (1948), to quantify the variability of binarized MC predictions across ICs. While among MC variation is expected, within-model prediction deviations with IC should be minimal. Otherwise, results will be sensitive to the stochasticity inherent to the subsampling procedures, which is the very definition of overfitting. H is a measure of disorder commonly used in information theory that is grounded on thermodynamics, hence it should naturally fit Araújo and New's (op. cit.) perspective of ENM. We show that model overfitting is not uniformly distributed in space and that its existence may be missed by the aforementioned performance measures (AUC, TSS, Kappa, etc.). Our approach weighs MC contributions inversely according to their degree of overfitting on a cell-by-cell basis, hence improving overall signal/noise ratio across the projection area, and highlighting consensual predictions among MC.

2 | MATERIAL AND METHODS

2.1 | The *Siderastrea* Atlantic complex

The genus *Siderastrea* Blainville 1830 (Scleractinia:Rhizangiidae) has three species in the Atlantic Ocean: *Siderastrea radians* (Pallas 1766), *Siderastrea siderea* (Ellis and Solander 1786) and *Siderastrea stellata* Verrill 1868, that make up the so called "*Siderastrea* Atlantic

complex" (SAC). Species boundaries definition within SAC is rather subjective due to overlap in diagnostic characters, although they are considered "good species" according to the Biological Species Concept (Menezes et al. 2013, 2014, Neves et al. 2010, Barbeitos et al. 2024). SAC species identification is biogeography reliant: a specimen not identified as *S. radians* must be *S. stellata* if it was collected off the Brazilian coast, or *S. siderea* if collected elsewhere on the western Atlantic Ocean (Laborel 1974). Nevertheless, Neves et al. (2010) re-identified a specimen collected by Hartt in the 19th century off the state of Bahia coast as *S. siderea*, yielding the first record of the species for Brazil. Additionally, Garcia et al. (2017) reported the occurrence of *S. stellata* in the Gulf of Mexico (Yucatán Peninsula) using morphological and genetic data. *S. siderea* and *S. stellata* may thus coexist in the Caribbean and in part of the Brazilian coast.

2.2 | Abiotic data

Environmental layers were obtained from the Bio-ORACLE (Assis et al. 2017, Tyberghein et al. 2012) with a resolution of 5 arcmin (~9.2 km at the equator). We selected bathymetry ($BTM - m$), average annual temperature ($SST - °C$), average annual seafloor salinity (using the practical salinity scale - $PSS\ g/kg$), chlorophyll A concentration ($Chl_a - mg \cdot m^{-3}$), and diffuse attenuation ($Dat - m^{-1}$). Because of obligatory endosymbiosis with algae know as zooxanthellae, whose excess photosynthates are traded for the host's nitrogenous excretes, *Siderastrea* spp. colonies are restricted to shallower waters. Depth, diffuse attenuation (i.e. the rate of light extinction as depth increases), and seawater chlorophyll A concentration are inversely correlated with photosynthetic capacity. Low temperatures at greater depths decrease calcareous carbonate accretion, hindering skeleton formation and thus colonies' growth rates(Lewis 1989, Sawall and Hochberg 2018).

When outside of optimal ranges, salinity increases physiological stress, interfering with photosynthesis, coral respiration, basal metabolism and mucus production (Ding et al. 2022, Pupier et al. 2024).

Variable redundancy was evaluated by examining a correlation matrix built using the R package *PerformanceAnalytics* (Peterson and Carl 2019), and retaining layers whose pairwise correlation coefficients were $-0.8 < r < 0.8$. To avoid inflating performance estimators (Barve et al. 2011, Lobo et al. 2008), layers were trimmed to fit the distribution of *S. stellata* (i.e. 45°-32°W;27°S-2°N), *S. siderea* (100°-60°W;2°S-45°N), and *S. radians* (100°W-30°E;31°S-34°N). Model training was performed using data from each species distribution's cutout, whereas generalizations were conducted across the whole area covered by the three species (see 2.5).

2.3 | Occurrence points

Occurrence points were obtained from peer-reviewed and "gray" literature. We collected 3,085 occurrences (1,572 for *S. siderea*, 1,060 for *S. radians* and 453 for *S. stellata*). Due to rounding errors, some occurrence points were placed on land when mapped

to environmental rasters and were thus "snapped" to the nearest cell with a complete abiotic dataset using the function `st_nearest_feature` from the package `sf` Pebesma and Bivand (2023). Only records shallower than 60 m were retained, in order to avoid projections onto cells beyond the recorded bathymetric range for species in the genus (Bongaerts et al. 2015).

2.4 | Model selection

Ensembles were built using nine different models, three from each of the labels proposed by Rangel and Loyola (2012), namely, "Fish Bowl", "Turbine", and "Vault" (for further details on each category, see the "Parameter selection and model tuning" in the SM). We chose BioClim (Booth et al. 2014), Gower (Carpenter et al. 1993), and Domain (Mahalanobis distance) (Farber and Kadmon 2003) as "Fish Bowl" (presence-only) models. Multivariate Adaptive Regression Splines (MARS) Friedman (1991), Generalized Linear (GLM - Nelder and Wedderburn 1972), and Additive (GAM - Guisan et al. 2002) Models were selected as "Turbine" representatives. We picked Deep Learning Artificial Neural Networks (ANN - LeCun et al. 2015), MaxEnt Phillips et al. (2006), and SVM Cortes and Vapnik (1995) from the "Vault" category.

2.5 | Model building and tuning

As some models need background points, 10,000 pseudoabsences were generated for the three cutouts (Barbet-Massin et al. 2012). In order to reduce the likelihood that pseudo-absences overlap with know occurrences, we used a 10 km buffer around all sampled occurrences within each cutout. The same threshold (see the section "Inter-reef distance and reef size estimation" in the SM for detailed justification of value choice) was used to filter occurrences using the `thin` function from the `spThin` package (Aiello-Lammens et al. 2015). The outcome of `spThin::thin` is stochastic, therefore, we replicated the selection of presences and pseudo-absences (P/PA) 120 times to get a diverse set of IC.

The package `blockCV` was used to reduce the influence of spatial autocorrelation on model tuning. With the aid of the R package `terra` Hijmans et al. (2026) we masked the rasters and employed the function `blockCV::cv_spatial_autocor` to estimate the median range at which autocorrelation of environmental variables breaks down in areas shallower than 60 m. We focused on shallow areas both because of potential coral growth and also because they tend display more variation in environmental conditions than the pelagic regions, due to their proximity to land masses. Ranges were separately estimated for each cutout and we used the function `blockCV::cv_spatial` to place a grid of hexagonal cells with sizes equivalent to the previously selected ranges onto the raster. This function was also used to randomly assign P/PA sampled points within these cells to different folds thus reducing spatial auto-correlation among the training and testing sets in downstream cross-validation experiments (CVEs).

P/PA replicates from 1 to 10 were used in parameter selection of "Turbine" models, i.e., the determination of the best parameter combination (see section "Parameter selection and model tuning" in the SM for the specifics on each model). Optimization was conducted separately on each replicate and distinct parameter combinations (=formulae) were retained. The best formula was selected via 5-fold cross validation experiments conducted on replicates 11-20. We first determined the suitability score that maximized True Skill Statistic (TSS - Allouche et al. 2006) for each combination of formula and replicate. Maximizing TSS also maximizes the sum of the model's specificity and sensitivity, thus, TSS is simply a different formulation of Youden's index, commonly used to set the diagnostic reference values in clinical tests (e.g. Kaivanto 2008). We then used this TSS maximization score (TSS_t) as a threshold value to binarize the continuous suitability scale and estimated the performance of each formula by computing its False Omission Rate (FOR). FOR is defined as:

$$FOR = \frac{FN}{FN + TN} \tag{1}$$

Where FN are the False Negatives (occurrences that were incorrectly scored as absences after binarization) and TN are true negatives (pseudo-absences that were correctly predicted by the model). This criterion favors models that are able to correctly predict all the occurrences in the training sets (True Positives - TP) while maintaining low commission error rates. Coral prevalence is low hence models should be able to detect rare instances without significant overprediction. Too many False Positives will indeed decrease the number of True Negatives in the denominator of equation 1, thus increasing FOR (we expect $TN \gg FN$ because the vast majority of the ocean is unsuitable to symbiotic corals). Because each formula was tested on all replicates, we first searched for the *highest* FOR scores among the 10 CVEs, thus establishing the worst-case scenario (i.e. the "floor" performance) for each formula within our evaluation universe. The selected combination of parameters was thus the one with the *best* (=lowest) FOR among the *worst* (=highest) values. The same approach was used to grid-search for best parameter combinations for BioClim and "Vault" models (see the SM for details).

2.6 | Model evaluation, training and generalization

After model tuning, each one of the remaining 100 P/PA replicates were sequentially submitted to all ENM algorithms, thus conditioning results returned by different models on the same training and testing data. Confusion matrices were generated at the end of each CVE (one per replicate), and used to compute the corresponding Area Under the Receiver Operating Characteristic Curve (AUC), sensitivity, specificity, FOR, TSS, and TSS_t. One generalization per replicate was conducted employing tuned models, that were trained on all corresponding occurrences and pseudo-absences. Binarization of generalized rasters was carried out using each replicate's TSS_t, yielding a set of 100 binary rasters for each species/model combination (9 models x 3 species x 100 replicates = 2,700 rasters). The same packages were used in tuning, training, and generalization (see SM for details). We employed

the R package `dismo` Hijmans et al. (2023) in Domain (Mahalanobis) and Gower CVEs, and `terra` in generalizations computed from these models. Mahalanobis distances were rescaled to the 0-1 range using the approach described in Etherington (2019).

2.7 | Ensemble learning

Occurrence suitability on cell c was estimated as

$$S(o) = \frac{\sum_{m=1}^{M}\sum_{i=1}^{N} b_{cmi}}{M \cdot N} \quad b \in \{0, 1\} \tag{2}$$

where b is model m's binarized outcome in the i-th cross-validation experiment (CVE), M is the number of models employed ($M = 9$) and N is the number of CVEs ($N = 100$). It is advisable to weight each model's contribution to the ensemble according to some criterion related to the model's global accuracy such as the AUC or TSS (e.g. Hao et al. 2020, Rose et al. 2024). When computed across all CVEs, species presence in cell c is the successful realization of a Bernoulli process with probability drawn from a beta distribution. The instability or uncertainty of model m may therefore be estimated as the expected entropy of the Bernoulli random variable X, given by Shannon's formula (Gill 2005)

$$H(X_m) = -\sum_{b=0}^{1} p_{cmb} \cdot log_2 p_{cmb} \tag{3}$$

where p_{cmb} is the proportion of model m's predicted absences/presences $(0/1)$ in cell c. Ecologists should recognize in this expression the familiar Shannon-Wiener diversity index. An equivalent formulation is the binary entropy function, expressed as

$$H(X_m) = -p_{cmb} \cdot log_2 p_{cmb} - (1 - p_{cmb}) \cdot log_2 p_{cmb} \tag{4}$$

When taken using log_2, entropy is measured in bits. Because there are only two possible outcomes (or microstates) in binarized rasters, the maximum possible entropy is given by $log_2 2 = 1 \quad bit \cdot cell^{-1}$. The entropy of model m in cell c is maximized ($H(X_m) = 1$) when $p_{cmi} = 0.5$. If such is the case, the model's ability to predict species presence or absence in a given cell is no better than a random classifier. Conversely, $H(X_m) = 0$ when all outcomes for model m and cell c are either 0 or 1 meaning that the model's forecast has maximum consistency across replicates. We can thus compute weights for the m-th model in cell c by using the expression (Wang and Zhan 2012):

$$w_{cm} = \frac{1 - H(X_m)}{M - \sum_{m=1}^{M} H(X_m)} \tag{5}$$

It is easy to see that

$$\sum_{m=1}^{M} H(X_m) \to 0 \implies w_{cm} \to \frac{1}{M} \tag{6}$$

i.e., when all models tend to minimum entropy, weights will be equally distributed among them. It is also easy to verify that

$$H(X_m) \to 1 \implies w_{cm} \to 0 \tag{7}$$

i.e., a random classifier will have zero weight on the ensemble. Additionally, the weights always satisfy two conditions:

$$0 \leq w_{cm} \leq 1 \tag{8}$$

$$\sum_{m=1}^{M} w_{cm} = 1 \tag{9}$$

Weighted suitability was computed as

$$S_w(o) = \frac{\sum_{m=1}^{M} \sum_{i=1}^{N} w_{cm} \cdot b_{cmi}}{M \cdot N} \quad b \in \{0, 1\} \tag{10}$$

From an information theory perspective, ENM models are (probabilistic) species detection sensors embedded in each cell. Entropy weighting is thus analogous to a noise filtering mechanism applied to a communication channel: if some models overfit the data, their contribution to the ensemble will be weighted down and results will depend on their upweighted counterparts. Ensemble efficiency is thus maximized when all models are equally weighted. Formally, efficiency is expressed as a ratio that quantifies the entropy deficit in the ensemble. Recall that the maximum entropy is computed as $log_b(\chi)$ where χ is the alphabet size (or the number of possible microstates) and b is the chosen base. In our case $\chi = M = 9$ and $b = 2$, hence ensemble efficiency on cell c is computed as the normalized Shannon entropy (Wilcox 1967)

$$\eta_c(E) = \frac{H_c(E)}{H_{max}(E)} = -\sum_{m=1}^{M} \frac{w_{cm} \cdot log_2 w_{cm}}{log_2 M} \tag{11}$$

Ensemble efficiency was combined with suitability in reliability maps analogous to the Maps of Biogeographical Ignorance (MoBIS) of Tessarolo et al. (2021). Reliability is defined in two dimensions: it decreases as weighted suitability approach 0.50, and it also decreases with the unevenness in model weight distribution. If the weighted suitability of a cell is close to 0.50, but ensemble efficiency is high, this means that models consistently disagreed about species presence on that cell across all CVEs because of fundamental algorithmic differences among them. On the other hand, if this probability is close to 1.0, but efficiency is low, this means that most models overfitted the data in the cell, and the high occurrence probability was issued by a minority of the ensemble, also making the outcome unreliable. High reliability thus requires extreme suitability issued by balanced ensembles.

2.7.1 | Entropy

Cell suitability may be presented in the usual form of one choropleth raster map per model and species. Likewise, it is also possible to generate entropy rasters that represent the overfitting surface of each model (e.g. Mesgaran et al. 2014, Tessarolo et al. 2021). Shannon entropy (H) is additive, so global uncertainty is the simple sum of cell H scores. Variances associated with these sums were approximated by bootstrapping, i.e., re-sampling the binarized rasters with replacement before computing H. Statistical significance of H differences among models and species was determined by fitting Generalized Linear Mixed Models (GLMMs) to the bootstrapped scores (see SM for details). As explained in 2.5, the same training and validation sets were sequentially used in generalizations across all models. Hence, although each model will yield different prediction rasters, they are not entirely independent from each other because they are derived from the same underlying CVE. Pseudo-replicates were constructed by re-sampling CVE *indices* and then using these indices to assemble the corresponding raster sets for each of the 9 models, maintaining the repeated measures design.

3 | RESULTS

3.1 | Unweighted vs. weighted suitabilities

Abiotic variables correlations were all between -0.80 and 0.80 for the three species, thus all environmental layers were kept during ENM (Supporting Information Figs. S1-S3). *S. radians* occurrences after thinning ranged from 294 to 298 across P/PA replicates, 360-363 in the case of *S. siderea* and 80 for all replicates in the case of *S. stellata*. Fig. 1 shows maps of unweighted (left column) and weighted (middle column) suitabilities. *S. radians* (Figs. 1A and 1B) and *S. siderea* (Figs. 1D and 1E) had similar projected distributions in areas with documented occurrences of coral populations, such as the Northeastern coast of Brazil, the Caribbean, Gulf of Mexico, both Florida coasts, Cape Verde, and the Gulf of Guinea. The stretch of African coast that goes from Senegal to Liberia was also recovered as a suitable area, although no corresponding occurrences were recovered from the literature.

Entropy weighting had clear effects over predicted occurrences for both species (Figs. 1A, 1B, 1D, and 1E). Figs. 1C and Figs. 1D show differences between unweighted and weighted suitability surfaces for this species. These results are in good agreement with the majority of individual models, shown in Figs. S5 and S6. The exceptions were Mahalanobis, which severely underpredicted occurrences of these species (Figs. S5C and S6C), and ANN, that overpredicted occurrences in pelagic areas (Figs. S5G and S6G) This was also true for SVM in the case of *S. siderea* (Fig. S6I).

Predictions for *S. stellata* were substantially different from the two other species. Higher suitability areas were concentrated along the Eastern Brazilian coast, Colombia, Venezuela, and the Yucatán Peninsula. A significant range across the Caribbean

was also highlighted as moderately suitable (Fig. 1G). Unlike the two other species, weighting did not substantially reduce the distribution of low suitability pixels in open Atlantic and Pacific Oceans, areas where symbiotic corals are not found (Figs 1E and 1H). Spurious occurrences were introduced by GAM, GLM, and MARS (Fig. S6D-F), and persisted even after entropy weighting (Fig. 1H) because they were consistently predicted across replicates (see overfitting surfaces in Fig S9D-F).

Ensemble efficiency and reliability maps are shown in Fig.2. It is interesting that ensembles are unbalanced along many coastal regions, but, with the exception of Mahalanobis, all methods contributed in predicting the occurrence of *S. radians* with high reliability across the Caribbean, off the Liberia-Sierra Leone stretch, and the tropical coast of South America (Fig.2A-B). Ensembles were less efficient in the case of *S. siderea* along the Western part of the Gulf of Mexico and northern Florida coasts but, otherwise, reliable predictions (Fig.2C-D) are well correlated with *S. radians*. Efficiency and reliability were the lowest in *S. stellata*, with the notable exception of the Yucatán peninsula and a narrower section along Colombia/Venezuela and the Eastern Brazilian coast (Fig.2E-F). Spurious predictions introduce by "Turbine" methods persist in these maps, on vast areas off Chile, Peru, Southwestern and Northwester African coasts. Efficiency rasters show that ensemble predictions are stable along these areas.

Figs. S11-S13 decompose efficiency into model weights. Fish bowl model votes were upweighted in all cases due to these models low uncertainty but, because Mahalanobis predicted species absence nearly everywhere, this had the side effect of decreasing reliability of occurrence prediction for all species (2). Figs. S11-S13 also show how ANN predictions were mostly downweighted for every species, followed by SVM, although this latter method did offer significant contributions in geographically restricted areas. "Turbine" models votes show a mixed pattern of up- and downweigthing, but the areas of spurious predictions in Fig.2F were slightly downweighted or even upweighted in some cases, indicating consistency in these methods predictions.

3.2 | Entropy vs. TSS

Fig. 3a shows the bootstrapped uncertainties (H) for all models grouped by species. Back transformed least-squares means (dots) estimated using Generalized Mixed Models (GLMM - see the section "GLMM analysis of Entropy and TSS scores" in the SM for details on model fitting) are close to observed H (crosses). BioClim was the most stable model (since Mahalanobis stability happened at the expense of prediction accuracy). There was no clear separation between the "Turbine" and Vault categories, with the exception of ANN, that was the only method consistently scoring above 100 kbits (see also Figs. S7-S9). "Turbine" and vault performances were more unstable for *S. stellata* than for other species, probably because of the low number of its occurrences (n = 80) used in model training (see also Fig. S9)

Boxplots in Fig. 3b represent distributions TSS scores obtained from the cross-validation experiments. The fitted GLMM was also in good agreement with the data (Fig. S15). Letters represent within species pairwise contrasts (using least-squares means,

see SM for details) that were **not significantly** different. For simplicity, results for cross-species comparisons are not shown, but every same-model pairwise contrast between different species was significant, with the exception of ANN for *S. radians* and *S. siderea* (as shown in Fig. 3b, see also Tab. S3). GLMM revealed significant model x species interaction, but boxplots suggest better fitting in *S. radians* than in other species. The very poor performance of Mahalanobis distance was captured by the boxplots, which also flagged ANN's instability. Nevertheless, spurious predictions by "Turbine" models, especially in the case of *S. stellata*, were entirely missed in this analysis since they were invariably ranked among top models, with median TSS scores equal or greater than 0.80. This is not surprising given that most overprediction in pelagic areas occurred outside of the training area for *S. stellata*, which was concentrated along the Brazilian coast (Fig. S6).

4 | DISCUSSION

In this study, we introduced a new ensemble learning approach to ecological niche modeling. By retaining all rasters generated during random sub-sampling, we managed to capture the effects of variation in initial conditions (IC) and model classes (MC), and to integrate these two sources of variation into a single probabilistic framework, as advocated by Araújo and New (2007). Estimating model's performance variability on each cell using information theory allowed for the construction of overfitting surfaces that highlighted model shortcomings across the projection area, even in the absence of real or simulated species occurrences. Ensemble contamination by overfitting was dampened by entropy weighting and reliability maps also reduced spurious overprediction due to over reliance on ensemble minorities. Finally, fine-grained weighting revealed that even models prone to overfitting may still offer significant local contributions to ensemble predictions, and that the traditional use of global performance measures as proxy to weight committee voting does not adequately capture spatial heterogeneity intrinsic to ENM predictions. The "No Free Lunch" Theorem (NFLT) (Friedman 1975) is a well-known principle in mathematical optimization. The theorem states that no optimization strategy can outperform its alternatives in every possible scenario (Ho and Pepyne 2002), and this is naturally extensible to ENM (Qiao et al. 2015). NFLT makes a strong argument in favor of ensembles over individual models, but we believe that our results have extended the theorem beyond its original scope: not only a model will not outperform its alternatives in every scenario, it will not outperform its alternatives in every cell of the raster, even if it is globally the best possible choice. Spatial variation in overfitting was not revealed by global measures normally used to benchmark ML algorithms, especially when occurring outside of the training area. This is particular troublesome given that most practitioners generalize predictions from one set of IC, which may not be at all representative of the extreme within-model variation observed in *S. siderea* and *S. stellata*.

4.1 | Uncertainty in ENM

The sources of variability in ENM and species distribution model (SDM) have been extensively discussed in the literature and go beyond what was originally proposed by (Araújo and New 2007). For instance, Iessarolo et al. (2021) accounted for variation in data quality (first discussed in Ladle and Hortal 2013) through Maps of Biogeographical Ignorance (MoBIS) that are integrated with SDM predictions. The basic idea is to increase the reliability of ENM/SDM projections by factoring data uncertainty into them. We proposed an analogous approach that integrates weighted suitability with ensemble efficiency (a measure of model uncertainty) in order to maximize prediction reliability. Algorithmic uncertainty is more closely related to area accessibility i.e., how reliable model extrapolations really are. Several methods based on this premise have been developed (e.g. *MESS* - Elith et al. 2010, *MOP* - Owens et al. 2013, Cobos et al. 2024, *ExDet* - Mesgaran et al. 2014, *SHAPE* Velazco et al. 2024, etc.) and (Meyer and Pebesma 2021) has coined the term Area of Applicability (AOA) to designate areas onto which model projections have high reliability. Large model overfitting surface means low applicability because the predictions of the models will vary widely with initial conditions. According to *dismo's* function *mess* (Fig. S16), the Caribean, South American and Western African coasts were accessible to all three species, but these results seem to be mainly dictated by depth (compare Fig. S16 and S13). In contrast, overfitting maps (Figs. S7-S9) show that it occurs even in areas of intermediate habitat suitability found **within** the range of known occurrences. This is actually not surprising because these are habitats with borderline abiotic conditions and hence models should return intermediate suitability values. As the cutoff threshold will naturally vary across CVEs, these cells are expected be binarized as either presence or absence with higher frequency than those with more extreme suitability scores. If all models in the ensemble are overfit equally, or not at all, equations (7) and (8) guarantee high ensemble efficiency as weights will converge to $1/M$. However, our results show the opposite in borderline cells, highlighting that predictions are being dominated by stabler models (Fig. 2). Readers should be aware that low model entropy does not guarantee good accuracy - stable models may be biased towards high commission or omission rates. This is a well known effect in machine learning, namely the "bias-variance" tradeoff. Models that approximate the target function using oversimplifying assumptions do not yield sufficient flexibility to capture underlying relationships in the training sets, leading to a large average error (bias) even though their predictions will have low variance (Drake 2014, Tourinho and Vale 2023). They may underfit the data and have low predictive power both within and outside the training areas. This is highlighted by Mahalanobis, that severely underfitted the data collected for every species, even within the training area, as flagged by the low TSS scores in (Fig. **??**b). We opted for keeping the model in the analysis, but practitioners would probably prune it from the ensemble.

4.2 | Entropy vs. TSS

According to TSS, most models exceeded the minimum acceptable threshold and within species average performances were either statistically indistinguishable or - as a consequence of the powerful repeated-measures design - most differences were too small in most pairwise comparisons to offer real guidance on model choice (Fig. 3E). AUCs were even more optimistic than TSS - note the right skewed distribution along the x-axis in the rug plot on Fig. S17, and the lower cluster that corresponds to Mahalanobis scores. Hence, had we employed AUCDIFF, the most widely used measure of overfitting in ENM (Warren and Seifert, 2011), this test would have very likely suggested excellent fit across our entire model suite because virtually all AUCTEST scores were over 0.90 (Fig. S20). AUCs provide inflated performance estimates as they are sensitive to geographic extent and prevalence: since absences are easily and correctly predicted, models appear better than they really are (Lobo et al., 2008). This will necessarily be the case of warm, shallow water symbiotic corals, found in a small fraction of the ocean floor. TSS estimates are supposedly insensitive to prevalence (Allouche et al., 2006), but empirical studies with simulated data (e.g. Leroy et al., 2018) do show a tendency towards overestimation of method performance when species prevalence is low, as in the case of *Siderastrea* spp. At any rate, even unbiased methods that relied on know occurrences would certainly fail to predict overfitting outside of the training area.

4.3 | Implications for the *Siderastrea* Atlantic complex

Yucatán peninsula was flagged by the reliability surface as an area as suitable for *S. stellata* as the Brazilian coast, supporting the findings by Garcia et al. (2017) (Fig. 2F). Conversely, the Brazilian coast was also recovered as adequate as the Caribbean for *S. siderea* (Fig. 2D), validating Neves et al. (2010) and the IA-based morphological analysis of Barbeitos et al. (2024), which provides evidence of the species presence in Rio Grande do Norte State, in Brazil. The stretch of Western African coast made up by Liberia, Sierra Leone, Guinea, Guinea-Bissau, Gambia, and Senegal, was also projected as suitable habitat for *S. radians* and *S. siderea*, but no corresponding record of the genus was recovered from the literature. Therefore, this region was not included among the occurrences used in model calibration. There is however one specimen (USNM 96978) vouchered at the National Museum of Natural History (Smithsonian Institution, Washington DC, USA) that was collected off Sierra Leone and identified as *S. radians* by Helmuth Zibrowius, in 1971. GBIF also holds two human observations from May 2023 (references 163444694 and 163342223), imported from the citizen-science network *iNaturalist*. Although species-level taxonomy of *Siderastrea* is controversial, the genus is easily recognizable and very distinctive from the other genera in its family. Thus, the images associated with the observations leave no doubt that *Siderastrea* was indeed present in West Africa as recently as 2023.

5 | CONCLUSIONS

The entropy weight method (EWM) is not a novel technique in itself, being extensively employed in other machine learning applications (e.g. Rokach 2009, Qu et al. 2022) but, to our knowledge, this is the first time it was applied to ENM. Even if accurate, the use of a single method (which is seemingly standard in ENM according to Qiao et al. 2015), always leaves room for the possibility that the observed results may be conditioned on the specific method choice. Ensembles of methods with varying underlying assumptions bring more confidence to ENM predictions, particularly when generalized beyond training areas. We thus believe that ecologists should benefit from the adoption of ensemble learning as standard practice in ENM.

