

Broad-scale factors shaping the ecological niche and geographic distribution of *Spirodela polyrhiza*

Marlon E. Cobos^{1*} and A. Townsend Peterson¹

¹Department of Ecology and Evolutionary Biology & Biodiversity Institute, University of Kansas,
Lawrence, Kansas, United States.

* Corresponding Author

E-mail: manubio13@gmail.com

Author Contributions

MEC and ATP conceived the idea of the project. MEC ran analyses. MEC wrote the manuscript.

MEC and ATP reviewed and improved the manuscript.

Abstract

The choice of appropriate independent variables to create models characterizing ecological niches of species is of critical importance in distributional ecology. This set of dimensions in which a niche is defined can inform about what factors limit the distributional potential of a species. We used a multistep approach to select relevant variables for modeling the ecological niche of the aquatic *Spirodela polyrriza*, taking into account variability arising from using distinct algorithms, calibration areas, and spatial resolutions of variables. We found that, even after an initial selection of meaningful variables, the final set of variables selected based on statistical inference varied considerably depending on the combination of algorithm, calibration area, and spatial resolution used. However, variables representing extreme temperatures and dry periods were more consistently selected than others, highlighting their importance in shaping the distribution of this species. Other variables related to seasonality of solar radiation, summer solar radiation, and some soil proxies of nutrients in water, were selected commonly but not as frequently as the ones mentioned above. We suggest that these later variables are also important to understanding the distributional potential of the species, but that their effects may be less pronounced at the scale at which they are represented for the needs of this type of modeling.

Introduction

Ecological niche modeling (ENM) includes a diverse set of tools used to study potential distributions of species via characterizations of their environmental requirements [1,2]. In particular, correlational ENMs use distributional data (occurrence records) and sets of environmental variables to calibrate models that are used to predict environmental suitability for a species across a region of interest [3]. Variables appropriate to characterizing and understanding a species' niche are those that allow identifying conditions that are favorable for the species, as well as detecting potential limits of what is or is not suitable for a species (e.g., temperatures that allow maximum growth rates, or maximum temperatures that can be tolerated). However, differences in the scale at which ecological processes occur and the grain and extent to which environmental variables are measured make it difficult to select predictors based on direct interpretations of their biological importance [4].

The challenge of selecting appropriate environmental variables when characterizing species' ecological niches using correlative models is well-known in the field of distributional ecology [5–7]. In general, models can be constructed with two main goals: (1) to improve predictions of the geographic distribution of the species, and (2) to understand which environmental variables are important constraints on species' niches. When the goal is to improve model predictive ability, variables can be selected based on how they improve predictions of independent testing data. In these cases, environmental data sets that efficiently summarize environmental variation across an area of interest (e.g., principal components) are commonly considered to be good choices [8]. However, when models are constructed to understand which and how environmental variables shape species' ecological niches and geographic distributions, use of biologically meaningful and interpretable variables becomes more relevant [4].

Common procedures for selecting environmental predictors in ENM include reducing multicollinearity, testing contribution of variables to models, selecting variables based on their

biological importance considering empirical evidence or the experience of researchers, or using a broad set of variables and letting the algorithm select important variables [9]. Other alternatives include selecting predictors by transforming original variables to summarize the variance explained by a set of principal components that are more information-rich and in general, are not correlated [10,11]. However, interpretation of the role of particular environmental factors in the characterization of species niches is complicated. More recently, different sets of variables have been used as part of the parameterizations to be tested during the process of model calibration [12]. After model selection, one or more sets of variables can be identified as more appropriate and powerful for use in creating final models.

The reality is that, regardless of the method used to determine the set of variables for modeling ecological niches, the decision is always difficult and the answer is rarely unique or unambiguous [9]. This complication exists because every step taken to define which variables are best may result in distinct sets of predictors at the end. For instance, when selecting one variable depending on correlation values and biological importance, the decision of which variable to keep depends on the researcher; many times, such an initial decision determines which other variables can or cannot be considered. Analyses like the variance inflation factor may end up identifying unique sets of variables; however, the set of variables selected depends on a predefined limit, which is not a biological consideration [13]. Distinct answers are obtained when different sets of variables are considered in model calibration, although such sets are usually subject to *a priori* processes of selection [14].