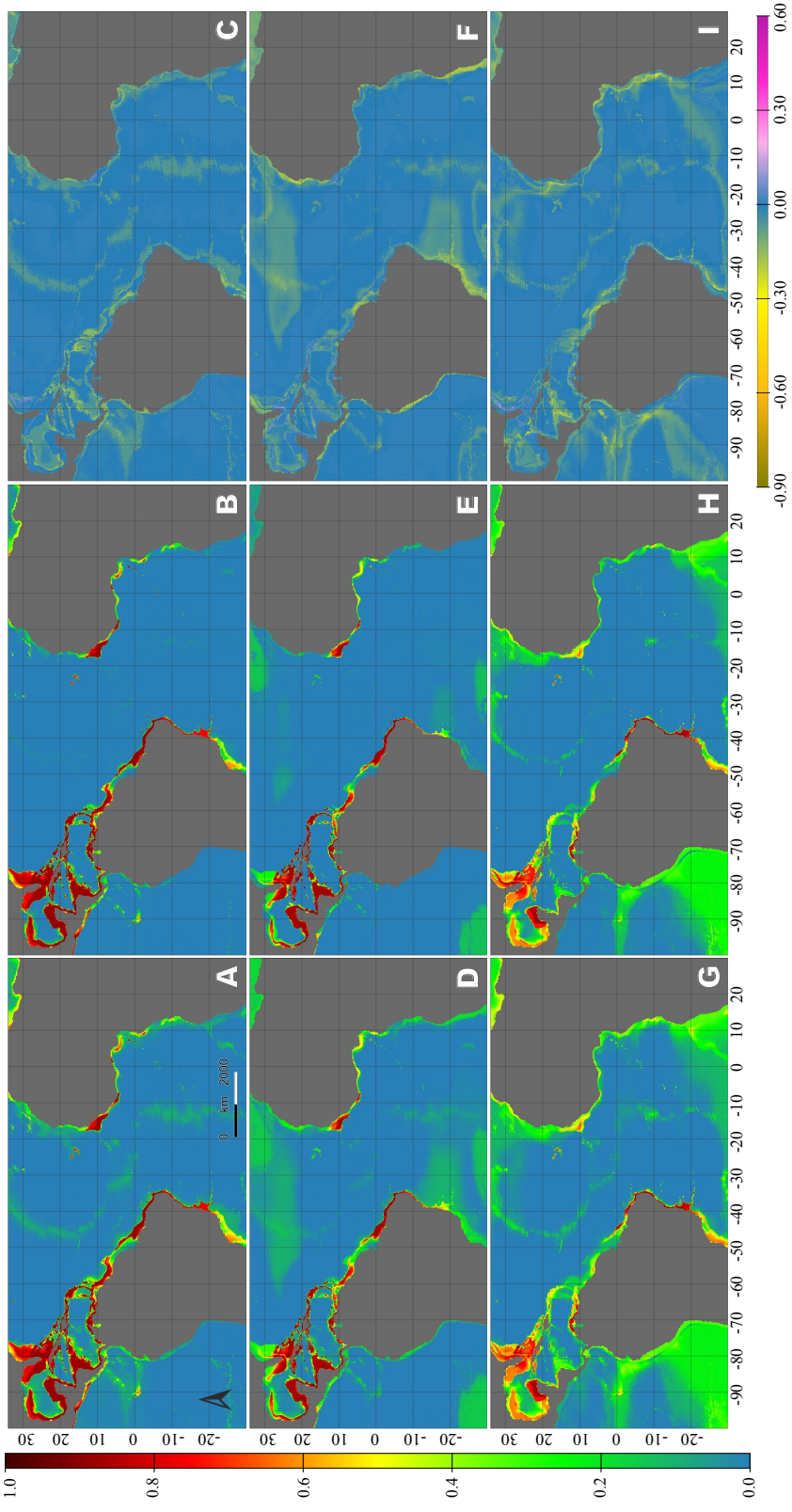

FIGURE 1 Maps of species occurrence probabilities computed from unweighted ensembles: A) *Siderastrea radians*. D) *S. siderea*. G) *S. stellata*. Maps of entropy-weighted ensembles: B) *Siderastrea radians*. E) *S. siderea*. H) *S. stellata*. Differences between results from weighted and unweighted ensembles for each species: C) *Siderastrea radians*. F) *S. siderea*. I) *S. stellata*. Negative values indicate a reduction in putative commission error (i.e. reduction in false positives) while positive values indicate a reduction in putative omission error (false negatives) after weighting. The color scale for the last column of images is shown at its bottom.

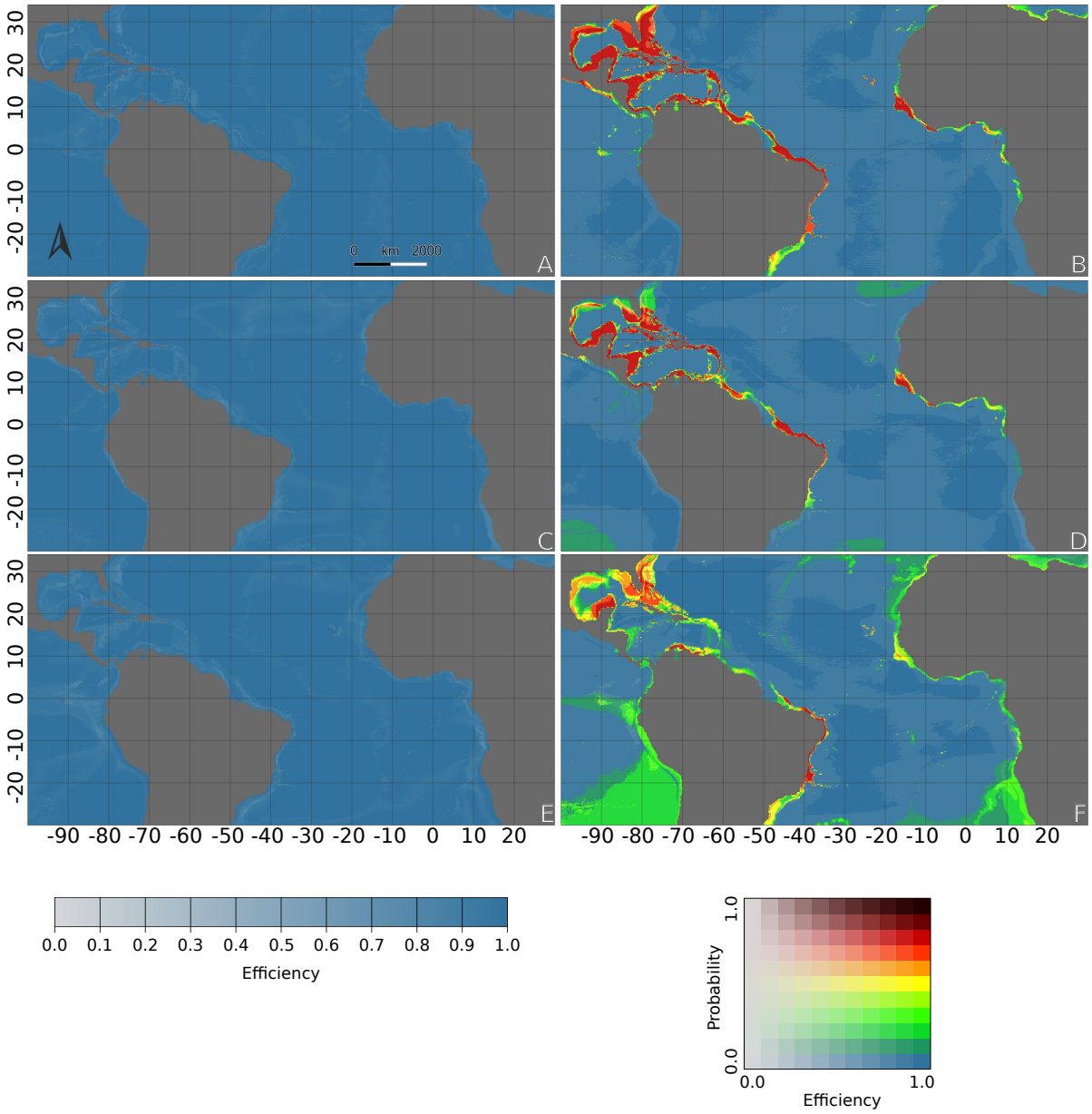

FIGURE 2 Maps of ensemble efficiency (left column) and prediction reliability (right column). Species are arranged in rows: A and B *S. radians*, C and D *S. siderea*, E and F *S. stellata*. Reliability is a combination of ensemble efficiency and entropy-weighted probabilities, represented by the bi-dimensional color key on the lower right side of the figure. Reliability in species occurrence increases as probabilities tend to either 0 or 1 and it also increases with efficiency, or the evenness of model weighting in the ensemble.

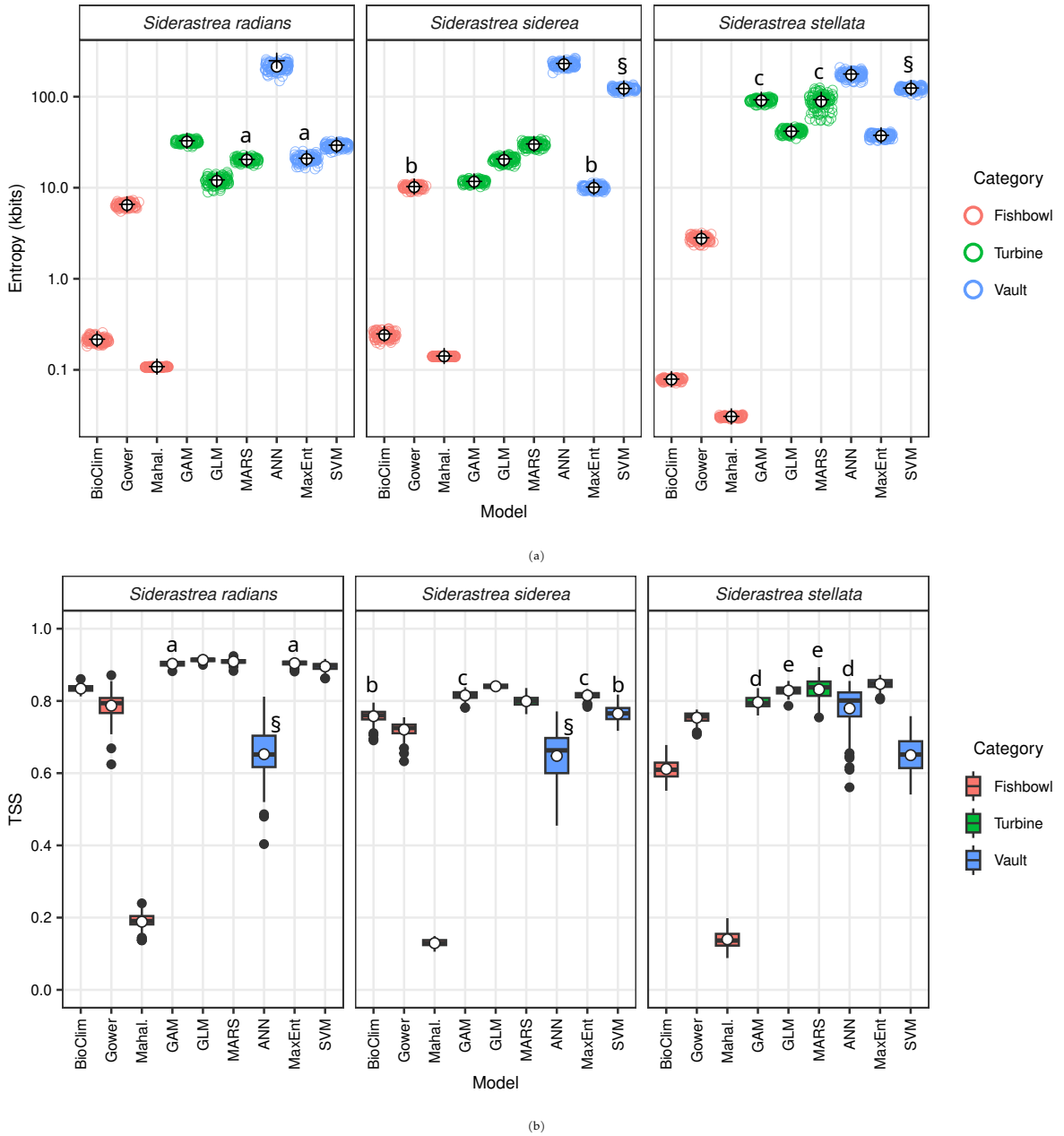

FIGURE 3 Performance evaluation. (a) Model entropy (H) expressed in kilobits (note the log scale in the y-axis). White-filled dots are the are the least-square means estimated from the best-fit GLMM model (see SM), crosses observed entropies for each raster and empty dots are the bootstrapped entropies (n = 100), color coded according to model category. Letters indicate **non-significant** pairwise comparisons of least-square means at $\alpha = 0.05$. (b) Distributions of TSS scores computed from cross-validation experiments. Horizontal lines are medians, vertical lines are non-outlier and boxes are interquartile ranges, black dots are outliers and white dots are the least-squares means estimated from the best-fit GLMM model (see SM). Letters indicate **non-significant, within species** pairwise comparisons of least-square means at $\alpha = 0.05$. Symbols indicate **non-significant, between species** comparison of the same method (see Table S2 for all pairwise comparisons). Standard error associated with least-squares means were too small to be displayed thus were not included in either graph.

AUTHOR CONTRIBUTIONS

H.J. Schipanski and M.S. Barbeitos designed the study. H.J. Schipanski collected the data. H.J. Schipanski, L.B. Perroni, C.F. Sguario and M.S. Barbeitos wrote computer code and analyzed the data. H.J. Schipanski, L.B. Perroni and M.S. Barbeitos wrote the manuscript.

ACKNOWLEDGMENTS

We thank Drs. Fernanda T. Brum and Maurício O. Moura for comments on an earlier draft of this manuscript.

FINANCIAL DISCLOSURE

None reported.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

REFERENCES

Abrahms, B., Welch, H., Brodie, S., Jacox, M. G., Becker, E. A., Bograd, S. J., Irvine, L. M., Palacios, D. M., Mate, B. R. & Hazen, E. L., 2019. Dynamic ensemble models to predict distributions and anthropogenic risk exposure for highly mobile species. – Divers. Distrib. 25: 1182–1193.

Ahmed, M. A. O., Didaci, L., Fumera, G. & Roli, F. 2015. An Empirical Investigation on the Use of Diversity for Creation of Classifier Ensembles. – In: Schwenker, F., Roli, F. & Kittler, J. (Eds.), Multiple Classifier Systems, vol. 9132. Springer International Publishing, pp. 206–219.

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B. & Anderson, R. P., 2015. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. – Ecography 38: 541–545.

Allouche, O., Tsoar, A. & Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). – J. Appl. Ecol. 43: 1223–1232.

Araújo, M. B. & New, M., 2007. Ensemble forecasting of species distributions. – Trends Ecol. Evol. 22: 42–47.

Assis, J., Tyberghein, L., Bosch, S., Verbruggen, H., Serrão, E. A. & Clerck, O. D., 2017. Bio-ORACLE v2.0: Extending marine data layers for bioclimatic modelling. – Global Ecol. Biogeogr. 27: 277–284.

Austin, R. A., Hawkes, L. A., Doherty, P. D., Henderson, S. M., Inger, R., Johnson, L., Pikesley, S. K., Solandt, J.-L., Speedie, C. & Witt, M. J., 2019. Predicting habitat suitability for basking sharks (*Cetorhinus maximus*) in UK waters using ensemble ecological niche modelling. – J. Sea Res. 153: 101767.

Barandela, R., Valdovinos, R. & Sánchez, J., 2003. New Applications of Ensembles of Classifiers. – Pattern Anal. Appl. 6: 245–256.

Barbeitos, M. S., Pérez, F. A., Olaya-Restrepo, J., Winter, A. P. M., Florindo, J. B. & Laureano, E. E., 2024. AI-based coral species discrimination: A case study of the *Siderastrea* Atlantic Complex. – PLoS ONE 19: e0312494.

Barbet-Massin, M., Jiguet, F., Albert, C. H. & Thuiller, W., 2012. Selecting pseudo-absences for species distribution models: how, where and how many?. – Methods Ecol. Evol. 3: 327–338.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., Soberón, J. & Villalobos, F., 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. – Ecol. Model. 222: 1810–1819.

Blainville, H. M. 1830. Zoophytes. – In: de Candolle, A. (Ed.), Dictionnaire des Sciences Naturelles, vol. 60. Paris, pp. 310–358.

Bolón-Canedo, V. & Alonso-Betanzos, A., 2019. Ensembles for feature selection: A review and future trends. – Inf. Fusion 52: 1–12.

Bongaerts, P., Carmichael, M., Hay, K. B., Tonk, L., Frade, P. R. & Hoegh-Guldberg, O., 2015. Prevalent endosymbiont zonation shapes the depth distributions of scleractinian coral species. – R. Soc. Open Sci. 2: 140297.

Booth, T. H., Nix, H. A., Busby, J. R. & Hutchinson, M. F., 2014. Bioclim: the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. – Divers. Distrib. 20: 1–9.

Carpenter, G., Gillison, A. N. & Winter, J., 1993. Domain: a flexible modelling procedure for mapping potential distributions of plants and animals. – Biodivers. Conserv. 2: 667–680.

Cerasoli, F., D'Alessandro, P. & Biondi, M., 2022. Fine-tuned ecological niche models unveil climatic suitability and association with vegetation groups for selected *Chaetocnema* species in South Africa (Coleoptera: Chrysomelidae). – Diversity 14: 100.

Cobos, M. E., Owens, H. L., Soberón, J. & Peterson, A. T., 2024. Detailed multivariate comparisons of environments with mobility oriented parity. – Front. Biogeogr. 17: e132916.
Coro, G., Sana, L. & Bove, P., 2024. An open science automatic workflow for multi-model species distribution estimation. – Int. J. Data Sci. Anal. 20: 1131–1150.
Cortes, C. & Vapnik, V., 1995. Support-vector networks. – Mach. Learn. 20: 273–297.
Ding, D.-S., Patel, A. K., Singhania, R. R., Chen, C.-W. & Dong, C.-D., 2022. Effects of Temperature and Salinity on Growth, Metabolism and Digestive Enzymes Synthesis of *Goniopora columna*. – Biology 11: 436.
Drake, J. M., 2014. Ensemble algorithms for ecological niche modeling from presence-background and presence-only data. – Ecosphere 5: art76.
Drake, J. M., Randin, C. & Guisan, A., 2006. Modelling ecological niches with support vector machines. – J. Appl. Ecol. 43: 424–432.
Elith, J., Kearney, M. & Phillips, S., 2010. The art of modelling range-shifting species. – Methods Ecol. Evol. 1: 330–342.
Ellis, J. & Solander, D. 1786. The natural history of many curious and uncommon zoophytes collected from various parts of the globe. – London: Benjamin White and Son.
Etherington, T. R., 2019. Mahalanobis distances and ecological niche modelling: correcting a chi-squared probability error. – PeerJ 7: e6678.
Farber, O. & Kadmon, R., 2003. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. – Ecol. Model. 160: 115–130.
Feng, X., Park, D. S., Walker, C., Peterson, A. T., Merow, C. & Papeş, M., 2019. A checklist for maximizing reproducibility of ecological niche models. – Nat. Ecol. Evol. 3: 1382–1395.
Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems?. – J. Mach. Learn. Res. 15: 3133–3181.
Fielding, A. H. & Bell, J. F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. – Environ. Conserv. 24: 38–49.
Friedman, J. H., 1991. Multivariate Adaptive Regression Splines. – Ann. Stat. 19: 1 – 67.
Friedman, M. 1975. There's no such thing as a free lunch. – Open Court Publishing Co ,U.S.
Garcia, C., Campos, J., Tello-Musi, J. L., Forsman, Z., Muñoz, J., Monsalvo, A. & Arias, E., 2017. Comparative Molecular and Morphological Variation Analysis of *Siderastrea* (Anthozoa, Scleractinia) Reveals the Presence of *Siderastrea stellata* in the Gulf of Mexico. – Biol. Bull. 232: 58–70.
Gibbs, J. W. 1902. Elementary principles in statistical mechanics: Developed with especial reference to the rational foundation of thermodynamics. – Cambridge University Press.
Gill, J., 2005. An entropy measure of uncertainty in vote choice. – Elect. Stud. 24: 371–392.
Guisan, A., Edwards, T. C. & Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. – Ecol. Model. 157: 89–100.
Hao, T., Elith, J., Guillera-Arroita, G. & Lahoz-Monfort, J. J., 2019. A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. – Divers. Distrib. 25: 839–852.
Hao, T., Elith, J., Lahoz-Monfort, J. J. & Guillera-Arroita, G., 2020. Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. – Ecography 43: 549–558.
Hijmans, R. J., Brown, A. & Barbosa, M., 2026. terra: Spatial Data Analysis. – <https://CRAN.R-project.org/package=terra>.
Hijmans, R. J., Phillips, S., Leathwick, J. & Elith, J., 2023. dismo: Species Distribution Modeling. – <https://CRAN.R-project.org/package=dismo>.
Ho, Y. & Pepyne, D., 2002. Simple explanation of the no-free-lunch theorem and its implications. – J. Optim. Theory Appl. 115: 549–570.
Kaivanto, K., 2008. Maximization of the sum of sensitivity and specificity as a diagnostic cutpoint criterion. – J. Clin. Epidemiol. 61: 517–518.
Laborel, J. 1974. West african reef corals: An hypothesis on their origin. – In: Proceedings of the 2nd International Coral Reef Symposium, vol. 1, pp. 425–443.
Ladle, R. & Hortal, J., 2013. Mapping species distributions: living with uncertainty. – Front. Biogeogr. 5: 8–9.
LeCun, Y., Bengio, Y. & Hinton, G., 2015. Deep learning. – Nature 521: 436–444.
Leroy, B., Delsol, R., Hugueny, B., Meynard, C. N., Barhoumi, C., Barbet-Massin, M. & Bellard, C., 2018. Without quality presence–absence data, discrimination metrics such as TSS can be misleading measures of model performance. – J. Biogeogr. 45: 1994–2002.
Lewis, J. B., 1989. Spherical growth in the caribbean coral *Siderastrea radians* (Pallas) and its survival in disturbed habitats. – Coral Reefs 7: 161–167.
Lobo, J. M., Jiménez-Valverde, A. & Real, R., 2008. AUC: A misleading measure of the performance of predictive distribution models. – Global Ecol. Biogeogr. 17: 145–151.
Mahajan, P., Uddin, S., Hajati, F. & Moni, M. A., 2023. Ensemble Learning for Disease Prediction: A Review. – Healthcare 11: 1808.
Melo-Merino, S. M., Reyes-Bonilla, H. & Lira-Noriega, A., 2020. Ecological niche models and species distribution models in marine environments: A literature review and spatial analysis of evidence. – Ecol. Model. 415: 108837.
Menezes, N., Neves, E., Barros, F., Kikuchi, R. K. P. & Johnsson, R., 2013. Intracolonial Variation in *Siderastrea* de Blainville, 1830 (Anthozoa, Scleractinia): Taxonomy under Challenging Morphological Constraints. – Biota Neotrop. 13: 108–116.

Menezes, N., Neves, E., Kikuchi, R. K. P. & Johnsson, R., 2014. Morphological Variation in the Atlantic Genus *Siderastrea* (Anthozoa, Scleractinia). – Pap. Avulsos Zool. 54: 199–208.
Mesgaran, M. B., Cousens, R. D. & Webber, B. L., 2014. Here be dragons: a tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. – Divers. Distrib. 20: 1147–1159.
Meyer, H. & Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. – Methods Ecol. Evol. 12: 1620–1633.
Mienye, I. D. & Sun, Y., 2022. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. – IEEE Access 10: 99129–99149.
Mohammed, A. & Kora, R., 2023. A comprehensive review on ensemble deep learning: Opportunities and challenges. – J. King Saud Univ. Comput. Inf. Sci. 35: 757–774.
Muscarella, R., Galante, P. J., Soley-Guardia, M., Boria, R. A., Kass, J. M., Uriarte, M. & Anderson, R. P., 2014. ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. – Methods Ecol. Evol. 5: 1198–1205.
Nelder, J. A. & Wedderburn, R. W. M., 1972. Generalized linear models. – J. R. Stat. Soc. Ser. A (Gen.) 135: 370–384.
Neves, E., Silveira, F., Pichon, M. & Johnsson, R., 2010. Cnidaria, Scleractinia, Siderastreidae, *Siderastrea siderea* (Ellis and Solander, 1786): Hartt Expedition and the first record of a Caribbean siderastreid in tropical Southwestern Atlantic. – Check List 6: 505–510.
Owens, H. L., Campbell, L. P., Dornak, L. L., Saupe, E. E., Barve, N., Soberón, J., Ingenloff, K., Lira-Noriega, A., Hensz, C. M., Myers, C. E. & Peterson, A. T., 2013. Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. – Ecol. Model. 263: 10–18.
Pallas, P. S. 1766. Elenchus zoophytorum. – Hagae-Comitum: Varrentrapp F.
Pearson, R. G., Thuiller, W., Araújo, M. B., Martinez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T. P. & Lees, D. C., 2006. Model-based uncertainty in species range prediction. – J. Biogeogr. 33: 1704–1711.
Pebesma, E. & Bivand, R. 2023. Spatial Data Science: With applications in R. – Chapman and Hall/CRC.
Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M. & Araújo, M. B. 2011. Ecological niches and geographic distributions. – Princeton University Press.
Peterson, B. G. & Carl, P., 2019. PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis. – <https://CRAN.R-project.org/package=PerformanceAnalytics>.
Phillips, S. J., Anderson, R. P. & Schapire, R. E., 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.
Phillips, S. J. & Dudík, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. – Ecography 31: 161–175.
Pupier, C. A., Grover, R., Rottier, C. & Ferrier-Pagès, C., 2024. Impact of seawater warming and nutrient deprivation on the physiology and energy metabolism of corals. – Front. Mar. Sci. 11: 1423034.
Qiao, H., Soberón, J. & Peterson, A. T., 2015. No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. – Methods Ecol. Evol. 6: 1126–1136.
Qu, W., Li, J., Song, W., Li, X., Zhao, Y., Dong, H., Wang, Y., Zhao, Q. & Qi, Y., 2022. Entropy-Weight-Method-Based Integrated Models for Short-Term Intersection Traffic Flow Prediction. – Entropy 24: 849.
Rangel, T. F. & Loyola, R. D., 2012. Labeling ecological niche models. – Nat. Conservacao 10: 119–126.
Rokach, L., 2009. Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. – Comput. Stat. Data Anal. 53: 4046–4072.
Rokach, L., 2010. Ensemble-based classifiers. – Artif. Intell. Rev. 33: 1–39.
Rose, M. B., Velazco, S. J. E., Regan, H. M., Flint, A. L., Flint, L. E., Thorne, J. H. & Franklin, J., 2024. Uncertainty in consensus predictions of plant species' vulnerability to climate change. – Divers. Distrib. 30: e13898.
Sarker, I. H., 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions. – SN Comput. Sci. 2: 160.
Sawall, Y. & Hochberg, E. J., 2018. Diel versus time-integrated (daily) photosynthesis and irradiance relationships of coral reef organisms and communities. – PLoS ONE 13: e0208607.
Shannon, C. E., 1948. A Mathematical Theory of Communication. – Bell Syst. Tech. J. 27: 379–423.
Soley-Guardia, M., Alvarado-Serrano, D. F. & Anderson, R. P., 2024. Top ten hazards to avoid when modeling species distributions: a didactic guide of assumptions, problems, and recommendations. – Ecography 2024: e06852.
Sunny, A., Marmolejo, C., Vidal-López, R., Falconi-Briones, F. A., Cuervo-Robayo, Á. P. & Bolom-Huet, R., 2025. EcoNicheS: enhancing ecological niche modeling, niche overlap and connectivity analysis using the shiny dashboard and R package. – PeerJ 13: e19136.
Tessarolo, G., Ladle, R. J., Lobo, J. M., Rangel, T. F. & Hortal, J., 2021. Using maps of biogeographical ignorance to reveal the uncertainty in distributional data hidden in species distribution models. – Ecography 44: 1743–1755.
Thuiller, W., 2024. Ecological niche modelling. – Curr. Biol. 34: R225–R229.
Thuiller, W., Lafourcade, B., Engler, R. & Araújo, M. B., 2009. Biomod – a platform for ensemble forecasting of species distributions. – Ecography 32: 369–373.
Tourinho, L. & Vale, M. M., 2023. Choosing among correlative, mechanistic, and hybrid models of species' niche and distribution. – Integr. Zool. 18: 93–109.
Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F. & Clerck, O. D., 2012. Bio-ORACLE: A global environmental data set for marine species distribution modelling. – Global Ecol. Biogeogr. 21: 272–281.

Velazco, S. J. E., Rose, M. B., De Andrade, A. F. A., Minoli, I. & Franklin, J., 2022. flexsdm: An R package for supporting a comprehensive and flexible species distribution modelling workflow. – Methods Ecol. Evol. 13: 1661–1669.
Velazco, S. J. E., Rose, M. B., De Marco Jr, P., Regan, H. M. & Franklin, J., 2024. How far can i extrapolate my species distribution model? exploring shape, a novel method. – Ecography 2024: e06992.
Verrill, A. E., 1868. Notice of the Corals and Echinoderms Collected by Prof. C. F. Hartt, at the Abrolhos Reefs, Province of Bahia, Brazil. – Trans. Conn. Acad. Arts Sci. 1: 351–371.
Wang, Z. & Zhan, W., 2012. Dynamic Engineering Multi-criteria Decision Making Model Optimized by Entropy Weight for Evaluating Bid. – Syst. Eng. Procedia 5: 49–54.
Warren, D. L. & Seifert, S. N., 2011. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. – Ecol. Appl. 21: 335–342.
Wilcox, A. 1967. Indices of qualitative variation. – U.S. Department of Energy, ORNL-TM–1919, 4167340.
Zeng, X., Wong, D. F. & Chao, L. S., 2014. Constructing Better Classifier Ensemble Based on Weighted Accuracy and Diversity Measure. – Sci. World J. 2014: 961747.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

Supporting Information for

Ensemble learning in ecological niche modeling of the *Siderastrea* (Cnidaria:Scleractinia) Atlantic complex: an information theory approach

Henrique J. Schipanski, Lucas B. Perroni, Caio F. Sguario and Marcos S. Barbeitos

Marcos S. Barbeitos PhD, Departamento de Zoologia, Universidade Federal do Paraná, Curitiba, PR, 81531980, Brazil.
email:msbarbeitos@gmail.com

This PDF file includes:

Supporting Information Text

Inter-reef distance and reef size estimation

In their paper on spatial properties of coral reefs, (1) surveyed over 1.5 million reefs (with areas over 1,000 m^2) cataloged in the Allen Coral Reef Database. The author derived several power-laws that describe distribution probabilities and scaling properties of coral reefs world wide. We derived part of our sampling strategy using their, results, as described below.

Inter-reef distance: probability density. Let D denote the inter-reef distance (in m). We assume that the distribution of D for $D \geq c_D$ follows a continuous power law with PDF exponent α_D:

$$p_D(d) = C_D\, d^{-\alpha_D}, \qquad d \geq c_D,$$

and $p_D(d) = 0$ for $d < c_D$.

We are given:

$$c_D = 100 \text{ m}, \qquad \alpha_D = 2.33.$$

where c_D is the minimum value for the upper tail of the distribution and α_D is the exponent of then power-law fit to the distance between nearest neighbors (Fig 1d. in their publication).

To normalize $p_D(d)$, we require

$$\int_{c_D}^{\infty} p_D(d)\, dd = 1.$$

Thus

$$1 = \int_{c_D}^{\infty} C_D\, d^{-\alpha_D}\, dd = C_D \int_{c_D}^{\infty} d^{-\alpha_D}\, dd = C_D \left[\frac{d^{1-\alpha_D}}{1-\alpha_D}\right]_{c_D}^{\infty} = C_D \frac{c_D^{1-\alpha_D}}{\alpha_D - 1},$$

and therefore

$$C_D = (\alpha_D - 1)\, c_D^{\alpha_D - 1}.$$

With $\alpha_D = 2.33$, we obtain

$$C_D = (2.33 - 1)\, c_D^{1.33} = 1.33\, c_D^{1.33}.$$

The normalized PDF for inter-reef distance is

$$p_D(d) = 1.33\, c_D^{1.33}\, d^{-2.33}, \qquad d \geq c_D.$$

This function $p_D(d)$ gives the expected probability density associated with distance d in the power-law tail of the inter-reef distance distribution. $p_D(10,000m) \approx 2.91 \cdot 10^{-7}$, hence lower than one in one million.

Reef size estimation: geometric considerations. Giménez-Romero and cols. report mean and median reef areas of 3.32 and 0.30 ha, respectively. Two points will only be sampled from the same reef when $D_{max} \geq 10\ km$ Reefs are fractal objects with a box counting dimension 95% confidence interval ranging from 1.13 to 1.35 according to their survey. The author report that the area x perimeter relationship of coral reefs may also be expressed by a power-law whose exponent is 1.26. We used their findings to estimate the probability that reefs with dimensions over 10 km exist in our sampling area.

Ellipse geometry and power-law area–perimeter relationship. We assume that coral reef patches can be approximated by ellipses with largest (major) diameter $D_{\max}$ and smallest (minor) diameter $D_{\min}$. The *elongation index* is defined as

$$e = \frac{D_{\min}}{D_{\max}}, \qquad 0 < e \leq 1.$$

The semi-major and semi-minor axes are

$$a = \frac{D_{\max}}{2}, \qquad b = \frac{D_{\min}}{2} = \frac{eD_{\max}}{2}.$$

The planar area of an ellipse with semi-axes a and b is

$$A_{\text{geom}} = \pi ab.$$

In terms of $D_{\max}$ and e,

$$A_{\text{geom}} = \pi \left(\frac{D_{\max}}{2}\right)\left(\frac{eD_{\max}}{2}\right) = \frac{\pi e}{4} D_{\max}^2.$$

There is no exact closed-form expression for the perimeter of an ellipse in elementary functions. A widely used approximation is Ramanujan's second formula. Let

$$h = \left(\frac{a-b}{a+b}\right)^2.$$

Then the approximate perimeter is

$$P_{\text{geom}} \approx \pi(a+b)\left(1 + \frac{3h}{10 + \sqrt{4-3h}}\right).$$

In our context, we compute a, b, and h from $D_{\max}$ and $e = D_{\min}/D_{\max}$ as above, and then evaluate P_{geom}.

Power-law area–perimeter relationship. Empirically, the relationship between reef area A and perimeter P follows a power law (Fig. 1c in (1)):

$$A = c\,P^{\alpha},$$

with exponent $\alpha = 1.26$. The parameter c is a scale factor (intercept in log–log space) that depends on the dataset and the units used.

Given a calibration point $(P_{\min}, A_{\min})$, we can estimate c by

$$A_{\min} = c\,P_{\min}^{\alpha} \quad \Rightarrow \quad c = \frac{A_{\min}}{P_{\min}^{\alpha}}.$$

From (1), we have that: $A_{\min} = 1000$ m^2 and $P_{\min} = 100$ m, with $\alpha = 1.26$, then

$$c = \frac{1000}{100^{1.26}}.$$

Given a geometric perimeter P_{geom} computed from the ellipse approximation, the corresponding power-law predicted area is

$$A_{\text{reef}} = c\,P_{\text{geom}}^{\alpha},$$

where c is estimated from the calibration point and $\alpha = 1.26$.

Thus, the workflow for a given maximum diameter $D_{\max}$ and elongation index e is:

1. Compute $a = D_{\max}/2$ and $b = eD_{\max}/2$.
2. Compute the geometric area $A_{\text{geom}} = \pi ab$.
3. Compute $h = \left(\frac{a-b}{a+b}\right)^2$.
4. Compute the geometric perimeter
$$P_{\text{geom}} \approx \pi(a+b)\left(1 + \frac{3h}{10+\sqrt{4-3h}}\right).$$
5. Estimate $c = A_{\min}/P_{\min}^{\alpha}$ from a minimal calibration reef.
6. Compute the power-law reef area $A_{\text{reef}} = c\,P_{\text{geom}}^{\alpha}$.

Power-law distributions for coral reef area: CCDF and CDF. We consider two continuous power-law distributions, one for reef planar area and one for inter-reef distance, each defined on a semi-infinite interval $[c, \infty)$ with lower cutoff c.

Let A denote reef planar area (in m^2). We assume that for $A \geq c_A$ the tail of the distribution follows a power law. The empirical analysis provides a power-law exponent for the *complementary cumulative distribution function* (CCDF),

$$\Pr(A > x) \propto x^{-\beta_A},$$

with exponent

$$\beta_A = 0.8,$$

and lower cutoff

$$c_A = 10{,}000 \text{ m}^2.$$

For a continuous power-law distribution on $[c_A, \infty)$ with PDF

$$p_A(a) = C_A\,a^{-\alpha_A}, \qquad a \geq c_A,$$

the normalization constant C_A and exponent α_A satisfy the condition

$$\int_{c_A}^{\infty} p_A(a)\,da = 1.$$

For $\alpha_A > 1$,

$$1 = \int_{c_A}^{\infty} C_A\,a^{-\alpha_A}\,da = C_A\int_{c_A}^{\infty} a^{-\alpha_A}\,da = C_A\left[\frac{a^{1-\alpha_A}}{1-\alpha_A}\right]_{c_A}^{\infty} = C_A\frac{c_A^{1-\alpha_A}}{\alpha_A - 1}.$$

Thus

$$C_A = (\alpha_A - 1)\,c_A^{\alpha_A - 1}.$$

The CCDF for $x \geq c_A$ is

$$\Pr(A > x) = \int_x^{\infty} p_A(a)\,da = C_A\int_x^{\infty} a^{-\alpha_A}\,da = C_A\frac{x^{1-\alpha_A}}{\alpha_A - 1}.$$

Substituting C_A gives

$$\Pr(A > x) = (\alpha_A - 1)\, c_A^{\alpha_A - 1} \frac{x^{1-\alpha_A}}{\alpha_A - 1} = \left(\frac{x}{c_A}\right)^{-(\alpha_A - 1)}.$$

Therefore, the CCDF exponent is $\beta_A = \alpha_A - 1$. In our case, the reported CCDF exponent is $\beta_A = 0.8$, so

$$\alpha_A = \beta_A + 1 = 1.8. \quad [1]$$

The normalized PDF is then

$$p_A(a) = 0.8\, c_A^{0.8}\, a^{-1.8}, \qquad a \geq c_A.$$

The CCDF and CDF are

$$\Pr(A > x) = \left(\frac{x}{c_A}\right)^{-\beta_A} = \left(\frac{x}{c_A}\right)^{-0.8}, \qquad x \geq c_A, \quad [2]$$

$$\Pr(A \leq x) = 1 - \Pr(A > x) = 1 - \left(\frac{x}{c_A}\right)^{-0.8}, \qquad x \geq c_A. \quad [3]$$

The study reports a median elongation index of approximately 0.5 (Fig. 2d (1)). Assuming an elliptical geometry compliant with this value, and applying the workflow outlined above, the expected area of a 10 km long reef is $x \approx 100$ ha, roughly 300 times the coral reefs median area (1). Table S4 (see Giménez-Romero and cols. Supplementary Information) lists PDF exponents (α_A) for different coral reef provinces. The region with the largest coral reefs in the Atlantic is the Southeastern Caribbean ($\alpha_A = 1.81$). We used 1 to calculate β_A, and verified that fewer than 2.4% of the reefs in this province are expected to have areas that exceed 100 ha according to equation 2. A more conservative estimate assumes an elongation index of 0.1, which is unlikely according to their analyses (see Fig 2d. in their paper). The expected area of such a narrow reef is $x \approx 81$ ha, and more than 97% of the reefs in this province should have areas under this threshold according to equation 3.

Parameter selection and model tuning

"Fish Bowl" models. "Fish Bowl" are simple envelope and distance-based ENMs (e.g., BIOCLIM, Gower or Mahalanobis distance, ENFA) that assume species distributions are constrained by environmental tolerances. These models have low complexity, assumptions deeply rooted in ecological and physiological theory, and their outcomes offer intuitive interpretation. Therefore, they are transparent and easy to understand. Because they tend to define broad suitability envelopes, "Fish Bowl" models commonly overpredict distributions, inflating commission errors (predicting presence where the species is actually absent) while usually keeping omission errors moderate. Overfitting is unlikely given their low complexity, but this comes at the cost of coarser, often overly generous predicted ranges. These models are "presence-only" and have no more than one tunable parameter, if any. (2). BioClim was the only model in this category that required tuning.

BioClim. BioClimatic envelope modeling was conducted using the function `bm_SRE` from the R package `biomod2` (3). The function has one tuning parameter which is the quantile of extreme observations to be excluded before training and testing. We tested four values: $(0.005, 0.010, 0.025, 0.050)$. Best quantiles were 0.010 for both *S. radians* and *S. siderea*, and 0.05 for *S. stellata*.

"Turbine" models. These models' biological significance is harder to assess because their assumptions are only partially based on ecological theory. They treat occurrence as a (possibly nonlinear) function of environmental predictors and fit explicit regression parameters. They are more complex than Fish Bowl models, can capture interactions and smooth responses, and typically have higher predictive power and precision. Overfitting risk is moderate: these models can overfit if over-parameterized or poorly regularized, but they usually offer a reasonable balance between flexibility and control, so both omission and commission errors are often intermediate between Fish Bowl and Vault models. As a result, Turbine models are often recommended for conservation planning and climate-change projections because they balance interpretability, predictive performance, and robustness to variation in IC. Their parameters may be manipulated to improve predictive power, as a car mechanic would tune up an engine to improve its performance, hence the analogy with a "turbine" (2).

GAM. GAM was fitted using Restricted Maximum Likelihood (REML) via `gam`, a function of the R package `mgcv` (4). We used a single parameter combination as the starting point of optimization, employing the standard mathematical structure:

$$\text{logit}(E(p)) = \beta_0 + s_1(\text{BTM}) + s_2(\text{Chl}_a) + s_3(\text{Dat}) + s_4(\text{PSS}) + s_5(\text{SST})$$

where s_p are smoothing functions and p is assumed to be drawn from a binomial distribution. Parameter selection was accomplished via shrinkage penalization of the smoothed terms, whose intensity is controlled by the *gamma* parameter. We tested $gamma \in \{1.4, 1.6, 1.8, 2.0\}$. Smoothed terms with effective degrees of freedom > 0.1 were retained (4). After the **parameter selection** stage (conducted on replicates 1-10), the smoothing parameter for diffuse attenuation *Dat* was dropped for all species, but the original formula was also retained for *S. radians* and *S. siderea*. After the **cross-validation** stage (conducted of replicates 11-20), only the simpler formula (i.e. lacking the non-linear component of *Dat*) was retained for all datasets:

$$\text{logit}(E(p)) = \beta_0 + \beta_1\text{BTM} + \beta_2\text{Chl}_a + \beta_3\text{Dat} + \beta_4\text{PSS} + \beta_5\text{SST} + s_1(\text{BTM}) + s_2(\text{Chl}_a) + s_4(\text{PSS}) + s_5(\text{SST})$$

GLM. GLM was fitted using R core function `glm`. The starting linear combination included all possible pairwise interactions between independent variables plus their quadratic terms. Formula selection was conducted using the `backward` searching algorithm implemented in the `stepAIC` function of the package `MASS` (5), setting `BIC` (Bayesian Information) as the selection criterion. The following models were retained after the **parameter selection** stage (replicates 1-10):

S. radians

1 $$\text{logit}(p) = \beta_0 + \beta_1\text{PSS}^2 + \beta_2\text{BTM} + \beta_3\text{Chl}_a + \beta_4\text{PSS} + \beta_5\text{SST} + \epsilon$$

2 $$\text{logit}(p) = \beta_0 + \beta_1\text{PSS}^2 + \beta_2(\text{BTM} \times \text{PSS}) + \beta_3(\text{Chl}_a \times \text{PSS}) + \beta_4\text{BTM} + \beta_5\text{Chl}_a + \beta_6\text{PSS} + \beta_7\text{SST} + \epsilon$$

3 $$\text{logit}(p) = \beta_0 + \beta_1\text{PSS}^2 + \beta_2(\text{BTM} \times \text{PSS}) + \beta_3(\text{Dat} \times \text{PSS}) + \beta_4\text{BTM} + \beta_5\text{Chl}_a + \beta_6\text{Dat} + \beta_7\text{PSS} + \beta_8\text{SST} + \epsilon$$

4 $$\text{logit}(p) = \beta_0 + \beta_1(\text{BTM} \times \text{PSS}) + \beta_2(\text{Chl}_a \times \text{PSS}) + \beta_3(\text{Dat} \times \text{PSS}) + \beta_4\text{BTM} + \beta_5\text{Chl}_a + \beta_6\text{Dat} + \beta_7\text{PSS} + \beta_8\text{SST} + \epsilon$$

5 $$\text{logit}(p) = \beta_0 + \beta_1\text{PSS}^2 + \beta_2(\text{BTM} \times \text{PSS}) + \beta_3(\text{Chl}_a \times \text{PSS}) + \beta_4(\text{Dat} \times \text{PSS}) + \beta_5\text{BTM} + \beta_6\text{Chl}_a + \beta_7\text{Dat} + \beta_8\text{PSS} + \beta_9\text{SST} + \epsilon$$

6 $$\text{logit}(p) = \beta_0 + \beta_1\text{PSS}^2 + \beta_2(\text{Chl}_a \times \text{PSS}) + \beta_3\text{BTM} + \beta_4\text{Chl}_a + \beta_5\text{PSS} + \beta_6\text{SST} + \epsilon$$

7 $$\text{logit}(p) = \beta_0 + \beta_1\text{PSS}^2 + \beta_2(\text{BTM} \times \text{PSS}) + \beta_3(\text{Chl}_a \times \text{PSS}) + \beta_4(\text{PSS} \times \text{SST}) + \beta_5\text{BTM} + \beta_6\text{Chl}_a + \beta_7\text{PSS} + \beta_8\text{SST} + \epsilon$$

S. siderea

1 $$\text{logit}(p) = \beta_0 + \beta_1(\text{BTM} \times \text{Chl}_a) + \beta_2(\text{BTM} \times \text{Dat}) + \beta_3(\text{BTM} \times \text{SST}) + \beta_4(\text{Chl}_a \times \text{PSS}) + \beta_5(\text{PSS} \times \text{SST}) + \beta_6\text{BTM} + \beta_7\text{Chl}_a + \beta_8\text{Dat} + \beta_9\text{PSS} + \beta_{10}\text{SST} + \epsilon$$

2 $$\text{logit}(p) = \beta_0 + \beta_1(\text{BTM} \times \text{Chl}_a) + \beta_2(\text{BTM} \times \text{Dat}) + \beta_3(\text{Chl}_a \times \text{PSS}) + \beta_4(\text{PSS} \times \text{SST}) + \beta_5\text{BTM} + \beta_6\text{Chl}_a + \beta_7\text{Dat} + \beta_8\text{PSS} + \beta_9\text{SST} + \epsilon$$

3 $$\text{logit}(p) = \beta_0 + \beta_1(\text{BTM} \times \text{Chl}_a) + \beta_2(\text{BTM} \times \text{SST}) + \beta_3(\text{Chl}_a \times \text{PSS}) + \beta_4(\text{Dat} \times \text{PSS}) + \beta_5(\text{PSS} \times \text{SST}) + \beta_6\text{BTM} + \beta_7\text{Chl}_a + \beta_8\text{Dat} + \beta_9\text{PSS} + \beta_{10}\text{SST} + \epsilon$$

S. stellata

1 $$\text{logit}(p) = \beta_0 + \beta_1(\text{Chl}_a \times \text{SST}) + \beta_2(\text{PSS} \times \text{SST}) + \beta_3\text{BTM} + \beta_4\text{Chl}_a + \beta_5\text{Dat} + \beta_6\text{PSS} + \beta_7\text{SST} + \epsilon$$

2 $$\text{logit}(p) = \beta_0 + \beta_1\text{SST}^2 + \beta_2(\text{BTM} \times \text{SST}) + \beta_3(\text{Chl}_a \times \text{Dat}) + \beta_4(\text{Chl}_a \times \text{SST}) + \beta_5(\text{Dat} \times \text{PSS}) + \beta_6(\text{Dat} \times \text{SST}) + \beta_7(\text{PSS} \times \text{SST}) + \beta_8\text{BTM} + \beta_9\text{Chl}_a + \beta_{10}\text{Dat} + \beta_{11}\text{PSS} + \beta_{12}\text{SST} + \epsilon$$

3 $$\text{logit}(p) = \beta_0 + \beta_1(\text{Dat} \times \text{PSS}) + \beta_2(\text{Dat} \times \text{SST}) + \beta_3(\text{PSS} \times \text{SST}) + \beta_4\text{BTM} + \beta_5\text{Dat} + \beta_6\text{PSS} + \beta_7\text{SST} + \epsilon$$

After the **cross-validations** stage (replicates 11-20), the final models were:

S. radians

5 $$\text{logit}(p) = \beta_0 + \beta_1\text{PSS}^2 + \beta_2(\text{BTM} \times \text{PSS}) + \beta_3(\text{Chl}_a \times \text{PSS}) + \beta_4(\text{Dat} \times \text{PSS}) + \beta_5\text{BTM} + \beta_6\text{Chl}_a + \beta_7\text{Dat} + \beta_8\text{PSS} + \beta_9\text{SST} + \epsilon$$

S. siderea

3 $$\text{logit}(p) = \beta_0 + \beta_1(\text{BTM} \times \text{Chl}_a) + \beta_2(\text{BTM} \times \text{SST}) + \beta_3(\text{Chl}_a \times \text{PSS}) + \beta_4(\text{Dat} \times \text{PSS}) + \beta_5(\text{PSS} \times \text{SST}) + \beta_6\text{BTM} + \beta_7\text{Chl}_a + \beta_8\text{Dat} + \beta_9\text{PSS} + \beta_{10}\text{SST} + \epsilon$$

S. stellata

3 $$\text{logit}(p) = \beta_0 + \beta_1(\text{Dat} \times \text{PSS}) + \beta_2(\text{Dat} \times \text{SST}) + \beta_3(\text{PSS} \times \text{SST}) + \beta_4\text{BTM} + \beta_5\text{Dat} + \beta_6\text{PSS} + \beta_7\text{SST} + \epsilon$$

MARS. Logistic MARS models were fitted using the function `earth` of the R package with the same name(6). Model selection was performed using `earth`'s built-in features, allowing for two-way interactions among variables (`degree` = 2). By default, `earth` employs cross-validation and forward/backward searching. Best models were subsequently subjected to analysis of variable importance using the function `evimp` in order to prune "unused" terms. Selected models were (B stands for the `basis function` that fits the splines): *S. radians*

1 $$\text{logit}(p) = \beta_0 + \beta_1 B_1\text{BTM} + \beta_2 B_2\text{Chl}_a + \beta_3 B_3\text{PSS} + \beta_4 B_4\text{SST} + \epsilon$$

2 $$\text{logit}(p) = \beta_0 + \beta_1 B_1\text{BTM} + \beta_2 B_2\text{Chl}_a + \beta_3 B_3\text{PSS} + \epsilon$$

S. siderea

1 $$\text{logit}(p) = \beta_0 + \beta_1 B_1 \text{BTM} + \beta_2 B_2 \text{Chl}_a + \beta_3 B_3 \text{PSS} + \beta_4 B_4 \text{SST} + \epsilon$$

S. stellata

1 $$\text{logit}(p) = \beta_0 + \beta_1 B_1 \text{BTM} + \beta_2 B_2 \text{Chl}_a + \beta_3 B_3 \text{Dat} + \beta_4 B_4 \text{PSS} + \beta_5 B_5 \text{SST} + \epsilon$$
2 $$\text{logit}(p) = \beta_0 + \beta_1 B_1 \text{BTM} + \beta_2 B_2 \text{Chl}_a + \beta_3 B_3 \text{PSS} + \beta_4 B_4 \text{SST} + \epsilon$$
3 $$\text{logit}(p) = \beta_0 + \beta_1 B_1 \text{BTM} + \beta_2 B_2 \text{PSS} + \beta_3 B_3 \text{SST} + \epsilon$$

After the **cross-validation** stage, retained models were: *S. radians and S. siderea*

1 $$\text{logit}(p) = \beta_0 + \beta_1 B_1 \text{BTM} + \beta_2 B_2 \text{Chl}_a + \beta_3 B_3 \text{PSS} + \beta_4 B_4 \text{SST} + \epsilon$$

S. stellata

1 $$\text{logit}(p) = \beta_0 + \beta_1 B_1 \text{BTM} + \beta_2 B_2 \text{Chl}_a + \beta_3 B_3 \text{Dat} + \beta_4 B_4 \text{PSS} + \beta_5 B_5 \text{SST} + \epsilon$$

"Vault" models. "Vault" is made up by "blackbox" models whose statistical machinery is hidden, as if locked in a safe. They are high-complexity machine-learning ENMs, including methods like Random Forests, MaxEnt, GARP, GBM, SVM and ANN, which use flexible algorithms to maximize fit between occurrences and environment with many parameters and often opaque internal structure. They usually achieve the highest statistical fit and precision among model classes, but their complexity makes parameter interpretation difficult and their realism and generality may be sacrificed for predictive performance. These models are particularly prone to overfitting, especially when extrapolated beyond the calibration domain, because there is often no strict constraint on maximum complexity or parameter count. Empirical work indicates that Vault models tend to predict smaller distributions than other model types (more restrictive suitability), which generally reduces commission errors but increases omission errors. (2).

ANN. We employed the function `h2o.deeplearning` (7) to estimate habitat suitability via artificial neural networks, setting the `distribution` parameter to "bernoulli". ANN was optimized through grid search of parameters `size` (the number of nodes in the hidden layer) and `l2` (weight-decay regularization parameter that shrinks weights towards 0, decreasing large weights and preventing overfitting). We explored number of nodes ranging from 2 to 8 $(2, 4, 6, 8)$ and regularization parameter ranging from 10^{-5} to 10^{-1} in 10^{-1} intervals. Selected parameters were $\{2, 10^{-1}\}$ for *S. radians*, $\{6, 10^{-1}\}$ for *S. siderea* and $\{2, 10^{-4}\}$ for *S. stellata*.

MaxEnt. MaxEnt modeling was performed using the function `maxnet` from the R package with the same name (8). Grid search was conducted over `regularization multiplier` values $(0.5, 1.0, 2.0, 3.0, 5.0)$ and `feature classes` $(l, lq, lqh, lqhp, lqht, lqhpt)$, where classes are l = linear, q = quadratic, h = hinge, p = product and t = threshold. The set $\{l, 0.5\}$ was selected for all species.

SVM. SVM was implemented with the aid of the function `svm` from the R package `e1071` (9). We employed a "radial" `kernel` and the `class_weights` parameter was set to "inverse" because of the imbalance between presences and pseudo-absences. Grid search was conducted over `cost` $(10^{-1}, 10^0, 10^1, 10^2, 10^3)$ and `gamma` values $(10^{-3}, 10^{-2}, 10^{-1}, 1, 2, 5)$. Selected sets were $\{10^3, 10^{-3}\}$ for *S. radians* and *S. siderea* and $\{10^3, 10^{-1}\}$ for *S. stellata*.

GLMM analysis of Entropy and TSS scores

GLMMs were fitted to the data via maximum-likelihood, using the R package *glmmTMB* (10). Replicates were treated as random factors nested within species, whereas species and models were considered fixed factors.

Entropy. Entropy is always positive and visual inspection of the data suggested a multiplicative relationship with the predictors (see Results in Main Text), so we employed either Gaussian or Gamma distributions, with variance modeled by a log link function in both cases. Visual inspection of bootstrapped instabilities (see Fig. 3 in Main Text) also suggested strong interaction between fixed effects, therefore, species and models were either treated independently, or an interaction term was fitted instead. We also fitted either the linear combination or product of fixed factors in the dispersion formula, yielding 8 different GLMMs, subsequently ranked using Bayesian Information Criterion (BIC). These GLMMs were then subjected to residual analysis using using 1,000 simulations in *DHARMa* (11) and results from the top chosen using this program was analyzed using *post-hoc* pairwise comparisons based on least-squares means using the R package *emmeans* (12, 13). The top GLMM design had interaction terms both for fixed effects and dispersion formulas. BIC preferred the gamma GLMM, which was also supported by *DHARMa* residual analyses (Fig. S14).

TSS. True Skill Statistic (TSS) scores computed for each CVE (see Material and Methods in Main Text) were used as an accuracy-based measure to evaluate performance. TSS scores theoretically range from -1 to 1, but were never lower than 0.7 (see Results in Main Text). It was thus unclear which probability distribution available in *glmmTMB* would be best suited to this variable. We fitted the same GLMM designs that were applied to instability scores, but assuming that TSS scores were either normally-, beta- or gamma-distributed. In order to improve fit, we raised TSS scores to a series of the most commonly used exponents in Box-Cox (14) transformations $(-2, -1, -1/2, 1/2, 1, 2)$, plus cubic terms $(-3, -1/3, 1/3, 3)$. Finally, we tested both logit and probit link functions when TSS scores were assumed to be sampled from a beta distribution (normal and gamma distributions were fit with identity and log link functions, respectively). A total of 100 models were tested and ranked using BIC and the top five GLMMs had their residuals analyzed using *DHARMa*. The best model had the same design selected for the entropy data, but assumed Gaussian-distributed scores on a cube-root scale. The second best TSS ($\Delta BIC = 22.3$) was a beta GLMM with a probit link function, also employing cube-root transformed TSS scores. This latter design passed all homoscedasticity tests and had fewer residual outliers according to *DHARMa* (Fig. S15), hence it was chosen over the top BIC pick for subsequent analyses.

References

1. A Gimenez-Romero, MA Matias, CM Duarte, Unravelling the universal spatial properties of coral reefs. *Glob. Ecol. Biogeogr.* **34**, e13939 (2025).
2. TF Rangel, RD Loyola, Labeling ecological niche models. *Natureza & Conserv.* **10**, 119–126 (2012).
3. W Thuiller, B Lafourcade, R Engler, MB Araújo, BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography* **32**, 369–373 (2009).
4. SN Wood, mgcv: GAMs and generalized ridge regression for r. *R news* **1/2** (2001).
5. WN Venables, BD Ripley, *Modern Applied Statistics with S.* (Springer, New York), Fourth edition, (2002) ISBN 0-387-95457-0.
6. S Milborrow, T Hastie, R Tibshirani, *earth: Multivariate Adaptive Regression Splines*, (2026) R package version 5.3.5.
7. T Fryda, et al., *h2o: R Interface for the 'H2O' Scalable Machine Learning Platform*, (2026) R package version 3.46.0.10.
8. S Phillips, *maxnet: Fitting 'Maxent' Species Distribution Models with 'glmnet'*, (2021) R package version 0.1.4.
9. D Meyer, E Dimitriadou, K Hornik, A Weingessel, F Leisch, *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, (2025) R package version 1.7-17.
10. ME Brooks, et al., glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R J.* **9**, 378–400 (2017).
11. F Hartig, *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*, (2022) R package version 0.4.6.
12. SR Searle, FM Speed, GA Milliken, Population marginal means in the linear model: An alternative to least squares means. *The Am. Stat.* **34**, 216–221 (1980).
13. RV Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, (2025) R package version 1.11.2.
14. GEP Box, DR Cox, An analysis of transformations. *J. Royal Stat. Soc. Ser. B (Methodological)* **26**, 211–243 (1964).