Implications of using one set or another set of environmental dimensions when creating models are not negligible, especially in applications in which model transfers to other geographic or temporal scenarios are required [15–17]. Other less explored complications are the effects that areas for model calibration, spatial resolution of raster layers used as predictors, and use of different modeling algorithms have on the variables that get selected. The area across which a model is calibrated has direct implications on predictions: models may be over- or under-fitted if

such an area is poorly defined [18,19]. Little is known about how changes in calibration areas affect decisions related to variable selection during the modeling process. The spatial resolution of variables is known to affect model calibration and model transfers [6,20]; however, little has been said about its effects on the final set of variables selected (but see [21]). Distinct modeling algorithms may also perform better or worse depending on the sets of variables used, as all predictors influence the model and interact with other variables differently. Again, however, this factor has not been considered deeply (but see [22]), and the set of variables for modeling is usually fixed when using multiple algorithms (e.g., [23,24]).

Here we explore the challenges in defining sets of environmental variables in ENM for *Spirodela polyrhiza* (greater duckweed), a freshwater plant species with a broad near cosmopolitan distribution [25]. Specifically, we used distinct methods for variable selection in a multi-step approach. We performed analyses at two spatial resolutions, used two algorithms for model calibration, and considered different options regarding calibration areas, to explore the consequences of these factors on variable selection. As little is known about macroecological factors driving the geographic distribution of greater duckweed, explorations of environmental variables can help to understand the distributional potential of this plant, environmental dimensions limiting its potential for expansion to other areas, and how climate change might affect this species' range.

Methods

Study organism

Spirodela polyrhiza is a species of duckweed that ranks among the smallest angiosperms known (sizes 0.5–18 mm). It is a free-floating aquatic plant that reproduces vegetatively in largest part [26]. To overcome unfavorable conditions, this species produces a starch-rich tissue called a turion that is denser than normal fronds, and sinks to the bottom of water sources until conditions

become favorable [27]. Similar to other duckweed species, under appropriate conditions, *S. polyrhiza* grows at high rates, which helps it to cover large portions of the surface of the water bodies where it is present [28].

Considering their high growth rates, small size, simple structures, and potential for industrial applications, duckweed species have been the subject of intensive and detailed research [29]. Among the most notable applications are possible utility in water treatment [30], bioenergy [31], animal feeding [27], human nutrition [32,33], and pharmaceutical applications [34]. Given the potential of duckweed species as model organisms [35,36], stock collections of these species have been established by several institutions around the world, which have aided substantially in promoting further research on these species [37]. As such, various aspects of *S. polyrhiza* physiology, genome, and potential industrial applications have been studied in detail [27,38–41]. A remarkable characteristic of the geographic distribution of *S. polyrhiza* is that it extends worldwide. According to Les et al. [42], duckweed species and other aquatic plants dispersed transoceanically in the recent past, which highlights the importance of external biotic dispersal for this species [43]. Duckweed dispersal is mainly via adhesion to aquatic birds and mammals [25,44,45]. However, little has been done to understand the macroecological factors that drive its distribution, and only general aspects of the regions occupied by *S. polyrhiza* have been characterized [29].

Occurrence data

We obtained geographic occurrence records for *S. polyrhiza* from the Global Biodiversity Information Facility (GBIF [46]) and the Botanical Information and Ecology Network (BIEN [47]). In all, a total of 85,923 georeferenced records were obtained (GBIF: 84,992; BIEN: 931; a complete set of references is in S1 Text). We cleaned data from each database independently to exclude records from before 1970, lacking coordinates, with zero values for longitude and latitude, or

duplicates [48]. Records marked as absent or uncommon were also removed from the GBIF data. After this initial cleaning, we had 45,913 records (GBIF: 45,459; BIEN: 454). We combined records from the two sources and excluded duplicates again. Records that were outside of, but closer than ~5' to the edge of environmental layers were moved to the nearest pixel with information; points falling farther outside layer borders were removed. To reduce bias from spatial autocorrelation, we thinned records using a minimum point-to-point distance of ~30'. The final number of records after these procedures was 964. Occurrence data download, cleaning, and thinning were accomplished using rgbif [49], spooc [50], BIEN [51], and ellipsenm [52] in R [53].