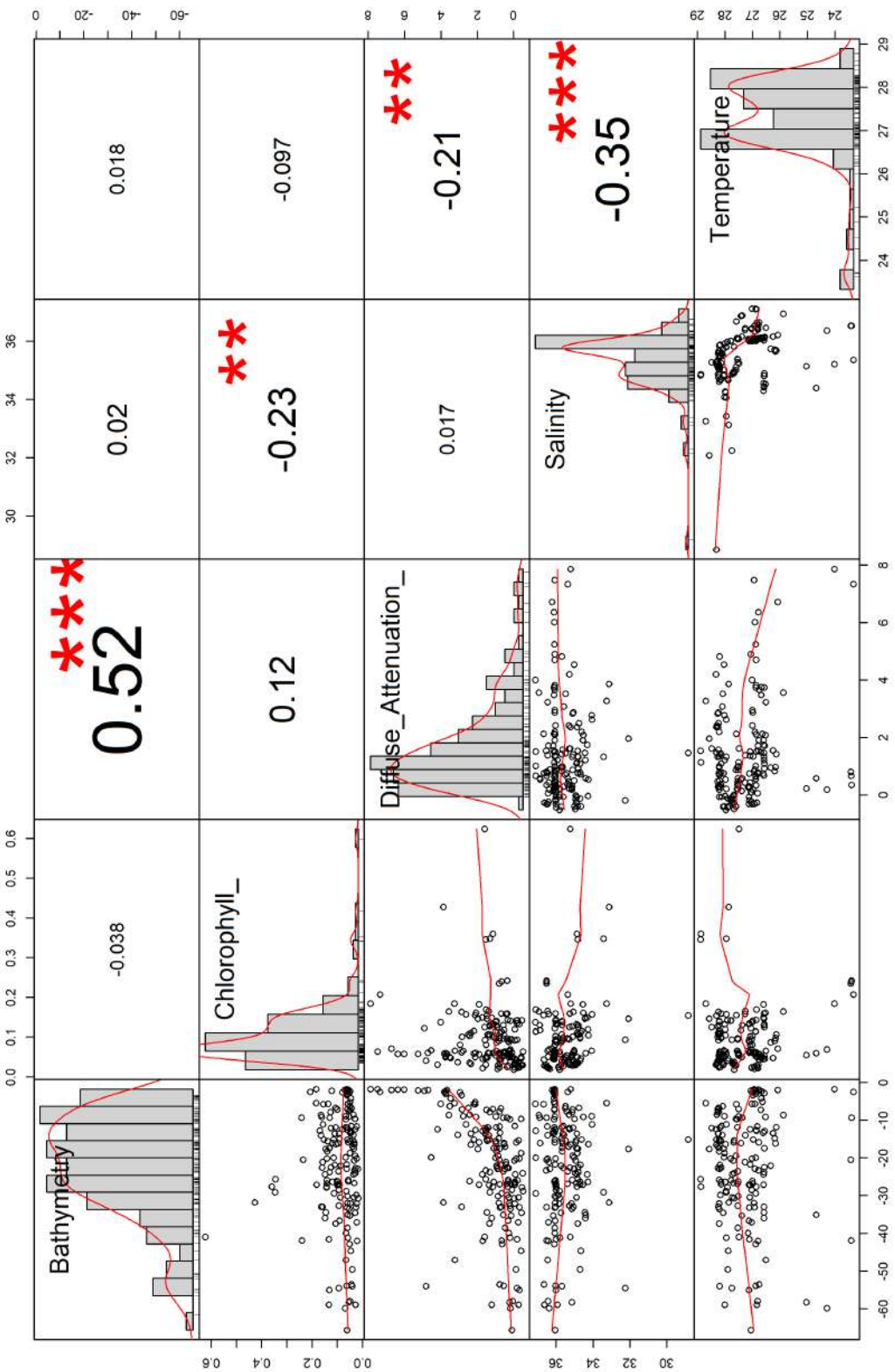

Fig. S1. Correlogram of environmental variables sampled at *S. radians* occurrence sites.

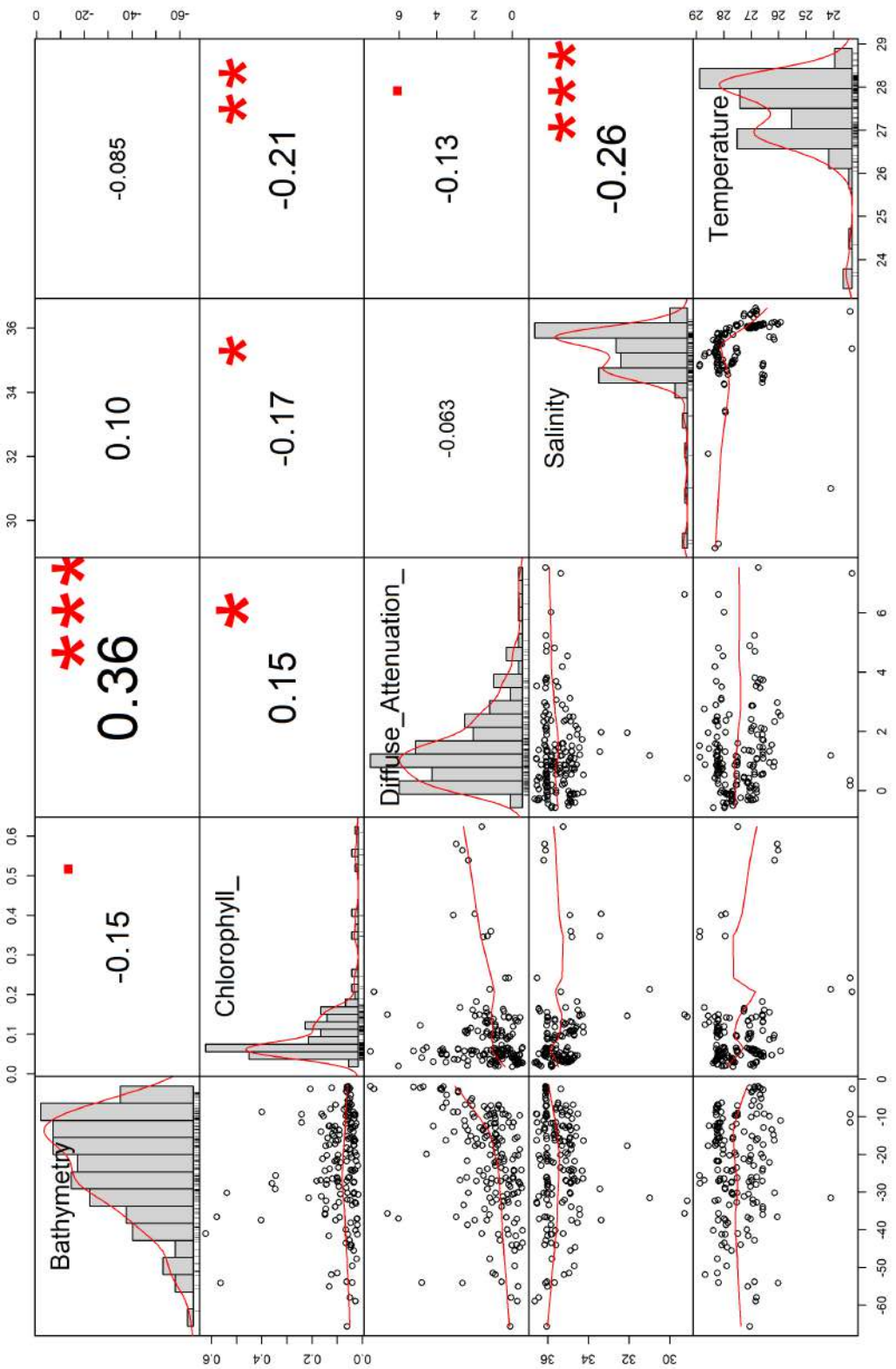

Fig. S2. Correlogram of environmental variables sampled at *S. siderea* occurrence sites.

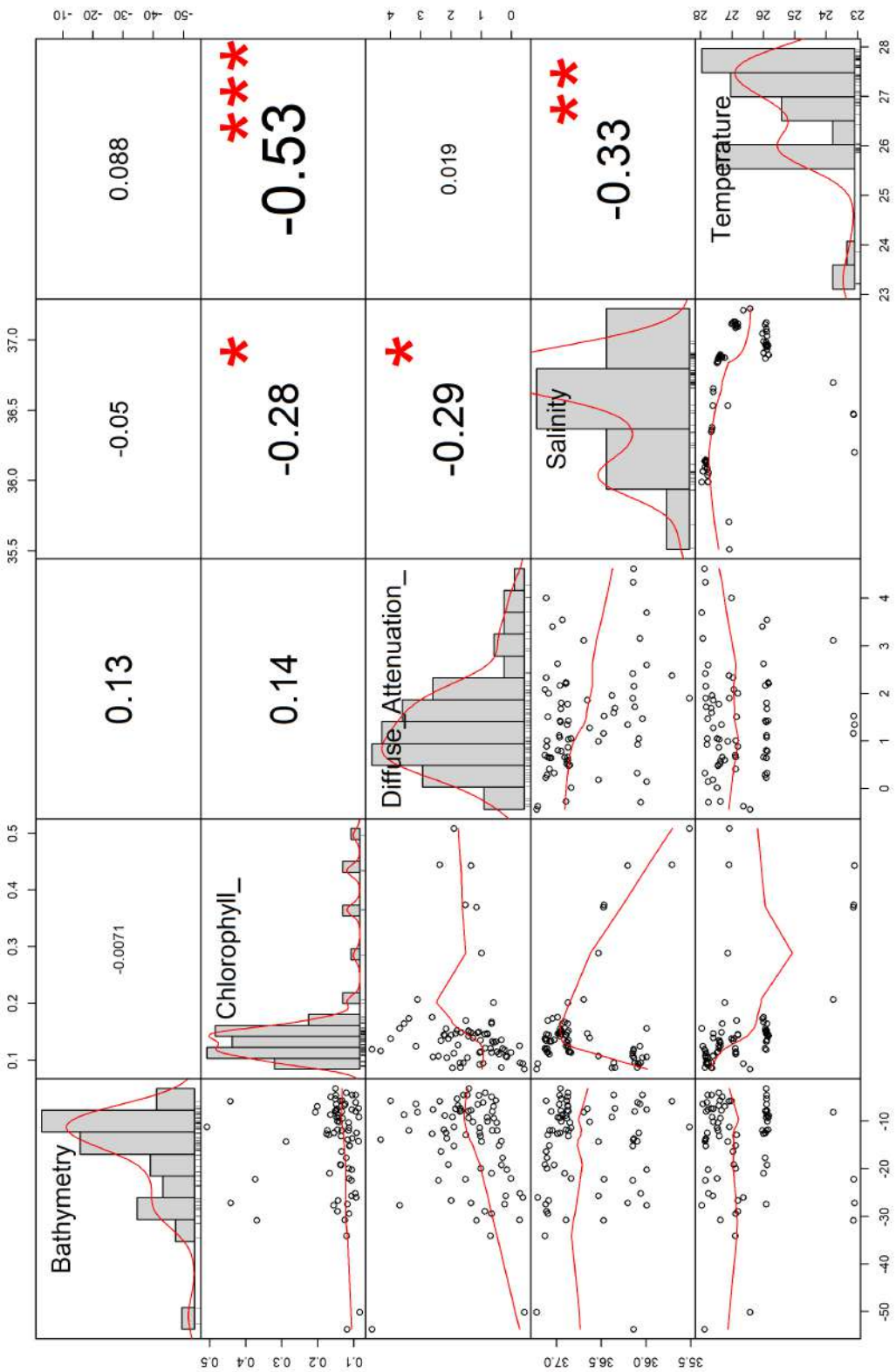

Fig. S3. Correlogram of environmental variables sampled at *S. stellata* occurrence sites.

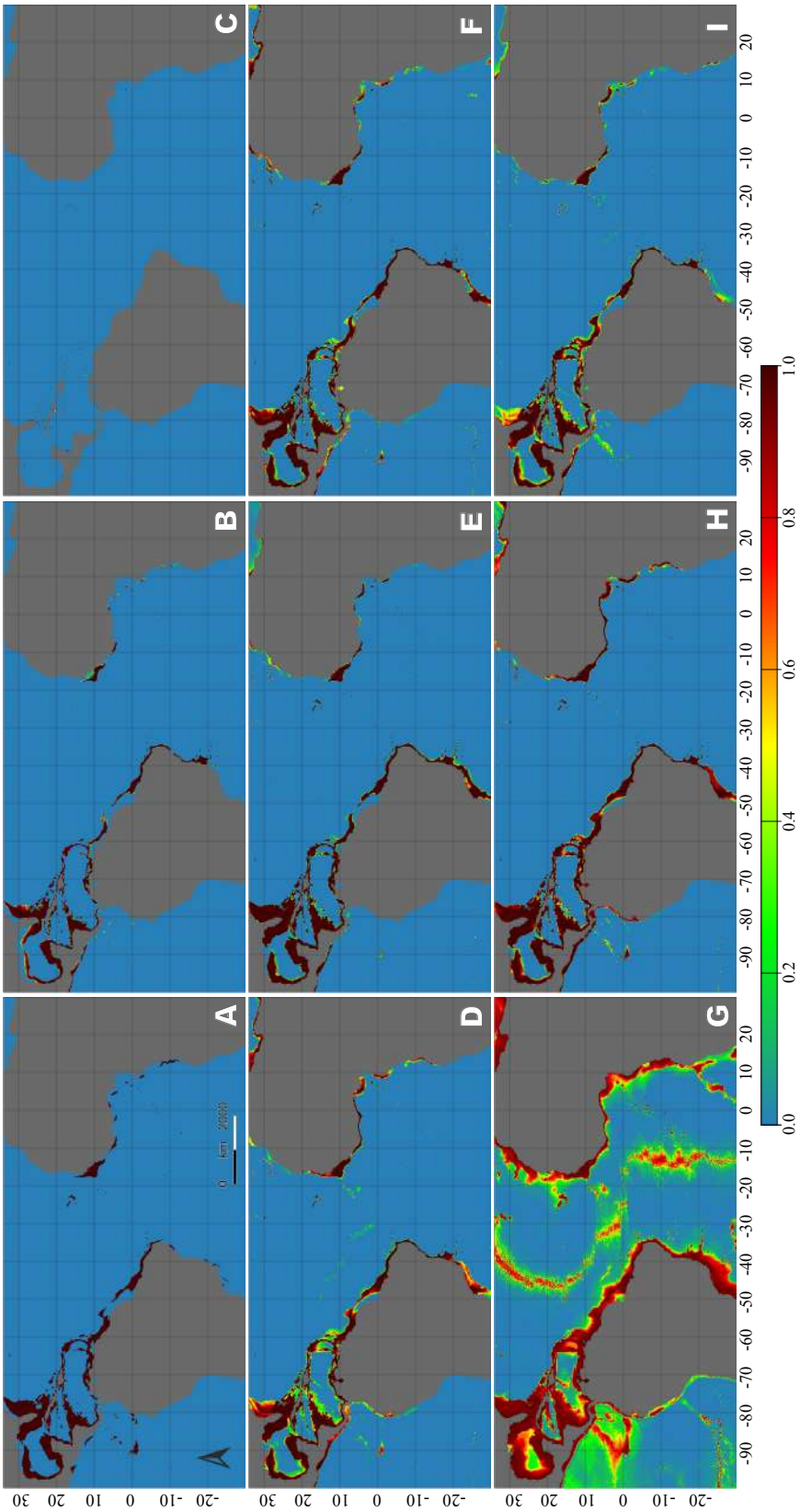

Fig. S4. Suitability maps for selected ENM methods applied to *S. radians* data. A-C: fishbowl methods - A. BioClim, B. Gower, C. Mahalanobis; D-F: turbine methods - D. GAM, E. GLM, F. MARS and G-I: blackbox - G. ANN, H. MaxEnt and I. SVM (Rangel & Loyola, 2012).

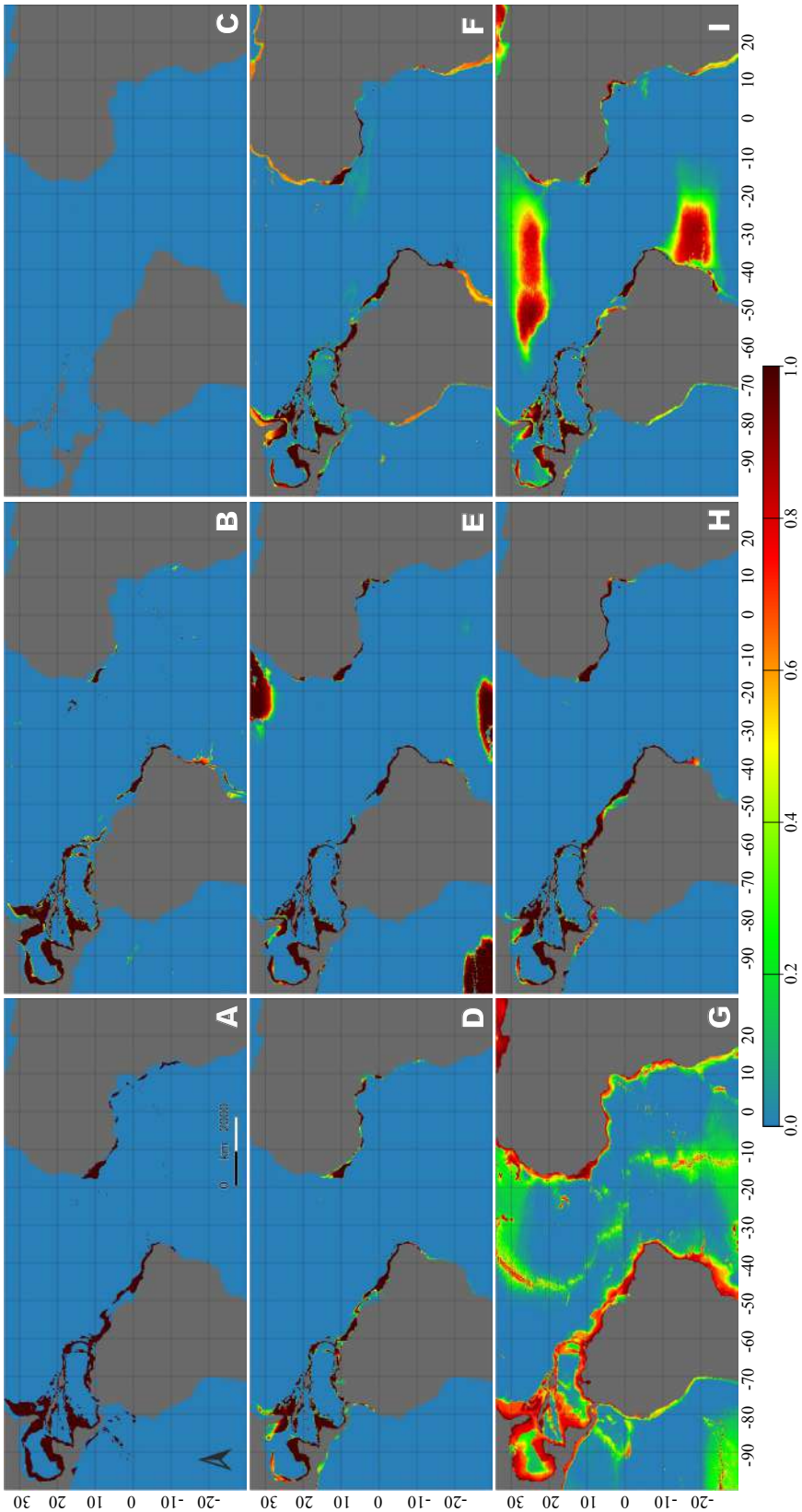

Fig. S5. Suitability maps for selected ENM methods applied to *S. siderea* data. A-C: fishbowl methods - A. BioClim, B. Gower, C. Mahalanobis; D-F: turbine methods - D. GAM, E. GLM, F. MARS and G-I: blackbox - G. ANN, H. MaxEnt and I. SVM (Rangel & Loyola, 2012).

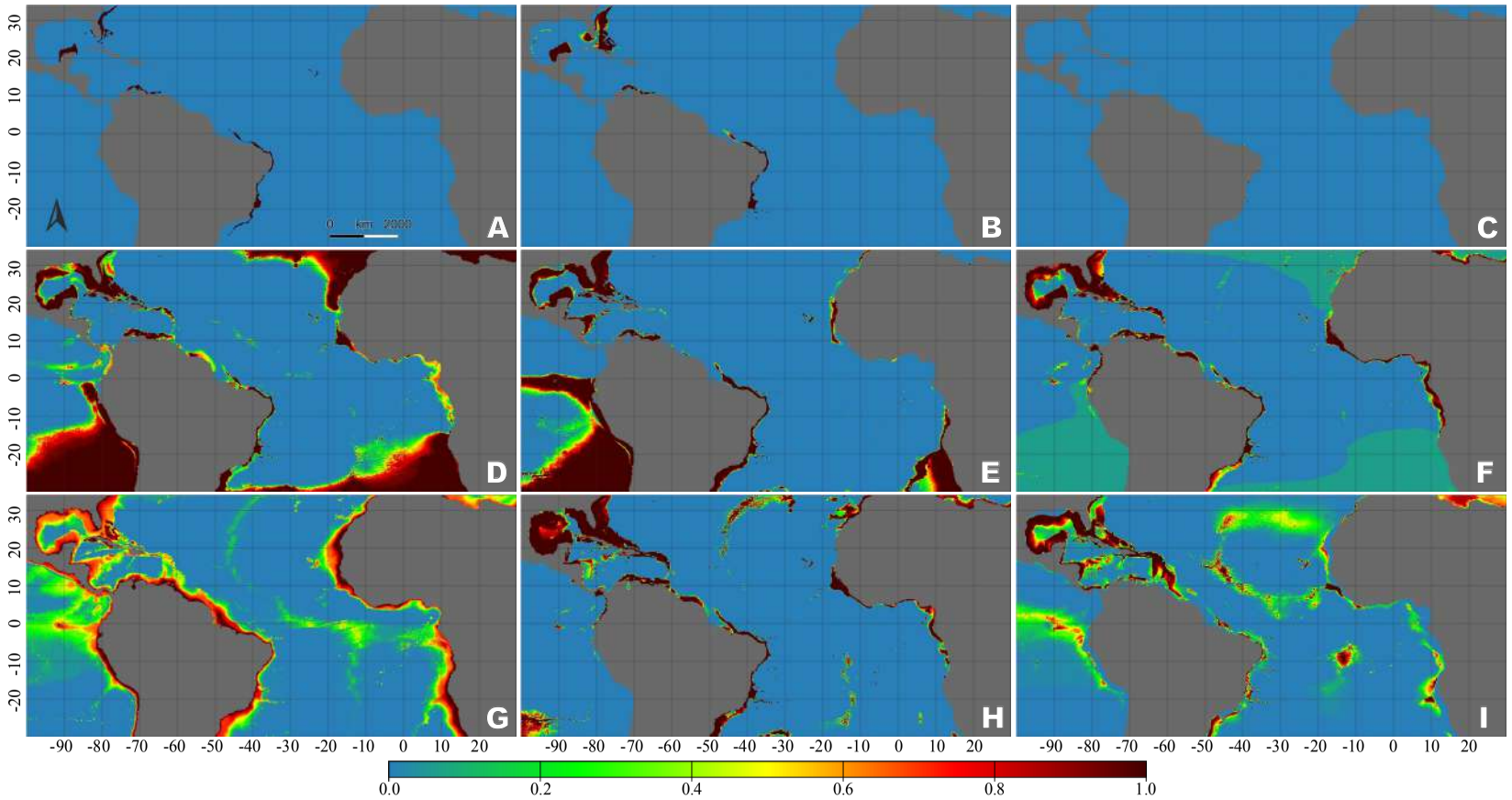

Fig. S6. Suitability maps for selected ENM methods applied to *S. stellata* data. A-C: fishbowl methods - A. BioClim, B. Gower, C. Mahalanobis; D-F: turbine methods - D. GAM, E. GLM, F. MARS and G-I: blackbox - G. ANN, H. MaxEnt and I. SVM (Rangel & Loyola, 2012).

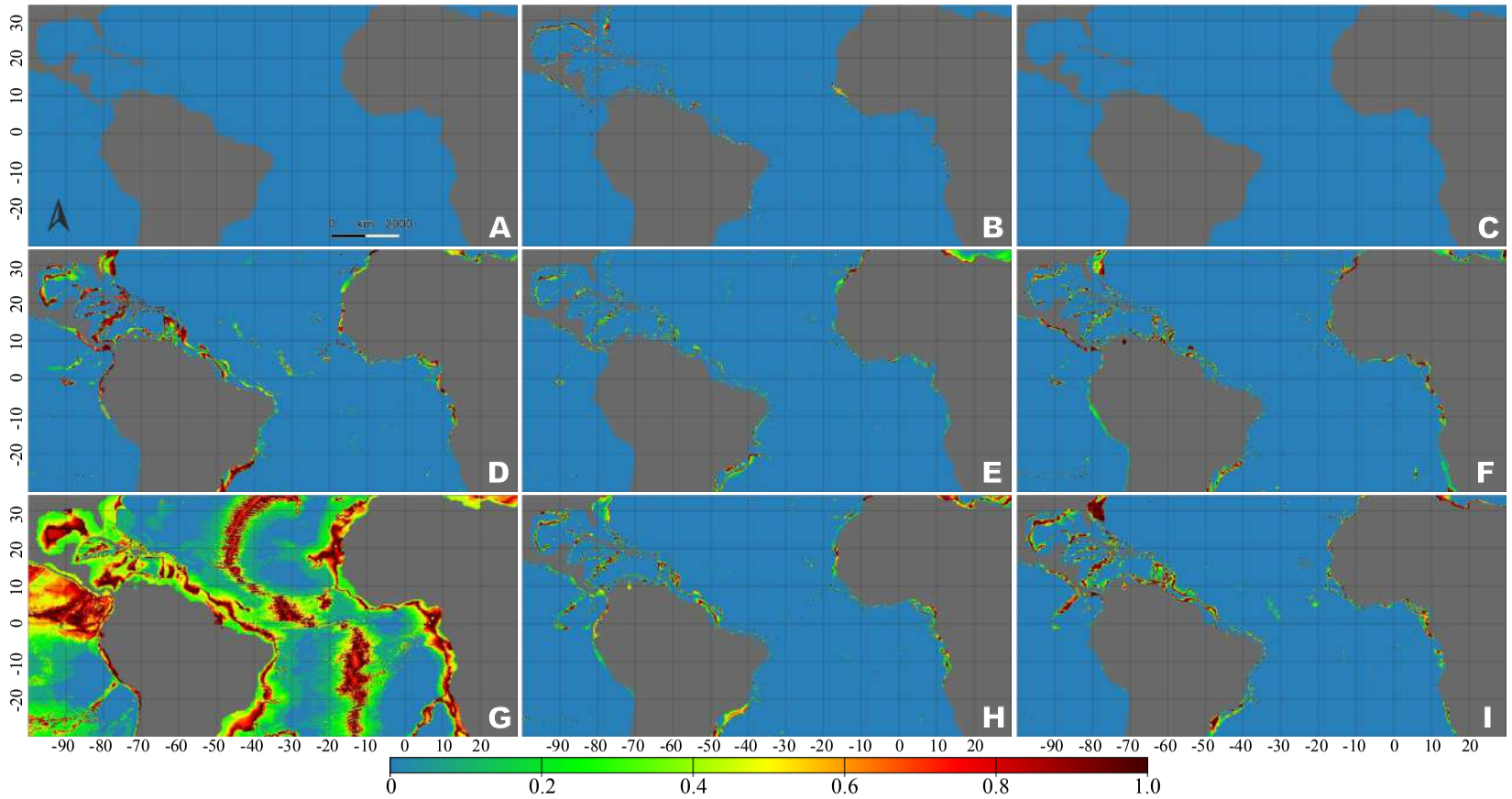

Fig. S7. Overfitting surfaces for selected ENM models, quantified using Shannon's entropy (H). *S. radians*. A-C: fishbowl methods - A. BioClim, B. Gower, C. Mahalanobis; D-F: turbine methods - D. GAM, E. GLM, F. MARS and G-I: blackbox - G. ANN, H. MaxEnt and I. SVM (Rangel & Loyola, 2012). For any given cell, when H = 1, the method predicted species occurrence in 50% of the CVEs and absence in the other 50%. Conversely, H = 0 means that the method consistently predicted either presence or absence of the species in all CVEs.

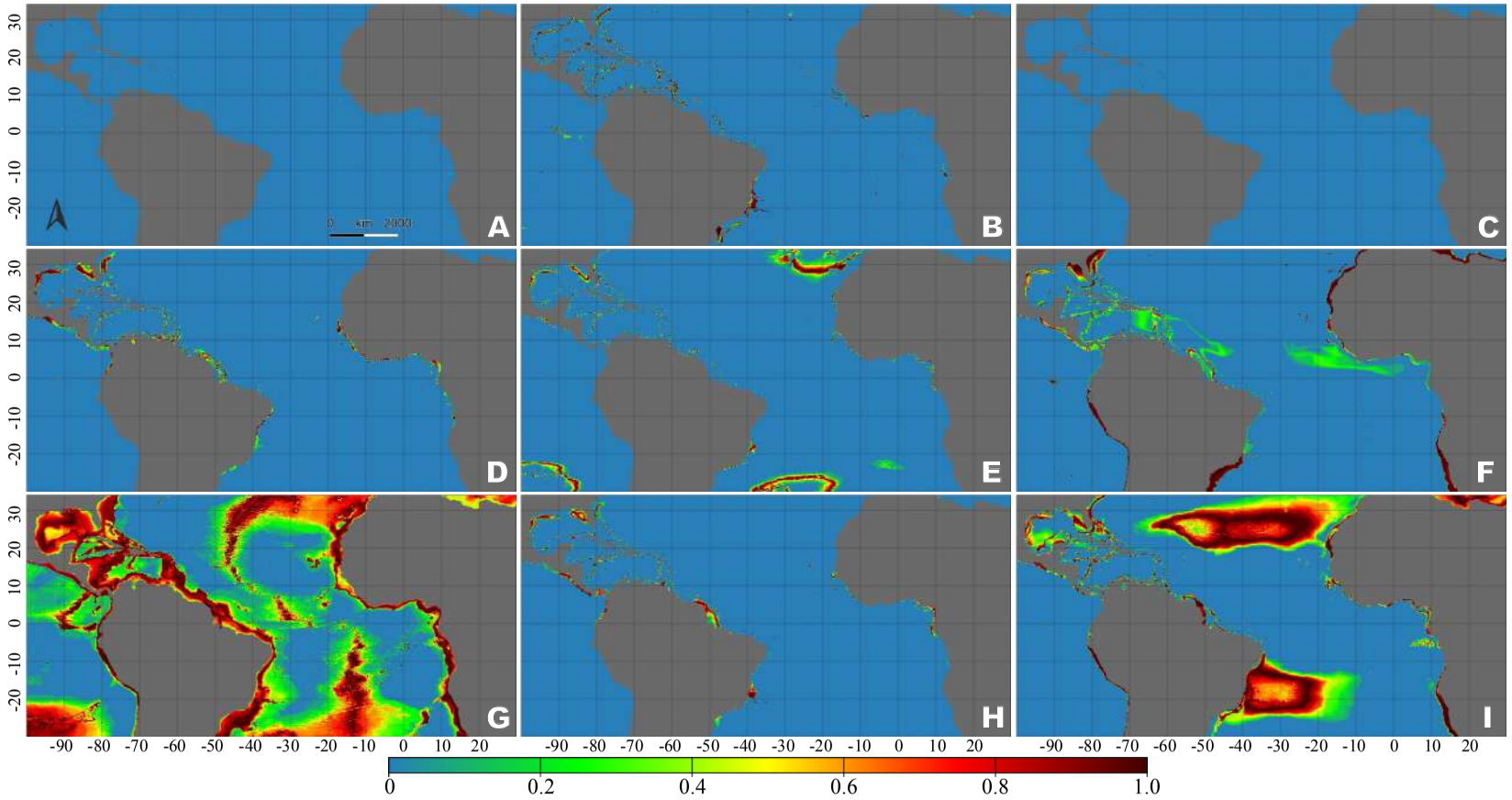

Fig. S8. Overfitting surfaces for selected ENM models, quantified using Shannon's entropy (H). *S. siderea*. A-C: fishbowl methods - A. BioClim, B. Gower, C. Mahalanobis; D-F: turbine methods - D. GAM, E. GLM, F. MARS and G-I: blackbox - G. ANN, H. MaxEnt and I. SVM (Rangel & Loyola, 2012). For any given cell, when H = 1, the method predicted species occurrence in 50% of the CVEs and absence in the other 50%. Conversely, H = 0 means that the method consistently predicted either presence or absence of the species in all CVEs.

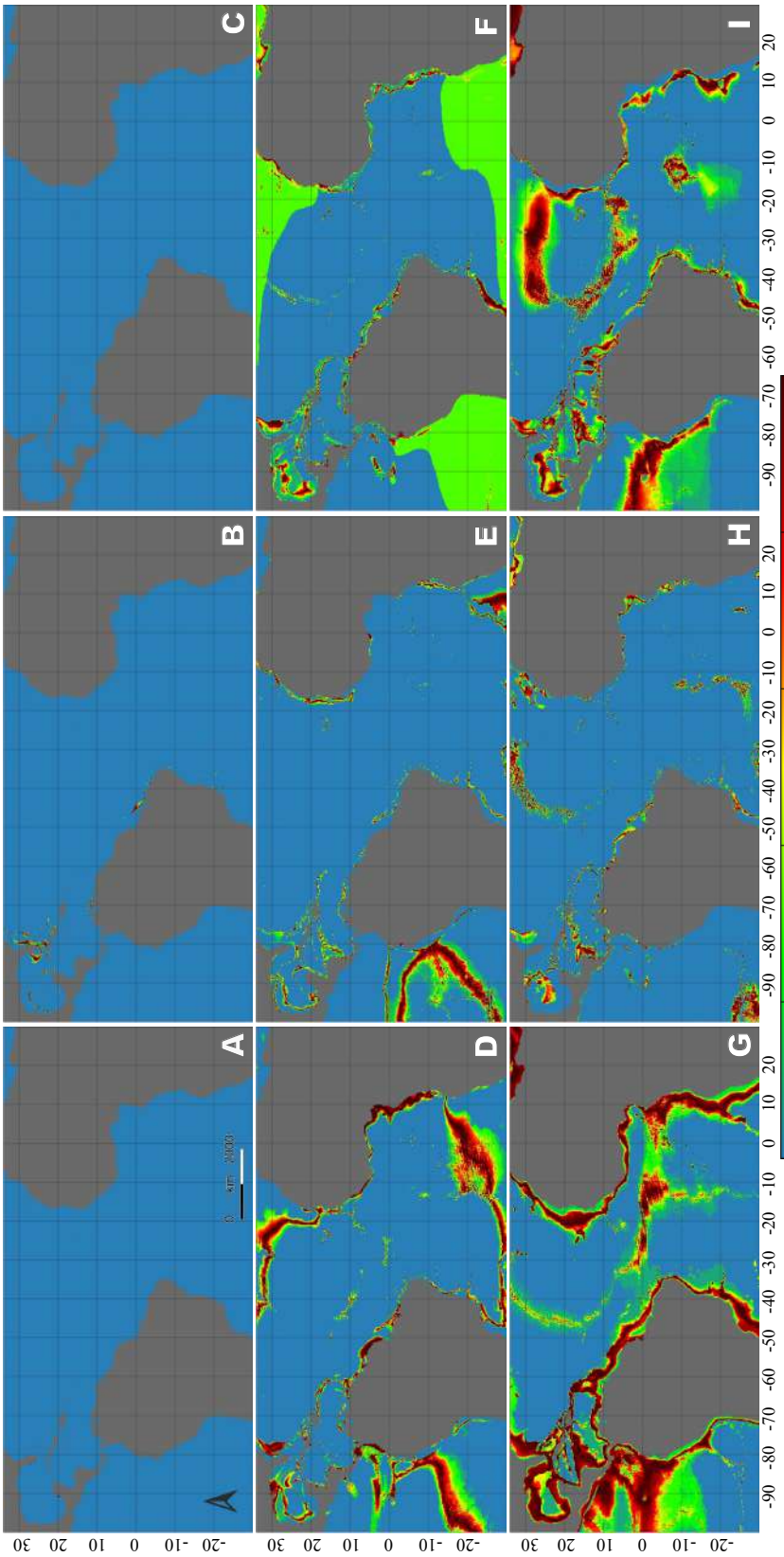

Fig. S9. Overfitting surfaces for selected ENM models, quantified using Shannon's entropy (H). *S. stellata*. A-C: fishbowl methods - A. BioClim, B. Gower, C. Mahalanobis; D-F: turbine methods - D. GAM, E. GLM, F. MARS and G-I: blackbox - G. ANN, H. MaxEnt and I. SVM (Rangel & Loyola, 2012). For any given cell, when H = 1, the method predicted species occurrence in 50% of the CVEs and absence in the other 50%. Conversely, H = 0 means that the method consistently predicted either presence or absence of the species in all CVEs.

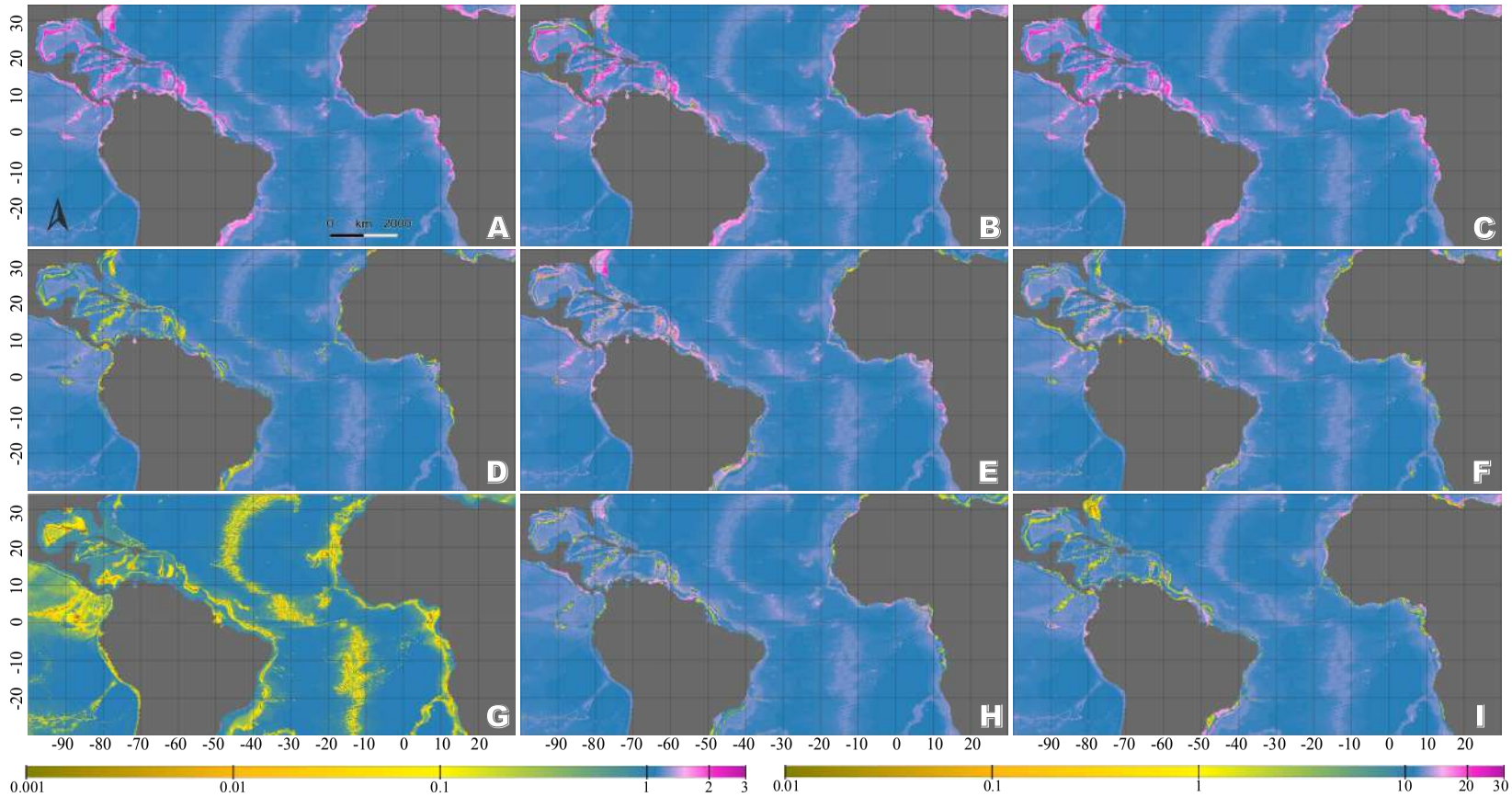

Fig. S10. Weight maps for selected ENM methods. *S. radians*. A-C: fishbowl methods - A. BioClim, B. Gower, C. Mahalanobis; D-F: turbine methods - D. GAM, E. GLM, F. MARS and G-I: blackbox - G. ANN, H. MaxEnt and I. SVM. Note the logarithmic color scales. In the scale on the left, 1 represents pixels in which all methods have equal weights (1/9) and the other values are the number of times methods were either downweighted or upweighted in other pixels. For instance, 2 means that a given method had a weight of 2/9 i.e., its votes had twice the amount of weight in the committee than what would be expected in an equal weight ensemble. The scale on the right represents the percent of the total ensemble weight that was assigned to each method. In an equal-weight ensemble, each method has 11% (1/9) of the total weight. The maximum value on the left scale is 7, which corresponds to 7/9 or 78% on the right hand scale.

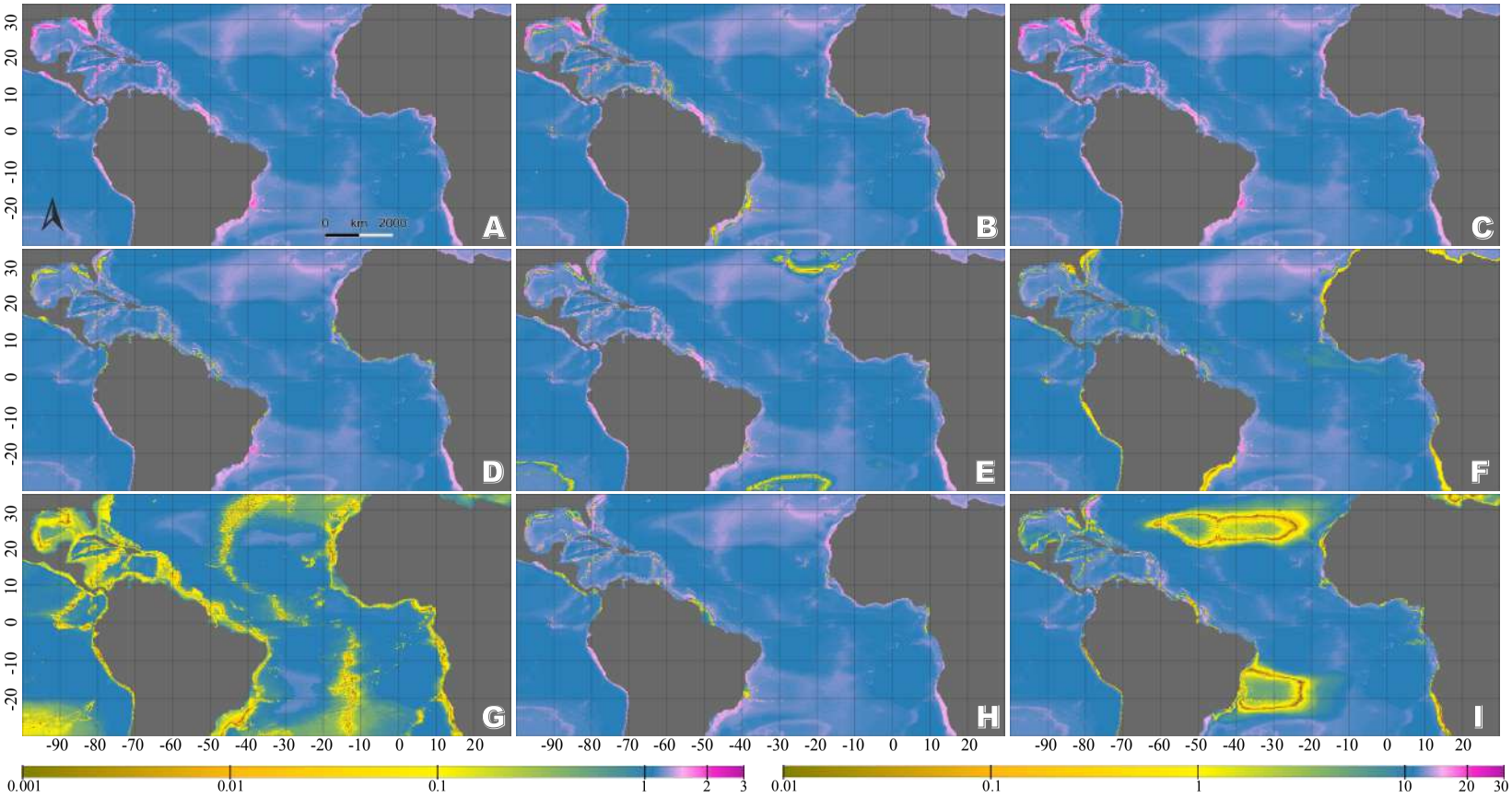

Fig. S11. Weight maps for selected ENM methods. *S. siderea*. A-C: fishbowl methods - A. BioClim, B. Gower, C. Mahalanobis; D-F: turbine methods - D. GAM, E. GLM, F. MARS and G-I: blackbox - G. ANN, H. MaxEnt and I. SVM. Note the logarithmic color scales. In the scale on the left, 1 represents pixels in which all methods have equal weights (1/9) and the other values are the number of times methods were either downweighted or upweighted in other pixels. For instance, 2 means that a given method had a weight of 2/9 i.e., its votes had twice the amount of weight in the committee than what would be expected in an equal weight ensemble. The scale on the right represents the percent of the total ensemble weight that was assigned to each method. In an equal-weight ensemble, each method has 11% (1/9) of the total weight. The maximum value on the left scale is 7, which corresponds to 7/9 or 78% on the right scale.

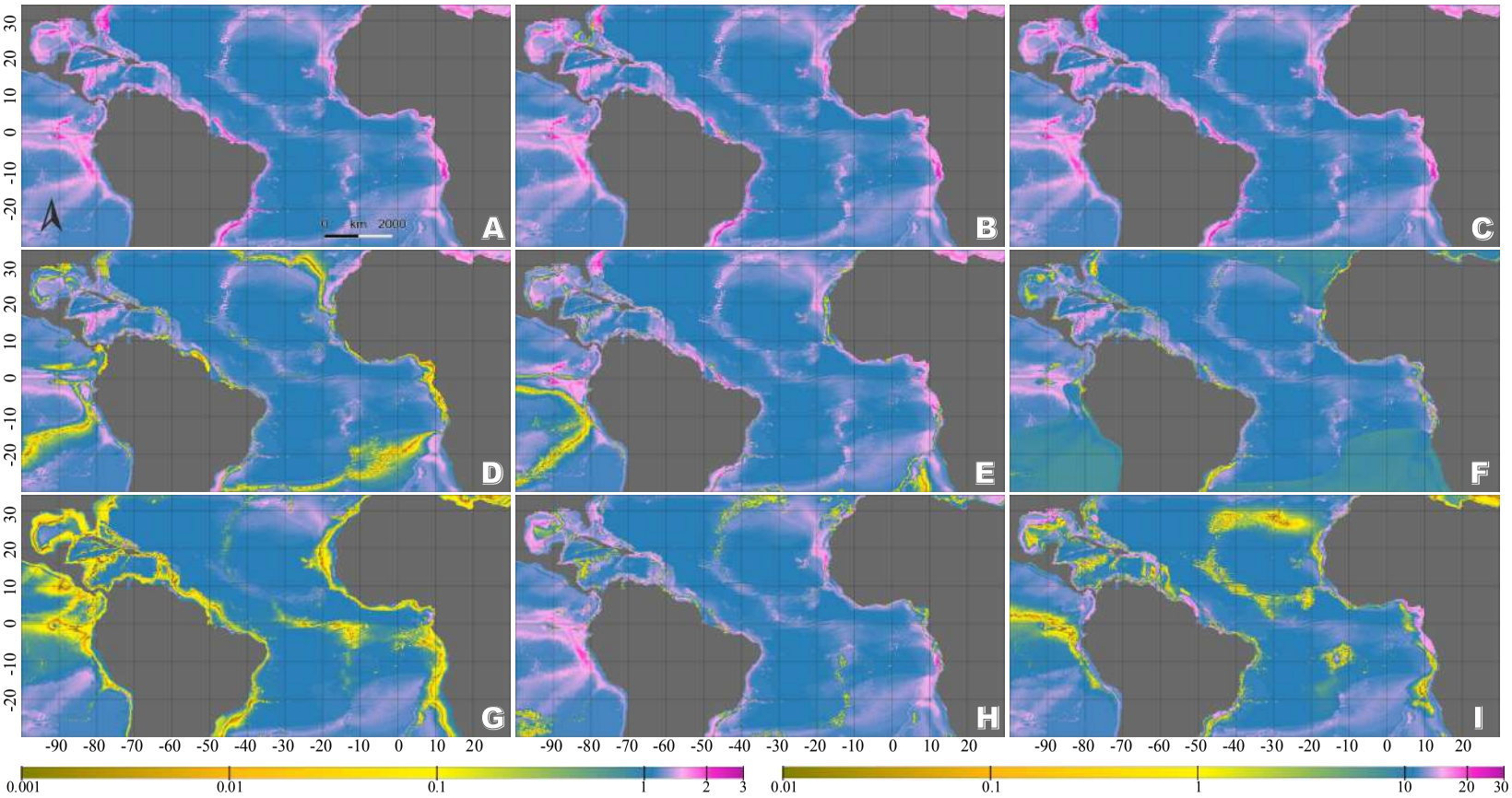

Fig. S12. Weight maps for selected ENM methods. S. *S. stellata*. A-C: fishbowl methods - A. BioClim, B. Gower, C. Mahalanobis; D-F: turbine methods - D. GAM, E. GLM, F. MARS and G-I: blackbox - G. ANN, H. MaxEnt and I. SVM. Note the logarithmic color scales. In the scale on the left, 1 represents pixels in which all methods have equal weights (1/9) and the other values are the number of times methods were either downweighted or upweighted in other pixels. For instance, 2 means that a given method had a weight of 2/9 i.e., its votes had twice the amount of weight in the committee than what would be expected in an equal-weight ensemble. The scale on the right represents the percent of the total ensemble weight that was assigned to each method. In an equal-weight ensemble, each method has 11% (1/9) of the total weight. The maximum value on the left scale is 7, which corresponds to 7/9 or 78% on the right scale.

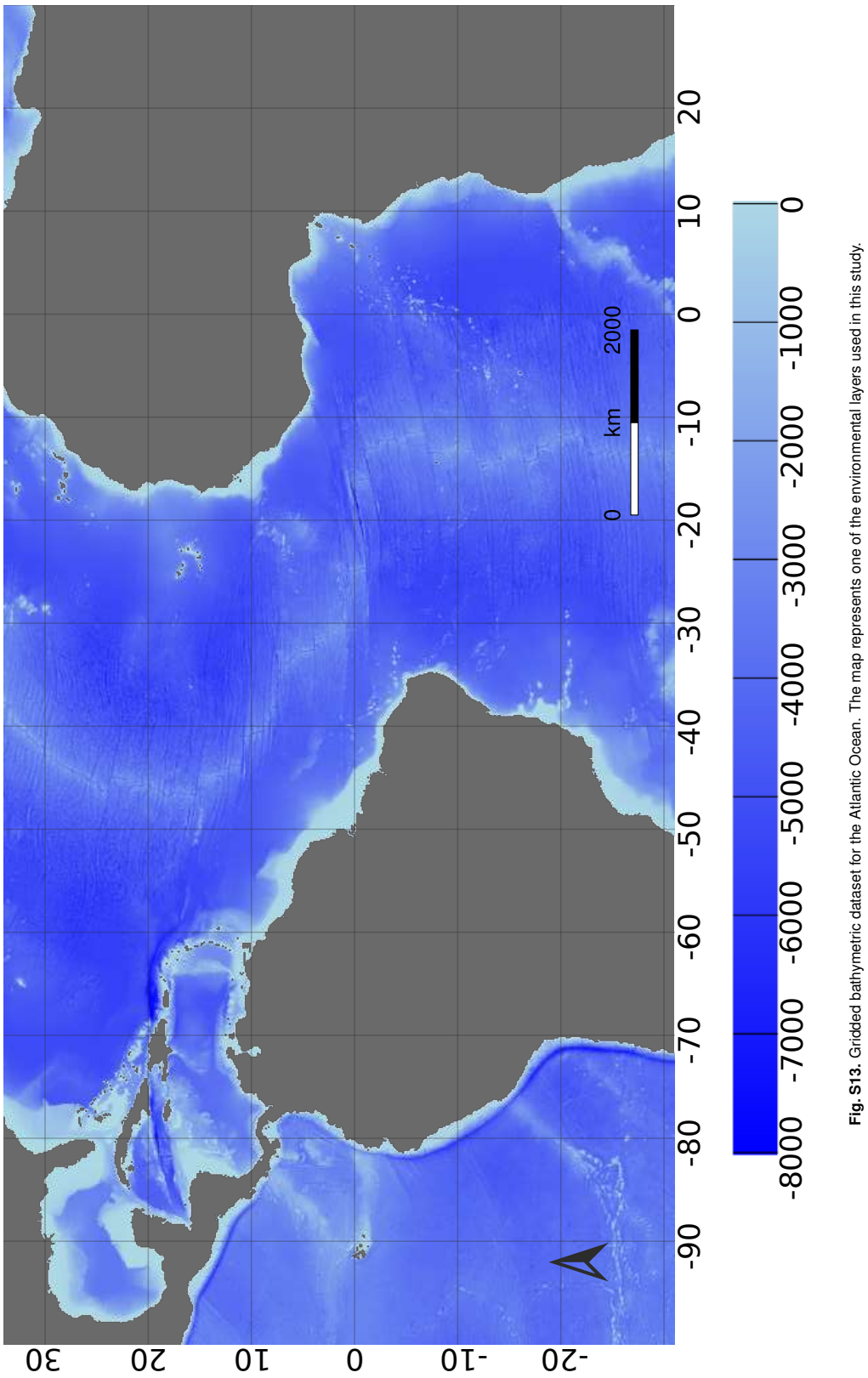

Fig. S13. Gridded bathymetric dataset for the Atlantic Ocean. The map represents one of the environmental layers used in this study.

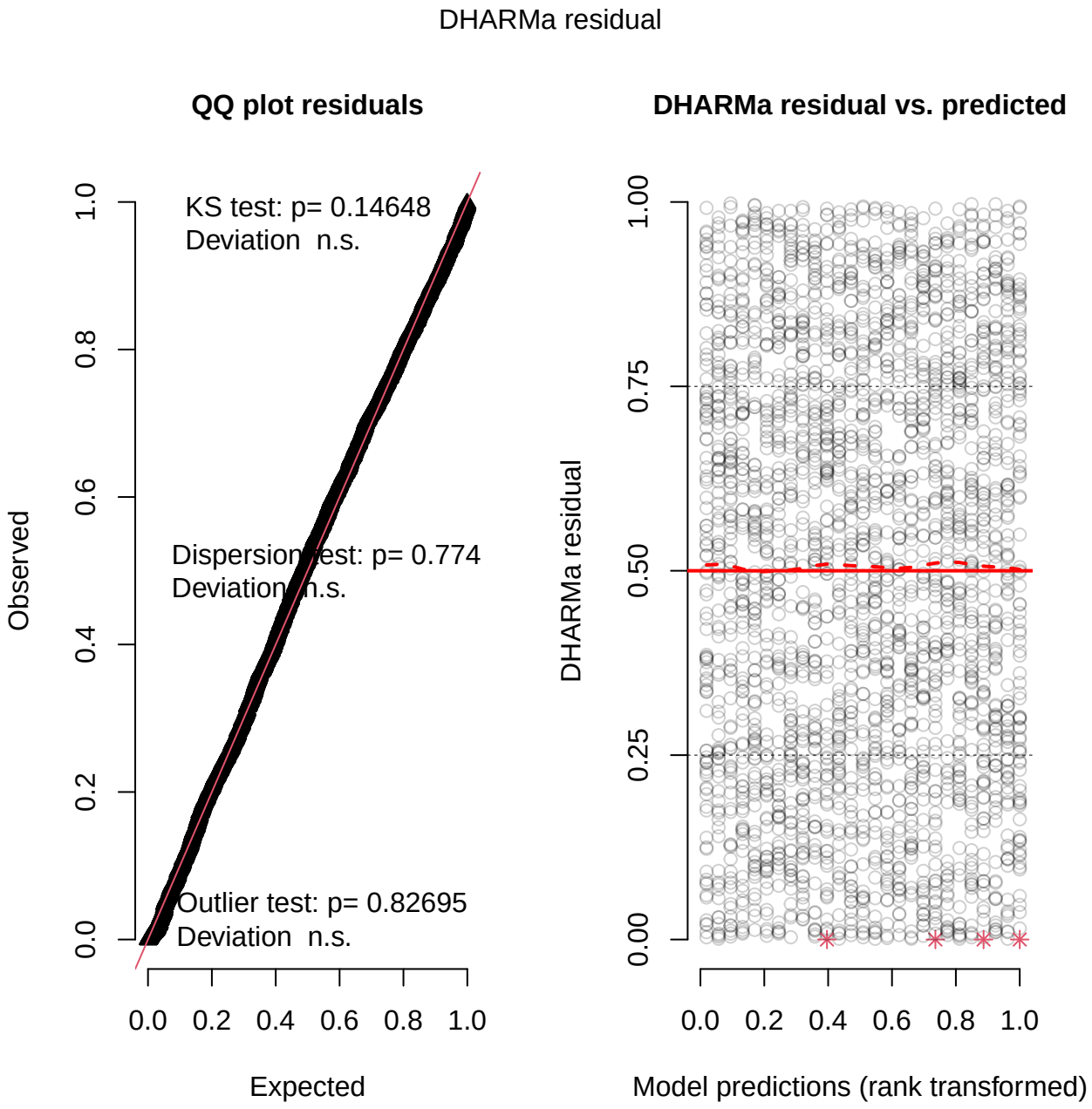

Fig. S14. DHARMA diagnostic residual analysis for the GLMM applied to entropy data in Fig 3a.