Environmental variables

We obtained raster environmental data layers from three sources: (1) bioclimatic (BIO) and solar radiation (SR) layers from WorldClim v2.1, at 10' resolution (available at www.worldclim.com [54]); (2) cation exchange capacity (CEC), organic carbon (OC), and pH, from the ISRIC – World Soil Information database, at 250 m resolution (available at www.soilgrids.org [55]); and (3) total phosphorus (TP), labile inorganic phosphorus (LIP), and organic phosphorus (OP) in soils, from Global Gridded Soil Phosphorus at 30' resolution (available at www.daac.ornl.gov/SOILS/guides/Global_Phosphorus_Dist_Map.html [56]).

Solar radiation layers were available as averages for the 12 months of the year. To create layers that better represented extremes and annual averages, we created the following “bioclimatic-like” layers: annual mean solar radiation (AMSR), maximum solar radiation of the month with maximum values (SRMax), minimum solar radiation of the month with minimum values (SRMin), range of solar radiation (RSR), average solar radiation of the quarter with highest values (ASRQH), and average solar radiation of the quarter with lowest values (ASRQL). We created these variables using the values for the 12 months obtained from WorldClim. Variable processing and calculations were done using the packages raster [57] and gdalUtilities [58], in R.

To test the effect of spatial resolution on the outcome of variable selection processes, we created two groups of variables, at distinct spatial resolutions: (1) a group at 10' resolution including BIO and SR variables, plus CEC, OC, and pH; and (2) a group at 30' resolution including BIO and SR variables, plus TP, LIP, and OP. We performed raster aggregation procedures (average of values) on CEC, OC, and pH to match the resolution of BIO variables, and on BIO and SR variables to match the resolution of variables at 30'. Although the set of variables representing soil conditions used at 10' differs from the one at 30', variable selection analyses will help to identify whether the variables selected differ between the two resolutions.

Geographic areas for model calibration

To explore the effect of the areas across which models are calibrated on the set of variables selected, we explored four options for such areas in our analysis: (1) buffers of ~5° (~500 km at the Equator) around occurrence records, (2) concave-hull polygons with a buffer of ~5°, (3) ecoregions occupied by the species buffered by ~1° (~100 km), and (4) the result of intersecting the previous three areas. Although a new simulation-based approach has been recently suggested as a reliable tool to estimate calibration areas [59], the broad distribution of this species makes it difficult for that method to be applied. Our chosen calibration areas are therefore reasonable options to calibrate models considering that such areas should reflect what regions could have been accessible to the species and present relevant environments for comparisons (Fig 1). The two groups of variables were masked to the four areas. We created these calibration areas using the packages *ellipsenm*, *rgeos* [60], and *rgdal* [61] in R.

Fig 1. Geographic representation of species occurrence data for *Spirodela polyrhiza* (upper panel) and areas for model calibration used to create ecological niche models.

Occurrence data represented are after filtering and spatial-thinning procedures. Buffer and concave areas are represented before masking them to land areas for purposes of representation.

Modeling algorithms

We used generalized linear models (GLMs) and Maxent [62,63] to estimate the ecological niche of the species. These algorithms are both used commonly in the literature and produce reliable and good-performing models [64,65]. For contrasts in model calibration, Maxent uses presences and a characterization of the background, whereas GLMs use presences and pseudo-absences; both background and the pseudo-absences were taken as a sample of available pixels across the calibration area. For purposes of comparison, the same points (20,000) were used in both algorithms, and were treated as both background and pseudo-absence data. The sample of 20,000 points was taken for each calibration area option independently. GLMs were performed as logistic models with a weight of 1 for presences and 10,000 for pseudo-absences (e.g., [14]). GLMs created in such a way are considered to be similar mathematically to Maxent models under certain conditions and assumptions [66].