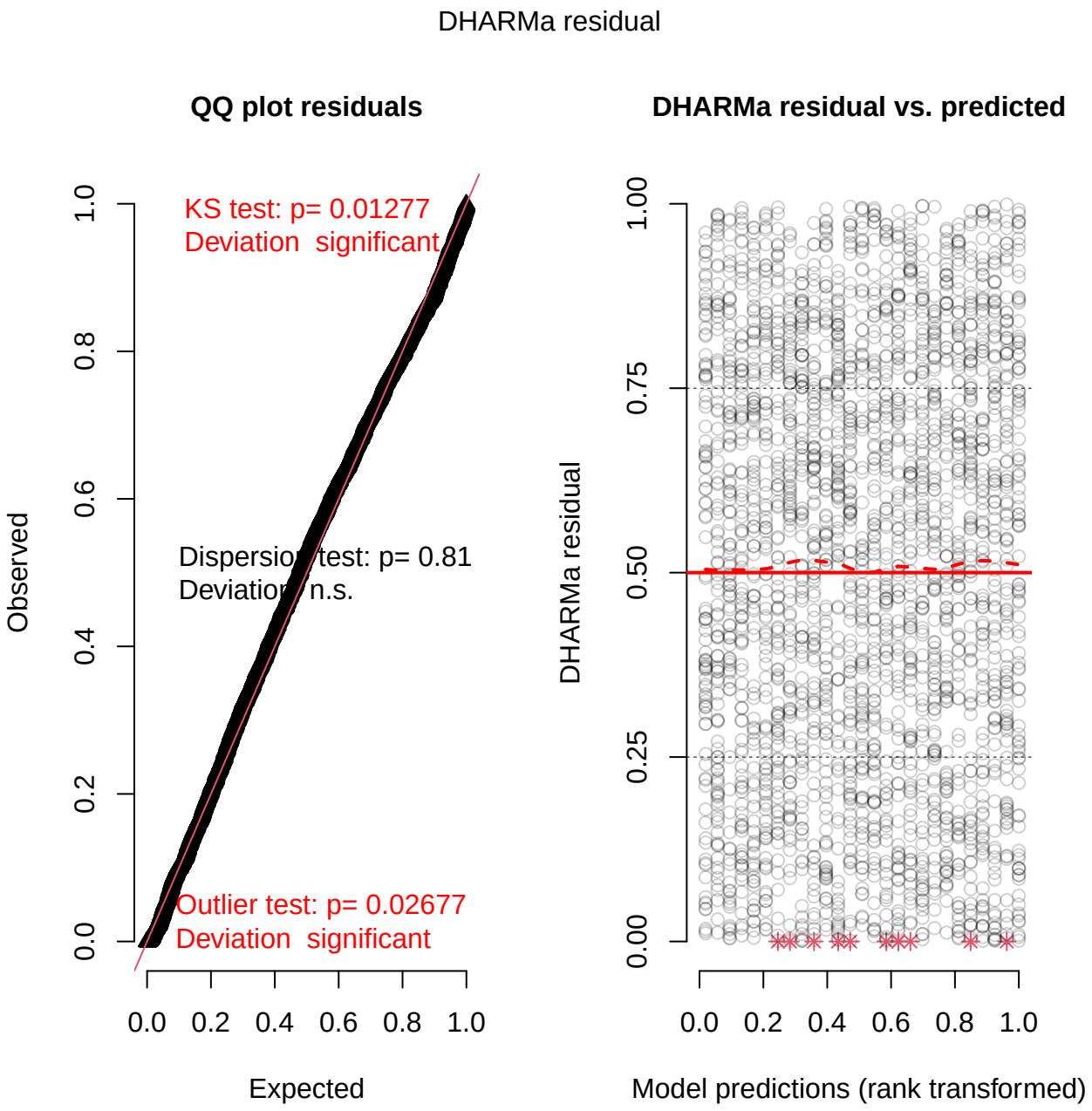

Fig. S15. DHARMA diagnostic residual analysis for the GLMM applied to TSS data in Fig. 3b.

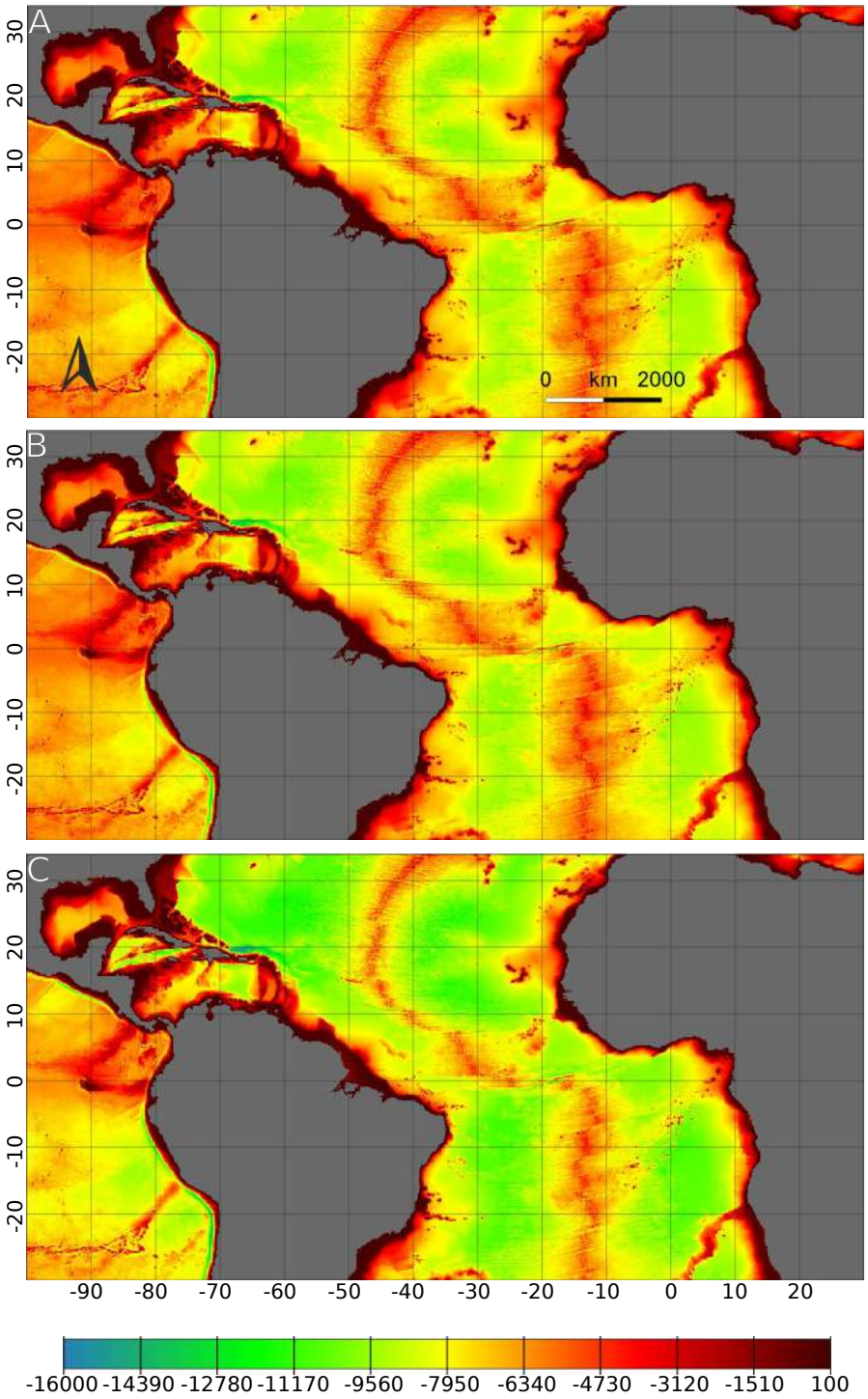

Fig. S16. Results of transferability (MESS analysis) for A. *S. radians*, B. *S. siderea*, C.*S. stellata*. Transferability increases with values in the color scale.

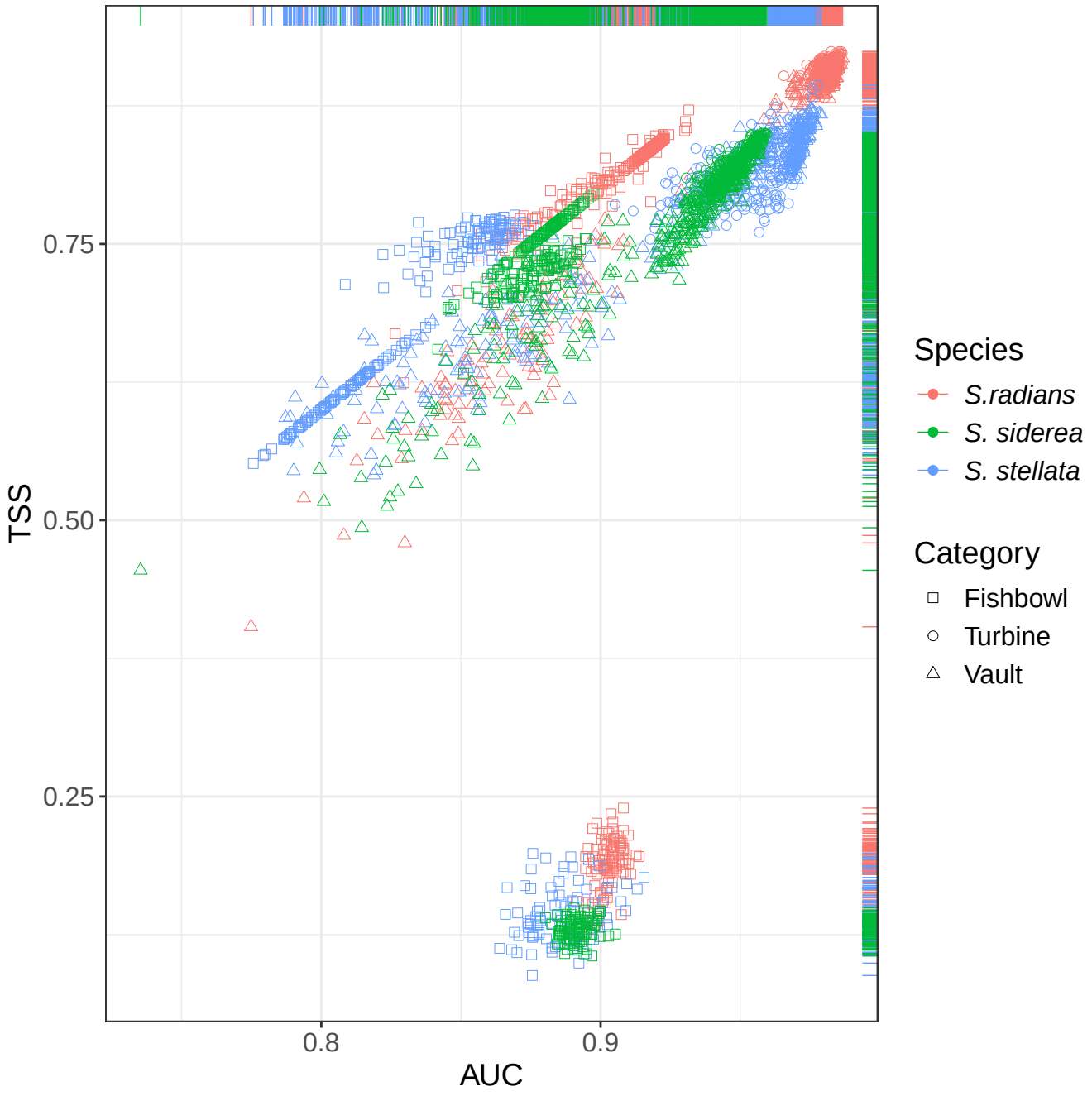

Fig. S17. Correlation rug plot between TSS scores and AUCs. Symbols vary according to categories and species are color coded according to the legend. Colored tick marks reflect point density along each axis.

Supporting Information for

Ensemble learning in ecological niche modeling of the *Siderastrea* (Cnidaria:Scleractinia) Atlantic complex: an information theory approach

Henrique J. Schipanski, Lucas B. Perroni, Caio F. Sguario and Marcos S. Barbeitos

Marcos S. Barbeitos PhD, Departamento de Zoologia, Universidade Federal do Paraná, Curitiba, PR, 81531980, Brazil.
email:msbarbeitos@gmail.com

This PDF file includes:

Table S1. List of available ENM algorithms.

Algorithm	Reference
Artificial Neural Networks (ANN)	(1)
Bayesian Additive Regression Trees (BART)	(2)
BIOCLIM (climatic envelope modeling)	(3)
Boosted Regression Trees (BRT)	(4)
Classification and Regression Trees (CART)	(5)
Classification Tree Analysis (CTA)	(6)
Domain/Gower Distance	(7)
Ecological Niche Factor Analysis (ENFA)	(8)
Generalized Additive Models (GAM)	(9)
Generalized Estimating Equations (GEE)	(10)
Generalized Linear Models (GLM)	(9)
Generalized Regression Analysis and Spatial Prediction (GRASP)	(11)
Genetic Algorithm for Rule-Set Production (GARP)	(12)
Gradient Boosting Machines (GBM)	(13)
Habitat Suitability Index (HSI)	(14)
Mahalanobis Distance	(15)
Maximum Entropy (MaxEnt)	(16)
Mixture Discriminant Analysis (MDA)	(17)
Multiple Regression (Linear or Logistic)	(9)
Multivariate Adaptive Regression Splines (MARS)	(18)
Random Forest (RF)	(19)
Regularized Discriminant Analysis (RDA)	(20)
Regularized Regression (RGLM)	(21)
Support Vector Machines (SVM)	(22)

Table S2. Post-hoc pairwise comparisons of bootstrapped BS scores graphed in main text's Fig. 3a, using least-square means computed from the fitted GLMM.

Contrast	Estimate	SE	Z ratio	p value
S. radians ANN - *S. siderea* ANN	-0.0621	0.0119	-5.2258	0.0001
S. radians ANN - *S. stellata* ANN	0.2059	0.0128	16.1086	0.0000
S. radians ANN - *S. radians* Maxent	2.3361	0.0132	177.3000	0.0000
S. radians ANN - *S. siderea* Maxent	3.0636	0.0114	268.4141	0.0000
S. radians ANN - *S. stellata* Maxent	1.7503	0.0113	154.8240	0.0000
S. radians ANN - *S. radians* SVM	2.0037	0.0115	174.8582	0.0000
S. radians ANN - *S. siderea* SVM	0.5587	0.0108	51.5147	0.0000
S. radians ANN - *S. stellata* SVM	0.5569	0.0112	49.7742	0.0000
S. radians Bioclim - *S. radians* ANN	-6.9104	0.0123	-563.1491	0.0000
S. radians Bioclim - *S. siderea* ANN	-6.9724	0.0095	-736.3725	0.0000
S. radians Bioclim - *S. stellata* ANN	-6.7045	0.0106	-633.5859	0.0000
S. radians Bioclim - *S. siderea* Bioclim	-0.1220	0.0110	-11.0656	0.0000
S. radians Bioclim - *S. stellata* Bioclim	1.0012	0.0077	130.1796	0.0000
S. radians Bioclim - *S. radians* GAM	-5.0157	0.0082	-611.7959	0.0000
S. radians Bioclim - *S. siderea* GAM	-4.0014	0.0082	-490.4652	0.0000
S. radians Bioclim - *S. stellata* GAM	-6.0544	0.0084	-721.3199	0.0000
S. radians Bioclim - *S. radians* GLM	-4.0218	0.0120	-334.6633	0.0000
S. radians Bioclim - *S. siderea* GLM	-4.5509	0.0097	-470.6762	0.0000
S. radians Bioclim - *S. stellata* GLM	-5.2707	0.0091	-581.1339	0.0000
S. radians Bioclim - *S. radians* Gower	-3.4105	0.0089	-381.3503	0.0000
S. radians Bioclim - *S. siderea* Gower	-3.8648	0.0085	-453.9449	0.0000
S. radians Bioclim - *S. stellata* Gower	-2.5596	0.0098	-260.5590	0.0000
S. radians Bioclim - *S. radians* Mahal	0.6827	0.0071	96.1593	0.0000
S. radians Bioclim - *S. siderea* Mahal	0.4141	0.0071	58.3249	0.0000
S. radians Bioclim - *S. stellata* Mahal	1.9403	0.0074	261.4584	0.0000
S. radians Bioclim - *S. radians* MARS	-4.5559	0.0085	-533.1263	0.0000

Continued on next page

Table S2 – continued from previous page

Contrast	Estimate	SE	Z ratio	p value
S. radians Bioclim - *S. siderea* MARS	-4.9313	0.0094	-525.0905	0.0000
S. radians Bioclim - *S. stellata* MARS	-6.0342	0.0203	-296.7280	0.0000
S. radians Bioclim - *S. radians* Maxent	-4.5743	0.0111	-413.7598	0.0000
S. radians Bioclim - *S. siderea* Maxent	-3.8467	0.0089	-433.1015	0.0000
S. radians Bioclim - *S. stellata* Maxent	-5.1601	0.0087	-590.2815	0.0000
S. radians Bioclim - *S. radians* SVM	-4.9067	0.0089	-548.8699	0.0000
S. radians Bioclim - *S. siderea* SVM	-6.3517	0.0081	-780.5530	0.0000
S. radians Bioclim - *S. stellata* SVM	-6.3534	0.0086	-739.5272	0.0000
S. radians GAM - *S. radians* ANN	-1.8947	0.0109	-173.9813	0.0000
S. radians GAM - *S. siderea* ANN	-1.9568	0.0076	-257.6561	0.0000
S. radians GAM - *S. stellata* ANN	-1.6888	0.0089	-188.8191	0.0000
S. radians GAM - *S. siderea* GAM	1.0142	0.0059	172.4738	0.0000
S. radians GAM - *S. stellata* GAM	-1.0387	0.0062	-167.4676	0.0000
S. radians GAM - *S. radians* GLM	0.9939	0.0106	93.7262	0.0000
S. radians GAM - *S. siderea* GLM	0.4648	0.0078	59.2611	0.0000
S. radians GAM - *S. stellata* GLM	-0.2550	0.0071	-35.9684	0.0000
S. radians GAM - *S. radians* MARS	0.4598	0.0064	71.7586	0.0000
S. radians GAM - *S. siderea* MARS	0.0844	0.0075	11.2565	0.0000
S. radians GAM - *S. stellata* MARS	-1.0185	0.0195	-52.1404	0.0000
S. radians GAM - *S. radians* Maxent	0.4414	0.0095	46.4656	0.0000
S. radians GAM - *S. siderea* Maxent	1.1689	0.0068	170.6688	0.0000
S. radians GAM - *S. stellata* Maxent	-0.1444	0.0067	-21.6625	0.0000
S. radians GAM - *S. radians* SVM	0.1090	0.0069	15.7373	0.0000
S. radians GAM - *S. siderea* SVM	-1.3360	0.0059	-228.3277	0.0000
S. radians GAM - *S. stellata* SVM	-1.3378	0.0065	-206.8412	0.0000
S. radians GLM - *S. radians* ANN	-2.8886	0.0140	-206.4277	0.0000
S. radians GLM - *S. siderea* ANN	-2.9506	0.0116	-254.0584	0.0000
S. radians GLM - *S. stellata* ANN	-2.6827	0.0125	-213.9610	0.0000
S. radians GLM - *S. siderea* GLM	-0.5291	0.0118	-44.9221	0.0000
S. radians GLM - *S. stellata* GLM	-1.2489	0.0113	-110.6096	0.0000
S. radians GLM - *S. radians* MARS	-0.5341	0.0109	-49.1137	0.0000
S. radians GLM - *S. siderea* MARS	-0.9095	0.0116	-78.7342	0.0000
S. radians GLM - *S. stellata* MARS	-2.0123	0.0214	-93.9520	0.0000
S. radians GLM - *S. radians* Maxent	-0.5525	0.0129	-42.6929	0.0000
S. radians GLM - *S. siderea* Maxent	0.1751	0.0111	15.7135	0.0000
S. radians GLM - *S. stellata* Maxent	-1.1383	0.0110	-103.2030	0.0000
S. radians GLM - *S. radians* SVM	-0.8849	0.0112	-79.1007	0.0000
S. radians GLM - *S. siderea* SVM	-2.3299	0.0106	-220.7008	0.0000
S. radians GLM - *S. stellata* SVM	-2.3316	0.0109	-213.7062	0.0000
S. radians Gower - *S. radians* ANN	-3.4999	0.0115	-305.3566	0.0000
S. radians Gower - *S. siderea* ANN	-3.5619	0.0084	-424.3844	0.0000
S. radians Gower - *S. stellata* ANN	-3.2940	0.0096	-342.0003	0.0000
S. radians Gower - *S. radians* GAM	-1.6052	0.0069	-231.6797	0.0000
S. radians Gower - *S. siderea* GAM	-0.5909	0.0069	-85.8753	0.0000
S. radians Gower - *S. stellata* GAM	-2.6439	0.0072	-369.3463	0.0000
S. radians Gower - *S. radians* GLM	-0.6113	0.0112	-54.6315	0.0000
S. radians Gower - *S. siderea* GLM	-1.1404	0.0086	-132.3197	0.0000
S. radians Gower - *S. stellata* GLM	-1.8602	0.0079	-234.2714	0.0000
S. radians Gower - *S. siderea* Gower	-0.4543	0.0073	-62.2467	0.0000
S. radians Gower - *S. stellata* Gower	0.8509	0.0088	96.7859	0.0000

Continued on next page

Table S2 – continued from previous page

Contrast	Estimate	SE	Z ratio	p value
S. radians Gower - *S. radians* Mahal	4.0932	0.0056	732.8053	0.0000
S. radians Gower - *S. siderea* Mahal	3.8246	0.0056	684.6045	0.0000
S. radians Gower - *S. stellata* Mahal	5.3508	0.0060	893.4996	0.0000
S. radians Gower - *S. radians* MARS	-1.1454	0.0073	-156.1309	0.0000
S. radians Gower - *S. siderea* MARS	-1.5208	0.0083	-183.0970	0.0000
S. radians Gower - *S. stellata* MARS	-2.6237	0.0199	-132.1227	0.0000
S. radians Gower - *S. radians* Maxent	-1.1638	0.0101	-114.6636	0.0000
S. radians Gower - *S. siderea* Maxent	-0.4363	0.0077	-56.4713	0.0000
S. radians Gower - *S. stellata* Maxent	-1.7496	0.0076	-231.3143	0.0000
S. radians Gower - *S. radians* SVM	-1.4962	0.0078	-192.0278	0.0000
S. radians Gower - *S. siderea* SVM	-2.9412	0.0069	-428.9811	0.0000
S. radians Gower - *S. stellata* SVM	-2.9429	0.0074	-398.2788	0.0000
S. radians Mahal - *S. radians* ANN	-7.5931	0.0101	-752.5814	0.0000
S. radians Mahal - *S. siderea* ANN	-7.6552	0.0064	-1197.3822	0.0000
S. radians Mahal - *S. stellata* ANN	-7.3872	0.0079	-929.2692	0.0000
S. radians Mahal - *S. radians* GAM	-5.6984	0.0043	-1327.3378	0.0000
S. radians Mahal - *S. siderea* GAM	-4.6841	0.0042	-1110.9201	0.0000
S. radians Mahal - *S. stellata* GAM	-6.7371	0.0047	-1447.2578	0.0000
S. radians Mahal - *S. radians* GLM	-4.7045	0.0098	-481.0613	0.0000
S. radians Mahal - *S. siderea* GLM	-5.2336	0.0067	-782.7403	0.0000
S. radians Mahal - *S. stellata* GLM	-5.9534	0.0058	-1028.9267	0.0000
S. radians Mahal - *S. siderea* Mahal	-0.2686	0.0013	-209.8046	0.0000
S. radians Mahal - *S. stellata* Mahal	1.2576	0.0025	501.3779	0.0000
S. radians Mahal - *S. radians* MARS	-5.2386	0.0049	-1063.8543	0.0000
S. radians Mahal - *S. siderea* MARS	-5.6140	0.0063	-894.2177	0.0000
S. radians Mahal - *S. stellata* MARS	-6.7169	0.0191	-351.6946	0.0000
S. radians Mahal - *S. radians* Maxent	-5.2570	0.0086	-613.4440	0.0000
S. radians Mahal - *S. siderea* Maxent	-4.5295	0.0055	-825.5003	0.0000
S. radians Mahal - *S. stellata* Maxent	-5.8428	0.0053	-1111.4214	0.0000
S. radians Mahal - *S. radians* SVM	-5.5894	0.0056	-1001.6799	0.0000
S. radians Mahal - *S. siderea* SVM	-7.0344	0.0042	-1684.5965	0.0000
S. radians Mahal - *S. stellata* SVM	-7.0362	0.0050	-1406.4527	0.0000
S. radians MARS - *S. radians* ANN	-2.3545	0.0112	-211.0841	0.0000
S. radians MARS - *S. siderea* ANN	-2.4165	0.0080	-303.2684	0.0000
S. radians MARS - *S. stellata* ANN	-2.1486	0.0093	-231.9384	0.0000
S. radians MARS - *S. siderea* MARS	-0.3754	0.0079	-47.6577	0.0000
S. radians MARS - *S. stellata* MARS	-1.4782	0.0197	-75.1070	0.0000
S. radians MARS - *S. radians* Maxent	-0.0184	0.0098	-1.8731	0.9852
S. radians MARS - *S. siderea* Maxent	0.7092	0.0073	97.6625	0.0000
S. radians MARS - *S. stellata* Maxent	-0.6042	0.0071	-85.2242	0.0000
S. radians MARS - *S. radians* SVM	-0.3508	0.0073	-47.8455	0.0000
S. radians MARS - *S. siderea* SVM	-1.7958	0.0063	-283.7412	0.0000
S. radians MARS - *S. stellata* SVM	-1.7975	0.0069	-260.4095	0.0000
S. radians Maxent - *S. siderea* Maxent	0.7275	0.0101	72.0640	0.0000
S. radians Maxent - *S. stellata* Maxent	-0.5858	0.0100	-58.7437	0.0000
S. radians Maxent - *S. radians* SVM	-0.3324	0.0101	-32.7645	0.0000
S. radians Maxent - *S. siderea* SVM	-1.7775	0.0094	-188.1487	0.0000
S. radians Maxent - *S. stellata* SVM	-1.7792	0.0098	-180.7993	0.0000
S. radians SVM - *S. siderea* SVM	-1.4450	0.0069	-210.8979	0.0000
S. radians SVM - *S. stellata* SVM	-1.4467	0.0074	-195.9047	0.0000

Continued on next page

Table S2 – continued from previous page

Contrast	Estimate	SE	Z ratio	p value
S. siderea ANN - *S. stellata* ANN	0.2680	0.0101	26.5178	0.0000
S. siderea ANN - *S. radians* Maxent	2.3982	0.0106	225.9170	0.0000
S. siderea ANN - *S. siderea* Maxent	3.1257	0.0083	377.1762	0.0000
S. siderea ANN - *S. stellata* Maxent	1.8124	0.0082	222.1722	0.0000
S. siderea ANN - *S. radians* SVM	2.0657	0.0084	246.2297	0.0000
S. siderea ANN - *S. siderea* SVM	0.6207	0.0075	82.9424	0.0000
S. siderea ANN - *S. stellata* SVM	0.6190	0.0080	77.4131	0.0000
S. siderea Bioclim - *S. radians* ANN	-6.7884	0.0131	-516.2993	0.0000
S. siderea Bioclim - *S. siderea* ANN	-6.8505	0.0105	-649.4096	0.0000
S. siderea Bioclim - *S. stellata* ANN	-6.5825	0.0116	-568.7805	0.0000
S. siderea Bioclim - *S. stellata* Bioclim	1.1232	0.0090	124.7121	0.0000
S. siderea Bioclim - *S. radians* GAM	-4.8937	0.0095	-517.2515	0.0000
S. siderea Bioclim - *S. siderea* GAM	-3.8795	0.0094	-413.1255	0.0000
S. siderea Bioclim - *S. stellata* GAM	-5.9324	0.0096	-617.1225	0.0000
S. siderea Bioclim - *S. radians* GLM	-3.8998	0.0129	-302.0343	0.0000
S. siderea Bioclim - *S. siderea* GLM	-4.4289	0.0107	-412.8042	0.0000
S. siderea Bioclim - *S. stellata* GLM	-5.1487	0.0102	-504.3440	0.0000
S. siderea Bioclim - *S. radians* Gower	-3.2885	0.0101	-325.1684	0.0000
S. siderea Bioclim - *S. siderea* Gower	-3.7429	0.0097	-385.8269	0.0000
S. siderea Bioclim - *S. stellata* Gower	-2.4376	0.0109	-223.9657	0.0000
S. siderea Bioclim - *S. radians* Mahal	0.8047	0.0085	94.3713	0.0000
S. siderea Bioclim - *S. siderea* Mahal	0.5361	0.0085	63.1634	0.0000
S. siderea Bioclim - *S. stellata* Mahal	2.0623	0.0088	234.9682	0.0000
S. siderea Bioclim - *S. radians* MARS	-4.4339	0.0098	-454.1320	0.0000
S. siderea Bioclim - *S. siderea* MARS	-4.8093	0.0105	-458.9302	0.0000
S. siderea Bioclim - *S. stellata* MARS	-5.9122	0.0209	-283.3055	0.0000
S. siderea Bioclim - *S. radians* Maxent	-4.4523	0.0120	-370.3591	0.0000
S. siderea Bioclim - *S. siderea* Maxent	-3.7248	0.0100	-371.5329	0.0000
S. siderea Bioclim - *S. stellata* Maxent	-5.0381	0.0099	-507.9480	0.0000
S. siderea Bioclim - *S. radians* SVM	-4.7847	0.0101	-473.2585	0.0000
S. siderea Bioclim - *S. siderea* SVM	-6.2297	0.0094	-664.6984	0.0000
S. siderea Bioclim - *S. stellata* SVM	-6.2315	0.0098	-636.7642	0.0000
S. siderea GAM - *S. radians* ANN	-2.9090	0.0109	-267.8512	0.0000
S. siderea GAM - *S. siderea* ANN	-2.9710	0.0075	-395.7917	0.0000
S. siderea GAM - *S. stellata* ANN	-2.7031	0.0089	-304.1054	0.0000
S. siderea GAM - *S. stellata* GAM	-2.0530	0.0061	-335.3305	0.0000
S. siderea GAM - *S. radians* GLM	-0.0204	0.0106	-1.9285	0.9785
S. siderea GAM - *S. siderea* GLM	-0.5495	0.0078	-70.8303	0.0000
S. siderea GAM - *S. stellata* GLM	-1.2693	0.0070	-180.7920	0.0000
S. siderea GAM - *S. radians* MARS	-0.5545	0.0064	-87.2384	0.0000
S. siderea GAM - *S. siderea* MARS	-0.9298	0.0074	-125.5077	0.0000
S. siderea GAM - *S. stellata* MARS	-2.0327	0.0195	-104.1990	0.0000
S. siderea GAM - *S. radians* Maxent	-0.5728	0.0095	-60.5215	0.0000
S. siderea GAM - *S. siderea* Maxent	0.1547	0.0068	22.9096	0.0000
S. siderea GAM - *S. stellata* Maxent	-1.1587	0.0066	-175.7780	0.0000
S. siderea GAM - *S. radians* SVM	-0.9053	0.0069	-131.6468	0.0000
S. siderea GAM - *S. siderea* SVM	-2.3503	0.0057	-409.6881	0.0000
S. siderea GAM - *S. stellata* SVM	-2.3520	0.0064	-368.0430	0.0000
S. siderea GLM - *S. radians* ANN	-2.3595	0.0120	-196.0291	0.0000
S. siderea GLM - *S. siderea* ANN	-2.4215	0.0091	-265.3579	0.0000