Variable selection process

As variable selection can be done in multiple ways and at distinct points in the process of data preparation or modeling, we followed a multi-step approach that considers quantitative and qualitative characteristics of predictors (Fig 2). Our approach consisted of (1) assessing linear correlations among variables; (2) exploring variable values in occurrences and across calibration areas; (3) an initial selection based on the criteria (1) and (2) and the biological relevance of variables; (4) creating variable sets resulting from all combinations of two or more initially selected variables; and (5) including all sets of variables in model calibration exercises to identify which algorithm parameters and variable sets, in concert, result in the best-performing models. We

performed exploration exercises 16 times: the combination of two environmental data sets, four calibration areas, and two algorithms.

Fig 2. Schematic representation of the process that was followed to select variables to model the ecological niche of the greater duckweed (*Spirodela polyrhiza*).

After the first two steps, initial selection of variables was done based on three considerations: (1) groups of variables with all variable-to-variable pairwise correlations $|r| \leq 0.8$ (Fig 3); (2) biological relevance of variables; and (3) variables for which the calibration area had wider limits in environmental dimensions than the occurrences [67], based on histogram plots of values (Fig 4). The latter consideration assumes that using variables for which the entire spectrum of responses can be characterized makes for better models. Biological relevance of variables was determined based on details about the species' natural history, phenology, and physiology in the literature, and our own experience with populations in the field and controlled environments. For simplicity, we selected the same initial set of variables based on the relevance criterion for the four calibration areas considered.

Fig 3. Histograms of variable values in calibration areas (M) and occurrence records. Results for variables at 10' resolution and calibration areas resulting from intersection are shown. See other results in S1-S7 Figs.

Fig 4. Results from linear correlation tests for initial variables. Values of correlation above $|0.8|$ are magnified threefold. Results for variables at 10' resolution are shown. See other results in S8 Fig.

Using the groups of variables remaining after the initial selection, we prepared subsets of variables representing all combinations of three or more variables [14]. Such sets of variables were then used as part of our process of model calibration in which other parameter settings were tested. For both Maxent and GLMs, we tested five response types (lq, lp, q, qp, lqp; l = linear, q = quadratic, and p = product). For Maxent, six regularization multiplier values were explored (0.1, 0.3, 0.6, 1.0, 2.5, 5.0). Performance of candidate models was evaluated based on three criteria: statistical significance of predictions (partial ROC; [68]), omission rate (allowing a 5% omission error; [69]), and model fit and complexity (based on the Akaike information criterion for GLMs, and the AICc proposed by Warren and Seifert [70] for Maxent).

In total, then, for each model calibration exercise, 10,180 and 5065 GLM models were tested at 10' and 30', respectively, and 61,080 and 30,390 Maxent models were tested at 10' and 30', respectively. Model calibration processes with Maxent were done using the kuenm R package [12], and model calibration using GLMs was done using stats in R [53].

Importance of variables and effects on models

To understand the effect of selected variables on models that could be used to represent species niches and/or potential distributions, we created a final model for each of the calibration areas at the two resolutions of variables, using parameters and sets of variables selected after model calibration. Then, we measured the effects of variables on such models: in Maxent, we used jackknife analysis to measure variable contributions [71], and for GLMs, we used an ANOVA to explore deviances for each of the predictors considered and whether deviance values were statistically significant (whether the deviance was larger than expected by chance).

We transferred all the models across the area comprising the union of the four calibration areas, and compared those models to assess whether patterns of suitability values differed as a

result of using distinct variables, calibration areas, and algorithms. Model transfers in Maxent were done using free extrapolation, no replicates, and a cloglog output format. Model transfers for GLMs were scaled 0–1. As ecological niches exist simultaneously in both geographic and environmental spaces [72], we created 3-dimensional visualizations of resulting predictions in a space defined in terms of some of the environmental variables with larger effects on our models (see below). Explorations in environmental space were used to detect how variation in suitability was associated with variable values.

Results

Results from the selection process

Graphical explorations of environmental conditions across calibration areas and occurrence records varied somewhat among the distinct options of areas for calibration (Fig 3, S1-S7 Figs). Despite such variations, these explorations allowed us to identify variables that appeared better for detecting suitable and unsuitable conditions based on distributions of values and confidence limits. Variable correlations also varied slightly among the distinct options of calibration areas tested, although we consistently found more highly correlated variables in calibration areas derived from ecoregions (Fig 4, S8 Fig). After considering the exploration of environmental conditions, correlation values, and biological importance, we retained 11 variables at 10' resolution and 10 variables at 30' resolution. The variables mean diurnal range (BIO 2), maximum temperature of warmest month (BIO 5), minimum temperature of coldest month (BIO 6), annual precipitation (BIO 12), precipitation of driest month (BIO 14), precipitation seasonality (BIO 15), range of solar radiation (RSR), and average solar radiation of the quarter with highest values (ASRQH), were in common between these sets; cation exchange capacity (CEC), organic carbon (OC), and pH were kept for the set at 10', whereas labile inorganic phosphorus (LIP) and organic phosphorus (OP) were kept at 30'.

All model calibration exercises found at least one parameter setting that produced a model that met all criteria for selection (S1 and S2 Tables). Variables selected contrasted markedly among treatments that considered distinct calibration areas, spatial resolutions, and modeling algorithms (Fig 5). None of the final sets of variables selected during model calibration used all of the variables initially selected. In general, fewer variables were selected for models created with Maxent at 10' resolution (2-4) than for the other algorithm/resolution combinations (6-7). Although the subsets of variables considered were not totally comparable between the tests at distinct resolutions, at least one variable representing soil conditions was consistently selected across all exercises using distinct calibration areas, using at least one of the modeling algorithms. Soil and solar radiation variables were more consistently selected at 10' resolution, especially when using Maxent, whereas at 30' resolution, bioclimatic and solar radiation variables were more consistently selected. Bioclimatic and solar radiation variables that represent extreme conditions or means of extreme periods appeared to be selected more consistently regardless of the differences in spatial resolution or algorithm (Fig 5).

Fig 5. Summary of variables retained after the multi-step approach for selection.

Results depending on spatial resolution of predictors, model calibration areas, and the algorithm used are shown. BIO2 = mean diurnal range of temperature; BIO5 = maximum temperature of warmest month; BIO6 = minimum temperature of coldest month; BIO12 = annual precipitation; BIO14 = precipitation of driest month; BIO15 = precipitation seasonality; RSR = range of solar radiation; ASRQH = average solar radiation of the quarter with highest values; CEC = cation exchange capacity; OC = organic carbon; LIP = labile inorganic phosphorus; OP = organic phosphorus.

Effects of variables on models

Bioclimatic and solar radiation variables had consistently larger effects than soil variables on Maxent models (S9 and S10 Figs), with the exception of CEC, which was the most important variable for the only model that selected this predictor (i.e., with variables at 10' using calibration areas that intersected the other three options; Fig S9). The most important predictor for Maxent models varied among BIO variables and CEC at 10' resolution, whereas at 30', BIO 6 was consistently selected as more important based on the contribution, permutation importance, and jackknife results. For GLMs, bioclimatic variables, a few quadratic versions, and products of such variables, as well as CEC, contributed most to the deviance in models at 10' resolution (S3-S6 Tables). Solar radiation variables were not particularly relevant to explain deviance in these models. At 30' resolution, deviance in models was mostly explained by bioclimatic and solar radiation variables, whereas soil variables did not explain large portions of the deviance (S7-S10 Tables).

Model projections

Geographic transfers of Maxent models at 10' resolution showed higher variability across distinct calibration areas than GLM projections (Fig 6, S11 Fig). Variation was greatest in northern and eastern Asia, central North America, eastern Australia, and northern and southern Africa. At 30' resolution, geographic transfers showed lower variability for both GLM and Maxent models.

Fig 6. Geographic projections of suitability values deriving from final models created with the variables selected. Results for variables at 10' resolution are shown. See other results in S11 Fig.