Continued on next page

Table S2 – continued from previous page

Contrast	Estimate	SE	Z ratio	p value
S. siderea GLM - *S. stellata* ANN	-2.1536	0.0103	-209.2400	0.0000
S. siderea GLM - *S. stellata* GLM	-0.7198	0.0087	-82.4505	0.0000
S. siderea GLM - *S. radians* MARS	-0.0050	0.0082	-0.6107	1.0000
S. siderea GLM - *S. siderea* MARS	-0.3804	0.0090	-42.0520	0.0000
S. siderea GLM - *S. stellata* MARS	-1.4833	0.0202	-73.4776	0.0000
S. siderea GLM - *S. radians* Maxent	-0.0234	0.0108	-2.1650	0.9209
S. siderea GLM - *S. siderea* Maxent	0.7041	0.0085	82.6928	0.0000
S. siderea GLM - *S. stellata* Maxent	-0.6092	0.0084	-72.6158	0.0000
S. siderea GLM - *S. radians* SVM	-0.3558	0.0086	-41.3019	0.0000
S. siderea GLM - *S. siderea* SVM	-1.8008	0.0077	-232.7972	0.0000
S. siderea GLM - *S. stellata* SVM	-1.8025	0.0082	-218.9647	0.0000
S. siderea Gower - *S. radians* ANN	-3.0455	0.0111	-273.6368	0.0000
S. siderea Gower - *S. siderea* ANN	-3.1076	0.0079	-393.7961	0.0000
S. siderea Gower - *S. stellata* ANN	-2.8396	0.0092	-308.1238	0.0000
S. siderea Gower - *S. radians* GAM	-1.1508	0.0064	-180.8183	0.0000
S. siderea Gower - *S. siderea* GAM	-0.1366	0.0063	-21.8203	0.0000
S. siderea Gower - *S. stellata* GAM	-2.1896	0.0066	-332.3313	0.0000
S. siderea Gower - *S. radians* GLM	-0.1570	0.0108	-14.4677	0.0000
S. siderea Gower - *S. siderea* GLM	-0.6861	0.0081	-84.3791	0.0000
S. siderea Gower - *S. stellata* GLM	-1.4059	0.0074	-189.1952	0.0000
S. siderea Gower - *S. stellata* Gower	1.3052	0.0083	156.6143	0.0000
S. siderea Gower - *S. radians* Mahal	4.5476	0.0049	934.0336	0.0000
S. siderea Gower - *S. siderea* Mahal	4.2790	0.0048	891.4759	0.0000
S. siderea Gower - *S. stellata* Mahal	5.8051	0.0053	1096.4893	0.0000
S. siderea Gower - *S. radians* MARS	-0.6911	0.0068	-101.5341	0.0000
S. siderea Gower - *S. siderea* MARS	-1.0664	0.0078	-136.7497	0.0000
S. siderea Gower - *S. stellata* MARS	-2.1693	0.0197	-110.3444	0.0000
S. siderea Gower - *S. radians* Maxent	-0.7094	0.0098	-72.5891	0.0000
S. siderea Gower - *S. siderea* Maxent	0.0181	0.0072	2.5208	0.7178
S. siderea Gower - *S. stellata* Maxent	-1.2952	0.0070	-184.3294	0.0000
S. siderea Gower - *S. radians* SVM	-1.0419	0.0073	-142.8227	0.0000
S. siderea Gower - *S. siderea* SVM	-2.4869	0.0062	-399.0541	0.0000
S. siderea Gower - *S. stellata* SVM	-2.4886	0.0068	-363.9070	0.0000
S. siderea Mahal - *S. radians* ANN	-7.3245	0.0101	-725.9234	0.0000
S. siderea Mahal - *S. siderea* ANN	-7.3866	0.0063	-1164.8972	0.0000
S. siderea Mahal - *S. stellata* ANN	-7.1186	0.0079	-897.8129	0.0000
S. siderea Mahal - *S. radians* GAM	-5.4298	0.0043	-1264.4093	0.0000
S. siderea Mahal - *S. siderea* GAM	-4.4156	0.0041	-1067.3957	0.0000
S. siderea Mahal - *S. stellata* GAM	-6.4686	0.0046	-1400.1818	0.0000
S. siderea Mahal - *S. radians* GLM	-4.4360	0.0098	-453.5727	0.0000
S. siderea Mahal - *S. siderea* GLM	-4.9650	0.0066	-748.1631	0.0000
S. siderea Mahal - *S. stellata* GLM	-5.6849	0.0058	-987.3494	0.0000
S. siderea Mahal - *S. stellata* Mahal	1.5262	0.0024	624.9356	0.0000
S. siderea Mahal - *S. radians* MARS	-4.9701	0.0049	-1009.0879	0.0000
S. siderea Mahal - *S. siderea* MARS	-5.3454	0.0062	-858.7200	0.0000
S. siderea Mahal - *S. stellata* MARS	-6.4483	0.0191	-337.7836	0.0000
S. siderea Mahal - *S. radians* Maxent	-4.9884	0.0086	-582.0612	0.0000
S. siderea Mahal - *S. siderea* Maxent	-4.2609	0.0054	-785.2819	0.0000
S. siderea Mahal - *S. stellata* Maxent	-5.5742	0.0052	-1066.6706	0.0000
S. siderea Mahal - *S. radians* SVM	-5.3209	0.0056	-953.3835	0.0000

Continued on next page

Table S2 – continued from previous page

Contrast	Estimate	SE	Z ratio	p value
S. siderea Mahal - *S. siderea* SVM	-6.7659	0.0041	-1652.1200	0.0000
S. siderea Mahal - *S. stellata* SVM	-6.7676	0.0050	-1361.7045	0.0000
S. siderea MARS - *S. radians* ANN	-1.9791	0.0118	-167.5147	0.0000
S. siderea MARS - *S. siderea* ANN	-2.0412	0.0088	-231.1407	0.0000
S. siderea MARS - *S. stellata* ANN	-1.7732	0.0100	-176.7544	0.0000
S. siderea MARS - *S. stellata* MARS	-1.1029	0.0201	-54.9930	0.0000
S. siderea MARS - *S. radians* Maxent	0.3570	0.0105	33.8510	0.0000
S. siderea MARS - *S. siderea* Maxent	1.0845	0.0082	132.2809	0.0000
S. siderea MARS - *S. stellata* Maxent	-0.2288	0.0081	-28.3616	0.0000
S. siderea MARS - *S. radians* SVM	0.0246	0.0083	2.9587	0.3696
S. siderea MARS - *S. siderea* SVM	-1.4204	0.0074	-192.3270	0.0000
S. siderea MARS - *S. stellata* SVM	-1.4222	0.0079	-179.9270	0.0000
S. siderea Maxent - *S. stellata* Maxent	-1.3133	0.0075	-175.8510	0.0000
S. siderea Maxent - *S. radians* SVM	-1.0600	0.0077	-137.2799	0.0000
S. siderea Maxent - *S. siderea* SVM	-2.5050	0.0067	-372.4304	0.0000
S. siderea Maxent - *S. stellata* SVM	-2.5067	0.0073	-343.7759	0.0000
S. siderea SVM - *S. stellata* SVM	-0.0017	0.0064	-0.2713	1.0000
S. stellata ANN - *S. radians* Maxent	2.1302	0.0116	183.3364	0.0000
S. stellata ANN - *S. siderea* Maxent	2.8577	0.0096	299.0226	0.0000
S. stellata ANN - *S. stellata* Maxent	1.5444	0.0094	164.1443	0.0000
S. stellata ANN - *S. radians* SVM	1.7978	0.0096	186.7194	0.0000
S. stellata ANN - *S. siderea* SVM	0.3528	0.0089	39.7738	0.0000
S. stellata ANN - *S. stellata* SVM	0.3510	0.0093	37.8718	0.0000
S. stellata Bioclim - *S. radians* ANN	-7.9116	0.0105	-752.4947	0.0000
S. stellata Bioclim - *S. siderea* ANN	-7.9737	0.0070	-1135.8542	0.0000
S. stellata Bioclim - *S. stellata* ANN	-7.7057	0.0084	-912.8151	0.0000
S. stellata Bioclim - *S. radians* GAM	-6.0169	0.0052	-1154.2117	0.0000
S. stellata Bioclim - *S. siderea* GAM	-5.0026	0.0051	-977.6304	0.0000
S. stellata Bioclim - *S. stellata* GAM	-7.0556	0.0055	-1293.8363	0.0000
S. stellata Bioclim - *S. radians* GLM	-5.0230	0.0102	-491.6451	0.0000
S. stellata Bioclim - *S. siderea* GLM	-5.5521	0.0073	-761.8334	0.0000
S. stellata Bioclim - *S. stellata* GLM	-6.2719	0.0064	-973.0459	0.0000
S. stellata Bioclim - *S. radians* Gower	-4.4117	0.0063	-698.0398	0.0000
S. stellata Bioclim - *S. siderea* Gower	-4.8661	0.0057	-858.7196	0.0000
S. stellata Bioclim - *S. stellata* Gower	-3.5608	0.0075	-476.7406	0.0000
S. stellata Bioclim - *S. radians* Mahal	-0.3185	0.0032	-98.8949	0.0000
S. stellata Bioclim - *S. siderea* Mahal	-0.5871	0.0032	-185.2341	0.0000
S. stellata Bioclim - *S. stellata* Mahal	0.9391	0.0038	247.8167	0.0000
S. stellata Bioclim - *S. radians* MARS	-5.5571	0.0057	-967.4851	0.0000
S. stellata Bioclim - *S. siderea* MARS	-5.9325	0.0069	-857.8833	0.0000
S. stellata Bioclim - *S. stellata* MARS	-7.0354	0.0193	-364.3637	0.0000
S. stellata Bioclim - *S. radians* Maxent	-5.5755	0.0091	-615.0240	0.0000
S. stellata Bioclim - *S. siderea* Maxent	-4.8480	0.0062	-781.1912	0.0000
S. stellata Bioclim - *S. stellata* Maxent	-6.1613	0.0060	-1031.1216	0.0000
S. stellata Bioclim - *S. radians* SVM	-5.9079	0.0063	-935.5116	0.0000
S. stellata Bioclim - *S. siderea* SVM	-7.3529	0.0051	-1446.3986	0.0000
S. stellata Bioclim - *S. stellata* SVM	-7.3547	0.0058	-1278.4268	0.0000
S. stellata GAM - *S. radians* ANN	-0.8560	0.0110	-77.5470	0.0000
S. stellata GAM - *S. siderea* ANN	-0.9180	0.0078	-117.9507	0.0000
S. stellata GAM - *S. stellata* ANN	-0.6501	0.0091	-71.5451	0.0000

Continued on next page

Table S2 – continued from previous page

Contrast	Estimate	SE	Z ratio	p value
S. stellata GAM - *S. radians* GLM	2.0326	0.0108	188.9823	0.0000
S. stellata GAM - *S. siderea* GLM	1.5035	0.0080	187.3396	0.0000
S. stellata GAM - *S. stellata* GLM	0.7837	0.0073	107.8069	0.0000
S. stellata GAM - *S. radians* MARS	1.4985	0.0067	225.1672	0.0000
S. stellata GAM - *S. siderea* MARS	1.1231	0.0077	146.0738	0.0000
S. stellata GAM - *S. stellata* MARS	0.0203	0.0196	1.0333	1.0000
S. stellata GAM - *S. radians* Maxent	1.4801	0.0097	153.0887	0.0000
S. stellata GAM - *S. siderea* Maxent	2.2077	0.0071	312.8036	0.0000
S. stellata GAM - *S. stellata* Maxent	0.8943	0.0069	130.4487	0.0000
S. stellata GAM - *S. radians* SVM	1.1477	0.0072	160.4292	0.0000
S. stellata GAM - *S. siderea* SVM	-0.2973	0.0061	-48.7849	0.0000
S. stellata GAM - *S. stellata* SVM	-0.2990	0.0067	-44.8813	0.0000
S. stellata GLM - *S. radians* ANN	-1.6397	0.0116	-141.8319	0.0000
S. stellata GLM - *S. siderea* ANN	-1.7017	0.0085	-200.0140	0.0000
S. stellata GLM - *S. stellata* ANN	-1.4338	0.0097	-147.5931	0.0000
S. stellata GLM - *S. radians* MARS	0.7148	0.0075	95.4367	0.0000
S. stellata GLM - *S. siderea* MARS	0.3394	0.0084	40.3058	0.0000
S. stellata GLM - *S. stellata* MARS	-0.7634	0.0199	-38.3676	0.0000
S. stellata GLM - *S. radians* Maxent	0.6965	0.0103	67.8733	0.0000
S. stellata GLM - *S. siderea* Maxent	1.4240	0.0078	181.4021	0.0000
S. stellata GLM - *S. stellata* Maxent	0.1106	0.0077	14.4273	0.0000
S. stellata GLM - *S. radians* SVM	0.3640	0.0079	45.8657	0.0000
S. stellata GLM - *S. siderea* SVM	-1.0810	0.0070	-154.5088	0.0000
S. stellata GLM - *S. stellata* SVM	-1.0827	0.0075	-144.4253	0.0000
S. stellata Gower - *S. radians* ANN	-4.3508	0.0122	-357.7667	0.0000
S. stellata Gower - *S. siderea* ANN	-4.4128	0.0093	-474.1237	0.0000
S. stellata Gower - *S. stellata* ANN	-4.1449	0.0104	-397.7205	0.0000
S. stellata Gower - *S. radians* GAM	-2.4561	0.0080	-305.7633	0.0000
S. stellata Gower - *S. siderea* GAM	-1.4418	0.0080	-180.8907	0.0000
S. stellata Gower - *S. stellata* GAM	-3.4948	0.0082	-426.6872	0.0000
S. stellata Gower - *S. radians* GLM	-1.4622	0.0119	-122.8229	0.0000
S. stellata Gower - *S. siderea* GLM	-1.9913	0.0095	-209.3664	0.0000
S. stellata Gower - *S. stellata* GLM	-2.7111	0.0089	-305.2297	0.0000
S. stellata Gower - *S. radians* Mahal	3.2423	0.0069	469.3633	0.0000
S. stellata Gower - *S. siderea* Mahal	2.9737	0.0069	431.9685	0.0000
S. stellata Gower - *S. stellata* Mahal	4.4999	0.0072	625.7944	0.0000
S. stellata Gower - *S. radians* MARS	-1.9963	0.0084	-238.0273	0.0000
S. stellata Gower - *S. siderea* MARS	-2.3717	0.0092	-256.9901	0.0000
S. stellata Gower - *S. stellata* MARS	-3.4746	0.0203	-171.5600	0.0000
S. stellata Gower - *S. radians* Maxent	-2.0147	0.0109	-184.2747	0.0000
S. stellata Gower - *S. siderea* Maxent	-1.2871	0.0087	-147.7825	0.0000
S. stellata Gower - *S. stellata* Maxent	-2.6005	0.0085	-304.2549	0.0000
S. stellata Gower - *S. radians* SVM	-2.3471	0.0088	-267.0806	0.0000
S. stellata Gower - *S. siderea* SVM	-3.7921	0.0079	-477.0435	0.0000
S. stellata Gower - *S. stellata* SVM	-3.7938	0.0084	-452.0245	0.0000
S. stellata Mahal - *S. radians* ANN	-8.8507	0.0103	-857.7967	0.0000
S. stellata Mahal - *S. siderea* ANN	-8.9127	0.0067	-1325.7090	0.0000
S. stellata Mahal - *S. stellata* ANN	-8.6448	0.0082	-1054.7040	0.0000
S. stellata Mahal - *S. radians* GAM	-6.9560	0.0048	-1447.4643	0.0000
S. stellata Mahal - *S. siderea* GAM	-5.9417	0.0047	-1263.8080	0.0000
S. stellata Mahal - *S. stellata* GAM	-7.9947	0.0051	-1578.3408	0.0000

Continued on next page

Table S2 – continued from previous page

Contrast	Estimate	SE	Z ratio	p value
S. stellata Mahal - *S. radians* GLM	-5.9621	0.0100	-595.3143	0.0000
S. stellata Mahal - *S. siderea* GLM	-6.4912	0.0070	-927.0160	0.0000
S. stellata Mahal - *S. stellata* GLM	-7.2110	0.0061	-1178.0962	0.0000
S. stellata Mahal - *S. radians* MARS	-6.4962	0.0054	-1208.1675	0.0000
S. stellata Mahal - *S. siderea* MARS	-6.8716	0.0066	-1039.0067	0.0000
S. stellata Mahal - *S. stellata* MARS	-7.9745	0.0192	-415.2778	0.0000
S. stellata Mahal - *S. radians* Maxent	-6.5146	0.0088	-737.1495	0.0000
S. stellata Mahal - *S. siderea* Maxent	-5.7871	0.0059	-986.2305	0.0000
S. stellata Mahal - *S. stellata* Maxent	-7.1004	0.0056	-1262.6290	0.0000
S. stellata Mahal - *S. radians* SVM	-6.8470	0.0060	-1144.3442	0.0000
S. stellata Mahal - *S. siderea* SVM	-8.2920	0.0047	-1777.5091	0.0000
S. stellata Mahal - *S. stellata* SVM	-8.2937	0.0054	-1539.7196	0.0000
S. stellata MARS - *S. radians* ANN	-0.8762	0.0216	-40.6369	0.0000
S. stellata MARS - *S. siderea* ANN	-0.9383	0.0201	-46.7005	0.0000
S. stellata MARS - *S. stellata* ANN	-0.6703	0.0206	-32.4910	0.0000
S. stellata MARS - *S. radians* Maxent	1.4599	0.0209	69.8706	0.0000
S. stellata MARS - *S. siderea* Maxent	2.1874	0.0198	110.3544	0.0000
S. stellata MARS - *S. stellata* Maxent	0.8741	0.0198	44.2557	0.0000
S. stellata MARS - *S. radians* SVM	1.1275	0.0199	56.7810	0.0000
S. stellata MARS - *S. siderea* SVM	-0.3176	0.0195	-16.2857	0.0000
S. stellata MARS - *S. stellata* SVM	-0.3193	0.0197	-16.2203	0.0000
S. stellata Maxent - *S. radians* SVM	0.2534	0.0076	33.5171	0.0000
S. stellata Maxent - *S. siderea* SVM	-1.1916	0.0066	-181.4975	0.0000
S. stellata Maxent - *S. stellata* SVM	-1.1934	0.0071	-168.1634	0.0000

Table S3. Post-hoc pairwise comparisons of TSS scores graphed in main text's Fig. 3b, using least-square means computed from the fitted GLMM.

Contrast	Estimate	SE	Z ratio	p value
S. radians ANN - *S. siderea* ANN	0.0092	0.0199	0.4635	1.0000
S. radians ANN - *S. stellata* ANN	-0.2931	0.0206	-14.2315	0.0000
S. radians ANN - *S. radians* Maxent	-0.7254	0.0140	-51.7463	0.0000
S. radians ANN - *S. siderea* Maxent	-0.3945	0.0141	-27.8897	0.0000
S. radians ANN - *S. stellata* Maxent	-0.4950	0.0147	-33.7191	0.0000
S. radians ANN - *S. radians* SVM	-0.6839	0.0144	-47.5188	0.0000
S. radians ANN - *S. siderea* SVM	-0.2551	0.0150	-16.9645	0.0000
S. radians ANN - *S. stellata* SVM	0.0052	0.0172	0.2999	1.0000
S. radians Bioclim - *S. radians* ANN	0.4526	0.0142	31.9630	0.0000
S. radians Bioclim - *S. siderea* ANN	0.4618	0.0147	31.4712	0.0000
S. radians Bioclim - *S. stellata* ANN	0.1595	0.0156	10.2408	0.0000
S. radians Bioclim - *S. siderea* Bioclim	0.2152	0.0063	34.1596	0.0000
S. radians Bioclim - *S. stellata* Bioclim	0.5348	0.0065	82.8035	0.0000
S. radians Bioclim - *S. radians* GAM	-0.2659	0.0042	-63.3558	0.0000
S. radians Bioclim - *S. siderea* GAM	0.0579	0.0046	12.4681	0.0000
S. radians Bioclim - *S. stellata* GAM	0.1131	0.0057	19.6824	0.0000
S. radians Bioclim - *S. radians* GLM	-0.3196	0.0037	-86.9573	0.0000
S. radians Bioclim - *S. siderea* GLM	-0.0209	0.0039	-5.3444	0.0000
S. radians Bioclim - *S. stellata* GLM	0.0169	0.0052	3.2348	0.1972
S. radians Bioclim - *S. radians* Gower	0.1388	0.0105	13.2603	0.0000

Continued on next page

Table S3 – continued from previous page

Contrast	Estimate	SE	Z ratio	p value
S. radians Bioclim - *S. siderea* Gower	0.3041	0.0058	52.3556	0.0000
S. radians Bioclim - *S. stellata* Gower	0.2251	0.0051	43.7977	0.0000
S. radians Bioclim - *S. radians* Mahal	1.3808	0.0065	211.7069	0.0000
S. radians Bioclim - *S. siderea* Mahal	1.5517	0.0050	311.2395	0.0000
S. radians Bioclim - *S. stellata* Mahal	1.5184	0.0082	185.5426	0.0000
S. radians Bioclim - *S. radians* MARS	-0.2969	0.0040	-75.1382	0.0000
S. radians Bioclim - *S. siderea* MARS	0.1049	0.0053	19.6731	0.0000
S. radians Bioclim - *S. stellata* MARS	0.0071	0.0094	0.7597	1.0000
S. radians Bioclim - *S. radians* Maxent	-0.2728	0.0039	-70.0690	0.0000
S. radians Bioclim - *S. siderea* Maxent	0.0581	0.0043	13.3995	0.0000
S. radians Bioclim - *S. stellata* Maxent	-0.0424	0.0058	-7.2532	0.0000
S. radians Bioclim - *S. radians* SVM	-0.2313	0.0051	-45.5547	0.0000
S. radians Bioclim - *S. siderea* SVM	0.1975	0.0067	29.5432	0.0000
S. radians Bioclim - *S. stellata* SVM	0.4578	0.0107	42.6186	0.0000
S. radians GAM - *S. radians* ANN	0.7185	0.0141	50.9375	0.0000
S. radians GAM - *S. siderea* ANN	0.7277	0.0146	49.7694	0.0000
S. radians GAM - *S. stellata* ANN	0.4254	0.0155	27.3954	0.0000
S. radians GAM - *S. siderea* GAM	0.3238	0.0045	72.3239	0.0000
S. radians GAM - *S. stellata* GAM	0.3790	0.0056	67.5652	0.0000
S. radians GAM - *S. radians* GLM	-0.0537	0.0035	-15.5423	0.0000
S. radians GAM - *S. siderea* GLM	0.2450	0.0037	66.1222	0.0000
S. radians GAM - *S. stellata* GLM	0.2828	0.0051	55.6314	0.0000
S. radians GAM - *S. radians* MARS	-0.0310	0.0037	-8.2714	0.0000
S. radians GAM - *S. siderea* MARS	0.3708	0.0052	71.5210	0.0000
S. radians GAM - *S. stellata* MARS	0.2730	0.0093	29.4274	0.0000
S. radians GAM - *S. radians* Maxent	-0.0069	0.0037	-1.8677	0.9857
S. radians GAM - *S. siderea* Maxent	0.3239	0.0041	78.0580	0.0000
S. radians GAM - *S. stellata* Maxent	0.2235	0.0057	39.1478	0.0000
S. radians GAM - *S. radians* SVM	0.0346	0.0049	7.0325	0.0000
S. radians GAM - *S. siderea* SVM	0.4634	0.0066	70.5436	0.0000
S. radians GAM - *S. stellata* SVM	0.7236	0.0107	67.8307	0.0000
S. radians GLM - *S. radians* ANN	0.7722	0.0140	55.3208	0.0000
S. radians GLM - *S. siderea* ANN	0.7814	0.0145	53.9651	0.0000
S. radians GLM - *S. stellata* ANN	0.4792	0.0154	31.1221	0.0000
S. radians GLM - *S. siderea* GLM	0.2987	0.0031	96.2830	0.0000
S. radians GLM - *S. stellata* GLM	0.3366	0.0047	72.1775	0.0000
S. radians GLM - *S. radians* MARS	0.0227	0.0032	7.2025	0.0000
S. radians GLM - *S. siderea* MARS	0.4245	0.0048	88.9563	0.0000
S. radians GLM - *S. stellata* MARS	0.3267	0.0091	36.0911	0.0000
S. radians GLM - *S. radians* Maxent	0.0469	0.0031	15.2023	0.0000
S. radians GLM - *S. siderea* Maxent	0.3777	0.0036	104.2658	0.0000
S. radians GLM - *S. stellata* Maxent	0.2772	0.0053	51.9391	0.0000
S. radians GLM - *S. radians* SVM	0.0884	0.0045	19.6979	0.0000
S. radians GLM - *S. siderea* SVM	0.5172	0.0062	82.7551	0.0000
S. radians GLM - *S. stellata* SVM	0.7774	0.0105	74.2177	0.0000
S. radians Gower - *S. radians* ANN	0.3138	0.0171	18.4027	0.0000
S. radians Gower - *S. siderea* ANN	0.3231	0.0175	18.4785	0.0000
S. radians Gower - *S. stellata* ANN	0.0208	0.0182	1.1388	1.0000
S. radians Gower - *S. radians* GAM	-0.4046	0.0104	-38.9463	0.0000
S. radians Gower - *S. siderea* GAM	-0.0808	0.0106	-7.6381	0.0000
S. radians Gower - *S. stellata* GAM	-0.0257	0.0111	-2.3112	0.8546