Projections of suitability in environmental space showed higher variability in Maxent projections than in GLMs, considering distinct calibration areas at 10' resolution (Fig 7, S12-S26

Figs). That is, suitability values varied highly across the regions of the environment detected as suitable (above the 5% omission threshold). In most Maxent projections of suitability in environmental space, and for various environmental variables, extreme environments were predicted to have high suitability (i.e., we observed truncated responses [2] in our models). GLM projections were more stable in both aspects; in these projections, and considering most variables, regions of high suitability tended to be surrounded by regions with decreasingly lower suitability (i.e., extreme environments were only rarely detected as the most suitable ones). At 30' resolution, projections of environmental space looked similar across distinct calibration areas and modeling algorithms. Perhaps the main difference is that Maxent constrained suitable environments a little more than did the GLMs.

Fig 7. Projections of suitability values in a three-dimensional environmental space.

Values of suitability derive from final models created with selected variables and parameters. GLM results for variables at 10' resolution and calibration areas resulting from intersection are shown. See other results in S12-S26 Figs.

Discussion

Determining sets of appropriate environmental variables with which to model ecological niches or potential distributions of species remains a major challenge in distributional ecology. Our results showed that multiple factors generated distinct outcomes regarding which variables are better for model development. Even after an initial selection of relevant variables, when we let statistical methods select the best subset of variables, choice of spatial resolution of layers, area for model calibration, and modeling algorithm, all affected the final subset of predictors selected. However, despite the fact that distinct groups of variables were selected when these

factors were changed, some predictors were more consistently selected than others, which hints at the relevance of such variables when understanding ecological limitations for the species.

Of particular interest is the recurring selection of variables representing maximum temperature (BIO 5) and minimum temperature (BIO 6), which, in most cases, helped to identify the limits of what is suitable for the species in regions with high and low temperatures. The consistent selection of these two variables could be attributed to the importance of the temperature in the natural history of the species, especially in that low temperatures are responsible for triggering production of turions [25]. If temperatures in an area are consistently low, the species will only produce these dormant fronds, and populations will stop growing (i.e., the species will be outside of its thermal niche, at least for reproduction).

Obtaining different sets of key variables in modeling exercises based on distinct calibration areas is concerning. One of the main aspects to be defined when creating such models is the area over which models will be calibrated [19,59]. This dependency has implications from both statistical and ecological perspectives. From a modeling point of view, the environmental values of the points selected as background or pseudoabsences affect how models are fitted to the occurrence records, sometimes resulting in overfitted models, which complicates model transfers [73,74]. In regard to the ecological relevance of these areas, because models are fitted within these regions, the associations to be found are only relevant if a species has had access to those environmental conditions [18]. As distinct calibration areas affected the set of variables selected and the effects of such variables in the models, correct definition of these areas becomes an even more important challenge. New methods to define calibration areas are now available that account for ecological, historical, and dispersal factors, which may result in more properly calibrated models and more consistent sets of variables [59]. However, this challenge persists in cases in which limited information exists about a species, or the distribution of species is close to global, as in this example.

The other two factors explored (modeling algorithm and spatial resolution of variables) also affected the set of environmental variables selected for niche models. As in previous explorations [22,75], the effects of these two factors were seen clearly in the transfers of models, both geographic and environmental, and thus cannot be neglected. Spatial resolution of layers has been noticed as a factor that can influence the sets of variables selected [21]. Depending on the spatial resolution of layers, the number of environmental combinations found in an area can change, with more numerous combinations at finer spatial resolutions. One of the complications deriving from these differences is that the ways in which variables are correlated can change at distinct resolutions due to changes in sample size [76], which can modify the initial selection of variables that is made, not necessarily related to the biological importance of such variables. Our inclusion of distinct modeling algorithms showed that combining these factors certainly increases the complexity of the process of selecting variables. The relevance of distinct variables has been shown to change depending on the algorithm used [22]. Although it is not clear whether the set of variables should be changed depending on the algorithm (if the variables have been preselected in some way), the fact that distinct algorithms work differently and that distinct predictors have distinct effects should not be overlooked [77].