Continued on next page

Table S3 – continued from previous page

Contrast	Estimate	SE	Z ratio	p value
S. radians Gower - *S. radians* GLM	-0.4584	0.0102	-44.9818	0.0000
S. radians Gower - *S. siderea* GLM	-0.1596	0.0103	-15.5348	0.0000
S. radians Gower - *S. stellata* GLM	-0.1218	0.0109	-11.2279	0.0000
S. radians Gower - *S. siderea* Gower	0.1653	0.0111	14.8405	0.0000
S. radians Gower - *S. stellata* Gower	0.0863	0.0108	7.9879	0.0000
S. radians Gower - *S. radians* Mahal	1.2421	0.0115	107.7521	0.0000
S. radians Gower - *S. siderea* Mahal	1.4129	0.0107	131.6491	0.0000
S. radians Gower - *S. stellata* Mahal	1.3796	0.0125	109.9999	0.0000
S. radians Gower - *S. radians* MARS	-0.4356	0.0103	-42.3250	0.0000
S. radians Gower - *S. siderea* MARS	-0.0339	0.0109	-3.1075	0.2686
S. radians Gower - *S. stellata* MARS	-0.1316	0.0133	-9.8689	0.0000
S. radians Gower - *S. radians* Maxent	-0.4115	0.0103	-40.0679	0.0000
S. radians Gower - *S. siderea* Maxent	-0.0807	0.0104	-7.7255	0.0000
S. radians Gower - *S. stellata* Maxent	-0.1811	0.0112	-16.2356	0.0000
S. radians Gower - *S. radians* SVM	-0.3700	0.0108	-34.3405	0.0000
S. radians Gower - *S. siderea* SVM	0.0588	0.0116	5.0582	0.0001
S. radians Gower - *S. stellata* SVM	0.3190	0.0143	22.2424	0.0000
S. radians Mahal - *S. radians* ANN	-0.9282	0.0150	-62.0361	0.0000
S. radians Mahal - *S. siderea* ANN	-0.9190	0.0155	-59.4799	0.0000
S. radians Mahal - *S. stellata* ANN	-1.2213	0.0163	-74.8715	0.0000
S. radians Mahal - *S. radians* GAM	-1.6467	0.0064	-257.2089	0.0000
S. radians Mahal - *S. siderea* GAM	-1.3229	0.0067	-197.2542	0.0000
S. radians Mahal - *S. stellata* GAM	-1.2677	0.0075	-168.8224	0.0000
S. radians Mahal - *S. radians* GLM	-1.7004	0.0061	-279.9800	0.0000
S. radians Mahal - *S. siderea* GLM	-1.4017	0.0062	-225.4489	0.0000
S. radians Mahal - *S. stellata* GLM	-1.3639	0.0071	-191.4101	0.0000
S. radians Mahal - *S. siderea* Mahal	0.1709	0.0069	24.6032	0.0000
S. radians Mahal - *S. stellata* Mahal	0.1375	0.0095	14.4713	0.0000
S. radians Mahal - *S. radians* MARS	-1.6777	0.0062	-268.6863	0.0000
S. radians Mahal - *S. siderea* MARS	-1.2759	0.0072	-177.2744	0.0000
S. radians Mahal - *S. stellata* MARS	-1.3737	0.0105	-130.3945	0.0000
S. radians Mahal - *S. radians* Maxent	-1.6536	0.0062	-266.3934	0.0000
S. radians Mahal - *S. siderea* Maxent	-1.3228	0.0065	-203.7399	0.0000
S. radians Mahal - *S. stellata* Maxent	-1.4232	0.0076	-187.6506	0.0000
S. radians Mahal - *S. radians* SVM	-1.6121	0.0070	-229.9473	0.0000
S. radians Mahal - *S. siderea* SVM	-1.1833	0.0083	-143.4066	0.0000
S. radians Mahal - *S. stellata* SVM	-0.9231	0.0118	-78.3655	0.0000
S. radians MARS - *S. radians* ANN	0.7495	0.0140	53.4052	0.0000
S. radians MARS - *S. siderea* ANN	0.7587	0.0146	52.1348	0.0000
S. radians MARS - *S. stellata* ANN	0.4564	0.0155	29.5152	0.0000
S. radians MARS - *S. siderea* MARS	0.4018	0.0050	80.5570	0.0000
S. radians MARS - *S. stellata* MARS	0.3040	0.0092	33.1579	0.0000
S. radians MARS - *S. radians* Maxent	0.0241	0.0034	7.0832	0.0000
S. radians MARS - *S. siderea* Maxent	0.3550	0.0039	90.9730	0.0000
S. radians MARS - *S. stellata* Maxent	0.2545	0.0055	46.0135	0.0000
S. radians MARS - *S. radians* SVM	0.0656	0.0047	13.9213	0.0000
S. radians MARS - *S. siderea* SVM	0.4944	0.0064	77.0707	0.0000
S. radians MARS - *S. stellata* SVM	0.7547	0.0106	71.3672	0.0000
S. radians Maxent - *S. siderea* Maxent	0.3308	0.0038	86.0961	0.0000
S. radians Maxent - *S. stellata* Maxent	0.2304	0.0055	41.9673	0.0000

Continued on next page

Table S3 – continued from previous page

Contrast	Estimate	SE	Z ratio	p value
S. radians Maxent - *S. radians* SVM	0.0415	0.0047	8.8954	0.0000
S. radians Maxent - *S. siderea* SVM	0.4703	0.0064	73.7218	0.0000
S. radians Maxent - *S. stellata* SVM	0.7305	0.0106	69.2277	0.0000
S. radians SVM - *S. siderea* SVM	0.4288	0.0072	59.8595	0.0000
S. radians SVM - *S. stellata* SVM	0.6890	0.0110	62.3882	0.0000
S. siderea ANN - *S. stellata* ANN	-0.3023	0.0209	-14.4969	0.0000
S. siderea ANN - *S. radians* Maxent	-0.7346	0.0145	-50.5317	0.0000
S. siderea ANN - *S. siderea* Maxent	-0.4038	0.0144	-28.0769	0.0000
S. siderea ANN - *S. stellata* Maxent	-0.5042	0.0150	-33.5210	0.0000
S. siderea ANN - *S. radians* SVM	-0.6931	0.0149	-46.5226	0.0000
S. siderea ANN - *S. siderea* SVM	-0.2643	0.0153	-17.3240	0.0000
S. siderea ANN - *S. stellata* SVM	-0.0041	0.0175	-0.2318	1.0000
S. siderea Bioclim - *S. radians* ANN	0.2374	0.0149	15.9716	0.0000
S. siderea Bioclim - *S. siderea* ANN	0.2467	0.0151	16.3470	0.0000
S. siderea Bioclim - *S. stellata* ANN	-0.0556	0.0161	-3.4549	0.1074
S. siderea Bioclim - *S. stellata* Bioclim	0.3197	0.0076	41.9191	0.0000
S. siderea Bioclim - *S. radians* GAM	-0.4810	0.0062	-77.9128	0.0000
S. siderea Bioclim - *S. siderea* GAM	-0.1572	0.0058	-26.9772	0.0000
S. siderea Bioclim - *S. stellata* GAM	-0.1021	0.0070	-14.5149	0.0000
S. siderea Bioclim - *S. radians* GLM	-0.5348	0.0058	-91.6904	0.0000
S. siderea Bioclim - *S. siderea* GLM	-0.2360	0.0053	-44.8990	0.0000
S. siderea Bioclim - *S. stellata* GLM	-0.1982	0.0066	-29.9415	0.0000
S. siderea Bioclim - *S. radians* Gower	-0.0764	0.0114	-6.7004	0.0000
S. siderea Bioclim - *S. siderea* Gower	0.0889	0.0068	13.0954	0.0000
S. siderea Bioclim - *S. stellata* Gower	0.0099	0.0065	1.5139	0.9994
S. siderea Bioclim - *S. radians* Mahal	1.1657	0.0079	146.8056	0.0000
S. siderea Bioclim - *S. siderea* Mahal	1.3365	0.0061	219.0929	0.0000
S. siderea Bioclim - *S. stellata* Mahal	1.3032	0.0091	142.7028	0.0000
S. siderea Bioclim - *S. radians* MARS	-0.5120	0.0060	-85.1999	0.0000
S. siderea Bioclim - *S. siderea* MARS	-0.1103	0.0064	-17.2651	0.0000
S. siderea Bioclim - *S. stellata* MARS	-0.2080	0.0102	-20.3961	0.0000
S. siderea Bioclim - *S. radians* Maxent	-0.4879	0.0060	-81.7058	0.0000
S. siderea Bioclim - *S. siderea* Maxent	-0.1571	0.0056	-28.1542	0.0000
S. siderea Bioclim - *S. stellata* Maxent	-0.2575	0.0071	-36.2125	0.0000
S. siderea Bioclim - *S. radians* SVM	-0.4464	0.0068	-65.6221	0.0000
S. siderea Bioclim - *S. siderea* SVM	-0.0176	0.0076	-2.3322	0.8430
S. siderea Bioclim - *S. stellata* SVM	0.2426	0.0115	21.1324	0.0000
S. siderea GAM - *S. radians* ANN	0.3946	0.0142	27.7029	0.0000
S. siderea GAM - *S. siderea* ANN	0.4039	0.0145	27.8949	0.0000
S. siderea GAM - *S. stellata* ANN	0.1016	0.0155	6.5433	0.0000
S. siderea GAM - *S. stellata* GAM	0.0551	0.0056	9.8445	0.0000
S. siderea GAM - *S. radians* GLM	-0.3776	0.0040	-94.5512	0.0000
S. siderea GAM - *S. siderea* GLM	-0.0788	0.0031	-25.4875	0.0000
S. siderea GAM - *S. stellata* GLM	-0.0410	0.0051	-8.0815	0.0000
S. siderea GAM - *S. radians* MARS	-0.3548	0.0042	-83.5235	0.0000
S. siderea GAM - *S. siderea* MARS	0.0469	0.0048	9.8500	0.0000
S. siderea GAM - *S. stellata* MARS	-0.0508	0.0093	-5.4827	0.0000
S. siderea GAM - *S. radians* Maxent	-0.3307	0.0042	-78.8527	0.0000
S. siderea GAM - *S. siderea* Maxent	0.0001	0.0036	0.0314	1.0000
S. siderea GAM - *S. stellata* Maxent	-0.1003	0.0057	-17.5971	0.0000

Continued on next page

Table S3 – continued from previous page

Contrast	Estimate	SE	Z ratio	p value
S. siderea GAM - *S. radians* SVM	-0.2892	0.0053	-54.4536	0.0000
S. siderea GAM - *S. siderea* SVM	0.1396	0.0062	22.3536	0.0000
S. siderea GAM - *S. stellata* SVM	0.3998	0.0107	37.4908	0.0000
S. siderea GLM - *S. radians* ANN	0.4735	0.0140	33.7670	0.0000
S. siderea GLM - *S. siderea* ANN	0.4827	0.0143	33.8543	0.0000
S. siderea GLM - *S. stellata* ANN	0.1804	0.0153	11.7765	0.0000
S. siderea GLM - *S. stellata* GLM	0.0378	0.0044	8.5788	0.0000
S. siderea GLM - *S. radians* MARS	-0.2760	0.0034	-80.5906	0.0000
S. siderea GLM - *S. siderea* MARS	0.1258	0.0040	31.0645	0.0000
S. siderea GLM - *S. stellata* MARS	0.0280	0.0089	3.1376	0.2504
S. siderea GLM - *S. radians* Maxent	-0.2519	0.0034	-75.0271	0.0000
S. siderea GLM - *S. siderea* Maxent	0.0790	0.0026	30.4048	0.0000
S. siderea GLM - *S. stellata* Maxent	-0.0215	0.0051	-4.1998	0.0076
S. siderea GLM - *S. radians* SVM	-0.2104	0.0047	-44.9659	0.0000
S. siderea GLM - *S. siderea* SVM	0.2184	0.0057	38.2120	0.0000
S. siderea GLM - *S. stellata* SVM	0.4787	0.0104	46.1851	0.0000
S. siderea Gower - *S. radians* ANN	0.1485	0.0147	10.1289	0.0000
S. siderea Gower - *S. siderea* ANN	0.1578	0.0149	10.5949	0.0000
S. siderea Gower - *S. stellata* ANN	-0.1445	0.0159	-9.0822	0.0000
S. siderea Gower - *S. radians* GAM	-0.5699	0.0057	-100.4781	0.0000
S. siderea Gower - *S. siderea* GAM	-0.2461	0.0053	-46.4978	0.0000
S. siderea Gower - *S. stellata* GAM	-0.1910	0.0066	-28.9536	0.0000
S. siderea Gower - *S. radians* GLM	-0.6237	0.0053	-117.7110	0.0000
S. siderea Gower - *S. siderea* GLM	-0.3249	0.0047	-69.7646	0.0000
S. siderea Gower - *S. stellata* GLM	-0.2871	0.0062	-46.6482	0.0000
S. siderea Gower - *S. stellata* Gower	-0.0790	0.0061	-13.0046	0.0000
S. siderea Gower - *S. radians* Mahal	1.0768	0.0076	142.4933	0.0000
S. siderea Gower - *S. siderea* Mahal	1.2476	0.0056	223.1144	0.0000
S. siderea Gower - *S. stellata* Mahal	1.2143	0.0088	137.9763	0.0000
S. siderea Gower - *S. radians* MARS	-0.6009	0.0055	-109.3984	0.0000
S. siderea Gower - *S. siderea* MARS	-0.1992	0.0059	-33.7407	0.0000
S. siderea Gower - *S. stellata* MARS	-0.2969	0.0099	-29.9806	0.0000
S. siderea Gower - *S. radians* Maxent	-0.5768	0.0055	-105.8134	0.0000
S. siderea Gower - *S. siderea* Maxent	-0.2460	0.0050	-49.0134	0.0000
S. siderea Gower - *S. stellata* Maxent	-0.3464	0.0067	-51.8550	0.0000
S. siderea Gower - *S. radians* SVM	-0.5353	0.0064	-84.2890	0.0000
S. siderea Gower - *S. siderea* SVM	-0.1065	0.0071	-14.8976	0.0000
S. siderea Gower - *S. stellata* SVM	0.1537	0.0112	13.7013	0.0000
S. siderea Mahal - *S. radians* ANN	-1.0991	0.0144	-76.5410	0.0000
S. siderea Mahal - *S. siderea* ANN	-1.0899	0.0146	-74.6962	0.0000
S. siderea Mahal - *S. stellata* ANN	-1.3921	0.0156	-89.0661	0.0000
S. siderea Mahal - *S. radians* GAM	-1.8176	0.0048	-376.5195	0.0000
S. siderea Mahal - *S. siderea* GAM	-1.4937	0.0044	-341.4033	0.0000
S. siderea Mahal - *S. stellata* GAM	-1.4386	0.0059	-244.4639	0.0000
S. siderea Mahal - *S. radians* GLM	-1.8713	0.0044	-427.0508	0.0000
S. siderea Mahal - *S. siderea* GLM	-1.5726	0.0036	-439.1504	0.0000
S. siderea Mahal - *S. stellata* GLM	-1.5348	0.0054	-284.9320	0.0000
S. siderea Mahal - *S. stellata* Mahal	-0.0333	0.0083	-4.0230	0.0153
S. siderea Mahal - *S. radians* MARS	-1.8486	0.0046	-400.5059	0.0000
S. siderea Mahal - *S. siderea* MARS	-1.4468	0.0051	-283.9009	0.0000

Continued on next page

Table S3 – continued from previous page

Contrast	Estimate	SE	Z ratio	p value
S. siderea Mahal - *S. stellata* MARS	-1.5446	0.0094	-163.5157	0.0000
S. siderea Mahal - *S. radians* Maxent	-1.8245	0.0046	-399.6022	0.0000
S. siderea Mahal - *S. siderea* Maxent	-1.4936	0.0040	-369.7498	0.0000
S. siderea Mahal - *S. stellata* Maxent	-1.5941	0.0060	-266.5579	0.0000
S. siderea Mahal - *S. radians* SVM	-1.7830	0.0056	-317.8560	0.0000
S. siderea Mahal - *S. siderea* SVM	-1.3541	0.0065	-208.3307	0.0000
S. siderea Mahal - *S. stellata* SVM	-1.0939	0.0108	-101.1408	0.0000
S. siderea MARS - *S. radians* ANN	0.3477	0.0145	24.0070	0.0000
S. siderea MARS - *S. siderea* ANN	0.3569	0.0147	24.2606	0.0000
S. siderea MARS - *S. stellata* ANN	0.0546	0.0157	3.4708	0.1025
S. siderea MARS - *S. stellata* MARS	-0.0978	0.0096	-10.1504	0.0000
S. siderea MARS - *S. radians* Maxent	-0.3777	0.0049	-76.4274	0.0000
S. siderea MARS - *S. siderea* Maxent	-0.0468	0.0045	-10.5008	0.0000
S. siderea MARS - *S. stellata* Maxent	-0.1473	0.0063	-23.4824	0.0000
S. siderea MARS - *S. radians* SVM	-0.3362	0.0059	-56.7920	0.0000
S. siderea MARS - *S. siderea* SVM	0.0926	0.0068	13.6861	0.0000
S. siderea MARS - *S. stellata* SVM	0.3529	0.0110	32.1381	0.0000
S. siderea Maxent - *S. stellata* Maxent	-0.1004	0.0054	-18.4363	0.0000
S. siderea Maxent - *S. radians* SVM	-0.2893	0.0050	-57.4264	0.0000
S. siderea Maxent - *S. siderea* SVM	0.1395	0.0060	23.1914	0.0000
S. siderea Maxent - *S. stellata* SVM	0.3997	0.0105	37.9545	0.0000
S. siderea SVM - *S. stellata* SVM	0.2602	0.0117	22.2459	0.0000
S. stellata ANN - *S. radians* Maxent	-0.4323	0.0154	-27.9819	0.0000
S. stellata ANN - *S. siderea* Maxent	-0.1015	0.0154	-6.5746	0.0000
S. stellata ANN - *S. stellata* Maxent	-0.2019	0.0158	-12.7827	0.0000
S. stellata ANN - *S. radians* SVM	-0.3908	0.0158	-24.7513	0.0000
S. stellata ANN - *S. siderea* SVM	0.0380	0.0163	2.3379	0.8397
S. stellata ANN - *S. stellata* SVM	0.2982	0.0182	16.3980	0.0000
S. stellata Bioclim - *S. radians* ANN	-0.0822	0.0149	-5.5046	0.0000
S. stellata Bioclim - *S. siderea* ANN	-0.0730	0.0153	-4.7727	0.0006
S. stellata Bioclim - *S. stellata* ANN	-0.3753	0.0160	-23.4041	0.0000
S. stellata Bioclim - *S. radians* GAM	-0.8007	0.0063	-126.3431	0.0000
S. stellata Bioclim - *S. siderea* GAM	-0.4769	0.0063	-75.3255	0.0000
S. stellata Bioclim - *S. stellata* GAM	-0.4217	0.0069	-61.2437	0.0000
S. stellata Bioclim - *S. radians* GLM	-0.8544	0.0060	-142.2837	0.0000
S. stellata Bioclim - *S. siderea* GLM	-0.5557	0.0058	-95.6435	0.0000
S. stellata Bioclim - *S. stellata* GLM	-0.5179	0.0065	-80.1017	0.0000
S. stellata Bioclim - *S. radians* Gower	-0.3960	0.0115	-34.4655	0.0000
S. stellata Bioclim - *S. siderea* Gower	-0.2308	0.0072	-31.9374	0.0000
S. stellata Bioclim - *S. stellata* Gower	-0.3097	0.0064	-48.4870	0.0000
S. stellata Bioclim - *S. radians* Mahal	0.8460	0.0081	104.8615	0.0000
S. stellata Bioclim - *S. siderea* Mahal	1.0169	0.0066	154.4776	0.0000
S. stellata Bioclim - *S. stellata* Mahal	0.9836	0.0090	109.0354	0.0000
S. stellata Bioclim - *S. radians* MARS	-0.8317	0.0062	-134.6297	0.0000
S. stellata Bioclim - *S. siderea* MARS	-0.4299	0.0068	-62.7733	0.0000
S. stellata Bioclim - *S. stellata* MARS	-0.5277	0.0101	-52.2463	0.0000
S. stellata Bioclim - *S. radians* Maxent	-0.8076	0.0061	-131.5167	0.0000
S. stellata Bioclim - *S. siderea* Maxent	-0.4767	0.0061	-78.1110	0.0000
S. stellata Bioclim - *S. stellata* Maxent	-0.5772	0.0070	-82.8382	0.0000
S. stellata Bioclim - *S. radians* SVM	-0.7661	0.0070	-110.2018	0.0000

Continued on next page

Table S3 – continued from previous page

Contrast	Estimate	SE	Z ratio	p value
S. stellata Bioclim - *S. siderea* SVM	-0.3373	0.0079	-42.4301	0.0000
S. stellata Bioclim - *S. stellata* SVM	-0.0770	0.0114	-6.7633	0.0000
S. stellata GAM - *S. radians* ANN	0.3395	0.0146	23.1890	0.0000
S. stellata GAM - *S. siderea* ANN	0.3487	0.0150	23.2431	0.0000
S. stellata GAM - *S. stellata* ANN	0.0464	0.0158	2.9472	0.3781
S. stellata GAM - *S. radians* GLM	-0.4327	0.0052	-82.7274	0.0000
S. stellata GAM - *S. siderea* GLM	-0.1340	0.0050	-26.7667	0.0000
S. stellata GAM - *S. stellata* GLM	-0.0962	0.0058	-16.7145	0.0000
S. stellata GAM - *S. radians* MARS	-0.4100	0.0054	-75.5334	0.0000
S. stellata GAM - *S. siderea* MARS	-0.0082	0.0062	-1.3261	0.9999
S. stellata GAM - *S. stellata* MARS	-0.1060	0.0097	-10.9711	0.0000
S. stellata GAM - *S. radians* Maxent	-0.3859	0.0054	-71.6482	0.0000
S. stellata GAM - *S. siderea* Maxent	-0.0550	0.0053	-10.2988	0.0000
S. stellata GAM - *S. stellata* Maxent	-0.1555	0.0063	-24.6299	0.0000
S. stellata GAM - *S. radians* SVM	-0.3444	0.0063	-54.7070	0.0000
S. stellata GAM - *S. siderea* SVM	0.0844	0.0074	11.4409	0.0000
S. stellata GAM - *S. stellata* SVM	0.3447	0.0110	31.3252	0.0000
S. stellata GLM - *S. radians* ANN	0.4357	0.0144	30.1545	0.0000
S. stellata GLM - *S. siderea* ANN	0.4449	0.0148	30.0289	0.0000
S. stellata GLM - *S. stellata* ANN	0.1426	0.0156	9.1524	0.0000
S. stellata GLM - *S. radians* MARS	-0.3138	0.0049	-64.2674	0.0000
S. stellata GLM - *S. siderea* MARS	0.0880	0.0057	15.4091	0.0000
S. stellata GLM - *S. stellata* MARS	-0.0098	0.0094	-1.0487	1.0000
S. stellata GLM - *S. radians* Maxent	-0.2897	0.0048	-59.9054	0.0000
S. stellata GLM - *S. siderea* Maxent	0.0411	0.0048	8.5885	0.0000
S. stellata GLM - *S. stellata* Maxent	-0.0593	0.0059	-10.1383	0.0000
S. stellata GLM - *S. radians* SVM	-0.2482	0.0058	-42.5628	0.0000
S. stellata GLM - *S. siderea* SVM	0.1806	0.0070	25.8363	0.0000
S. stellata GLM - *S. stellata* SVM	0.4408	0.0107	41.0275	0.0000
S. stellata Gower - *S. radians* ANN	0.2275	0.0144	15.7861	0.0000
S. stellata Gower - *S. siderea* ANN	0.2368	0.0148	16.0169	0.0000
S. stellata Gower - *S. stellata* ANN	-0.0655	0.0155	-4.2139	0.0072
S. stellata Gower - *S. radians* GAM	-0.4909	0.0050	-98.4780	0.0000
S. stellata Gower - *S. siderea* GAM	-0.1671	0.0050	-33.5794	0.0000
S. stellata Gower - *S. stellata* GAM	-0.1120	0.0057	-19.7616	0.0000
S. stellata Gower - *S. radians* GLM	-0.5447	0.0046	-119.5683	0.0000
S. stellata Gower - *S. siderea* GLM	-0.2460	0.0043	-57.2638	0.0000
S. stellata Gower - *S. stellata* GLM	-0.2081	0.0051	-40.4389	0.0000
S. stellata Gower - *S. radians* Mahal	1.1558	0.0071	163.8062	0.0000
S. stellata Gower - *S. siderea* Mahal	1.3266	0.0053	250.6063	0.0000
S. stellata Gower - *S. stellata* Mahal	1.2933	0.0081	159.1207	0.0000
S. stellata Gower - *S. radians* MARS	-0.5220	0.0048	-109.1813	0.0000
S. stellata Gower - *S. siderea* MARS	-0.1202	0.0056	-21.3790	0.0000
S. stellata Gower - *S. stellata* MARS	-0.2180	0.0093	-23.4068	0.0000
S. stellata Gower - *S. radians* Maxent	-0.4978	0.0047	-105.1953	0.0000
S. stellata Gower - *S. siderea* Maxent	-0.1670	0.0047	-35.6514	0.0000
S. stellata Gower - *S. stellata* Maxent	-0.2674	0.0058	-46.3888	0.0000
S. stellata Gower - *S. radians* SVM	-0.4563	0.0057	-79.4192	0.0000
S. stellata Gower - *S. siderea* SVM	-0.0275	0.0069	-3.9783	0.0182
S. stellata Gower - *S. stellata* SVM	0.2327	0.0107	21.7504	0.0000
S. stellata Mahal - *S. radians* ANN	-1.0658	0.0158	-67.6347	0.0000

Continued on next page

Table S3 – continued from previous page

Contrast	Estimate	SE	Z ratio	p value
S. stellata Mahal - *S. siderea* ANN	-1.0565	0.0161	-65.6414	0.0000
S. stellata Mahal - *S. stellata* ANN	-1.3588	0.0168	-80.8668	0.0000
S. stellata Mahal - *S. radians* GAM	-1.7842	0.0081	-220.6050	0.0000
S. stellata Mahal - *S. siderea* GAM	-1.4604	0.0081	-180.6853	0.0000
S. stellata Mahal - *S. stellata* GAM	-1.4053	0.0085	-164.8502	0.0000
S. stellata Mahal - *S. radians* GLM	-1.8380	0.0078	-234.7270	0.0000
S. stellata Mahal - *S. siderea* GLM	-1.5393	0.0077	-200.3779	0.0000
S. stellata Mahal - *S. stellata* GLM	-1.5014	0.0082	-183.3586	0.0000
S. stellata Mahal - *S. radians* MARS	-1.8153	0.0080	-227.9491	0.0000
S. stellata Mahal - *S. siderea* MARS	-1.4135	0.0085	-166.3984	0.0000
S. stellata Mahal - *S. stellata* MARS	-1.5113	0.0113	-133.9620	0.0000
S. stellata Mahal - *S. radians* Maxent	-1.7911	0.0079	-225.7369	0.0000
S. stellata Mahal - *S. siderea* Maxent	-1.4603	0.0079	-184.7092	0.0000
S. stellata Mahal - *S. stellata* Maxent	-1.5608	0.0086	-181.6774	0.0000
S. stellata Mahal - *S. radians* SVM	-1.7496	0.0086	-203.9749	0.0000
S. stellata Mahal - *S. siderea* SVM	-1.3208	0.0094	-140.4532	0.0000
S. stellata Mahal - *S. stellata* SVM	-1.0606	0.0125	-85.1841	0.0000
S. stellata MARS - *S. radians* ANN	0.4455	0.0164	27.1641	0.0000
S. stellata MARS - *S. siderea* ANN	0.4547	0.0167	27.1883	0.0000
S. stellata MARS - *S. stellata* ANN	0.1524	0.0174	8.7568	0.0000
S. stellata MARS - *S. radians* Maxent	-0.2799	0.0091	-30.6101	0.0000
S. stellata MARS - *S. siderea* Maxent	0.0509	0.0091	5.5875	0.0000
S. stellata MARS - *S. stellata* Maxent	-0.0495	0.0097	-5.0930	0.0001
S. stellata MARS - *S. radians* SVM	-0.2384	0.0097	-24.5584	0.0000
S. stellata MARS - *S. siderea* SVM	0.1904	0.0104	18.2328	0.0000
S. stellata MARS - *S. stellata* SVM	0.4506	0.0133	34.0015	0.0000
S. stellata Maxent - *S. radians* SVM	-0.1889	0.0064	-29.5877	0.0000
S. stellata Maxent - *S. siderea* SVM	0.2399	0.0075	32.1715	0.0000
S. stellata Maxent - *S. stellata* SVM	0.5001	0.0111	45.2440	0.0000

References

1. S Lek, et al., Application of neural networks to modelling nonlinear relationships in ecology. *Ecol. Model.* **90**, 39–52 (1996).
2. HA Chipman, EI George, RE McCulloch, Bart: Bayesian additive regression trees. *Annals Appl. Stat.* **4**, 266–298 (2010).
3. HA Nix, A biogeographic analysis of australian elapid snakes in *Atlas of Elapid Snakes of Australia.* (Australian Government Publishing Service), p. 4–15 (1986).
4. J Elith, JR Leathwick, T Hastie, A working guide to boosted regression trees. *J. Animal Ecol.* **77**, 802–813 (2008).
5. L Breiman, JH Friedman, RA Olshen, CJ Stone, *Classification and Regression Trees.* (Wadsworth and Brooks), (1984).
6. A Linden, Cta: Stata module for conducting classification tree analysis, (Boston College Department of Economics), Technical report (2020).
7. G Carpenter, AN Gillison, J Winter, Domain: A flexible modeling procedure for mapping potential distributions of plants and animals. *Biodivers. & Conserv.* **2**, 667–680 (1993).
8. AH Hirzel, J Hausser, D Chessel, N Perrin, Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecol. Model.* **157**, 83–90 (2002).
9. A Guisan, TC Edwards, T Hastie, Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecol. Model.* **157**, 89–100 (2002).
10. KY Liang, SL Zeger, Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22 (1986).
11. A Lehmann, JM Overton, JR Leathwick, Grasp: generalized regression analysis and spatial prediction. *Ecol. Model.* **160**, 165–183 (2003).
12. DR Stockwell, D Peters, The garp modelling system: problems and solutions to automated spatial prediction. *Int. J. Geogr. Inf. Sci.* **13**, 143–158 (1999).
13. JH Friedman, Greedy function approximation: a gradient boosting machine. *The Annals Stat.* **29**, 1189–1232 (2001).
14. U.S. Fish and Wildlife Service, Habitat suitability index models, (U.S. Department of the Interior), Technical report (1981).

15. O Farber, R Kadmon, Assessment of alternative approaches for bioclimatic modeling with special emphasis on the mahalanobis distance. *Ecol. Model.* **160**, 115–130 (2003).
16. SJ Phillips, RP Anderson, RE Schapire, Maximum entropy modeling of species geographic distributions. *Ecol. Model.* **190**, 231–259 (2006).
17. GJ McLachlan, KE Basford, *Mixture Models: Inference and Applications to Clustering.* (Dekker, New York), pp. xii + 254 (1988).
18. JH Friedman, Multivariate adaptive regression splines. *Annals Stat.* **19**, 1–67 (1991).
19. L Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).
20. JH Friedman, Regularized discriminant analysis. *J. Am. Stat. Assoc.* **84**, 165–175 (1989).
21. J Friedman, T Hastie, R Tibshirani, Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
22. JM Drake, C Randin, A Guisan, Modelling ecological niches with support vector machines. *J. Appl. Ecol.* **43**, 424–432 (2006).