Exploring environmental conditions within calibration areas and in occurrence records beforehand helped to identify variables for which truncated responses could be found. Although this point may not be related to the biological role of this environmental dimension, it is crucial in being able to transfer a model to other conditions with less ambiguity [67,78]. Maximum temperature of warmest month (BIO 5) and minimum temperature of coldest month (BIO 6) are examples of variables that contributed importantly to models, and, as expected, values of suitability were higher at intermediate values, with decreasing suitability towards extreme environmental values. Performing these explorations can help to select predictors appropriately when the goal is to understand why species are distributed the way they are. However, other variables should not be discarded only based on these graphical explorations, because they may

be important environmental constraints despite the truncation. For instance, cation exchange capacity (CEC), a soil variable, showed truncation towards lower values in our examples (Fig 7), but still was selected across various of the experiments, and its contribution to models was not negligible.

In spite of the variability found in the results, we found that variables related to temperature extremes were critical in characterizing the greater duckweed ecological niche, which concurs with findings from experimental work done with this species [25,79]. In fact, temperature may be the main factor shaping the distribution of this species, especially considering its distributional limits at high latitudes. Models created using precipitation variables (particularly those using precipitation of driest quarter; BIO 14) correctly discarded suitability in xeric regions, showing the importance of considering a factor that represents water availability [80]. Solar radiation of quarters periods with higher values (ASRQH) and range of solar radiation (RSR) were also potentially helpful in limiting the distribution of the species towards higher latitudes, as solar radiation informs about a crucial resource for photosynthesis, and experimental work has confirmed the importance of this factor [31,38]. Factors related to soil variables that served as proxies for nutrient availability and water conditions also showed high importance in some of the results. Although nutrients are critical for the development of this species, the fact that soil variables are only indirect proxies for such information [81] and the complications of representing this type of information at the scale of our analyses may explain why these variables were not selected as consistently as others.

In sum, we showed that selecting relevant variables to characterize ecological niches and potential distributions becomes even more complicated when multiple factors related to data processing and model development are considered. However, if a series of criteria and approaches is applied in concert, certain variables are selected more consistently than others. Such variables may in effect be the ones that shape and constrain the species' distribution from a macroecological point of view. Variables representing extreme temperatures, dry periods,

seasonality of solar radiation, summer solar radiation, and some soil proxies of nutrients in water were among the factors that contributed the most to shaping the distribution of *S. polyrriza*.

Acknowledgments

MEC thanks his doctoral dissertation committee for the initial suggestion of this set of analyses. We thank the members of the KUENM working group in the University of Kansas Biodiversity Institute, for their thinking and work on these topics over the years.

References

1. Franklin J. Mapping Species Distributions: Spatial Inference and Prediction. Cambridge University Press; 2010.
2. Peterson AT, Soberón J, Pearson RG, Anderson RP, Martínez-Meyer E, Nakamura M, et al. Ecological Niches and Geographic Distributions. Princeton: Princeton University Press; 2011.
3. Peterson AT, Papeş M, Soberón J. Mechanistic and correlative models of ecological niches. Eur J Ecol. 2015;1: 28–38. doi:10.1515/eje-2015-0014
4. Smith AB, Santos MJ. Testing the ability of species distribution models to infer variable importance. Ecography. 2020;43: 1801–1813. doi:https://doi.org/10.1111/ecog.05317
5. Araújo MB, Guisan A. Five (or so) challenges for species distribution modelling. J Biogeogr. 2006;33: 1677–1688. doi:10.1111/j.1365-2699.2006.01584.x
6. Austin MP, Van Niel KP. Improving species distribution models for climate change studies: Variable selection and scale. J Biogeogr. 2011;38: 1–8. doi:10.1111/j.1365-2699.2010.02416.x

- 454 7. Bradie J, Leung B. A quantitative synthesis of the importance of variables used in MaxEnt
455 species distribution models. *J Biogeogr.* 2017;44: 1344–1361. doi:10.1111/jbi.12894
- 456 8. Cruz-Cárdenas G, López-Mata L, Villaseñor JL, Ortiz E. Potential species distribution
457 modeling and the use of principal component analysis as predictor variables. *Rev Mex*
458 *Biodivers.* 2014;85: 189–199. doi:10.7550/rmb.36723
- 459 9. Simões M, Romero-Alvarez D, Nuñez-Penichet C, Jiménez L, Cobos ME. General theory
460 and good practices in ecological niche modeling: A basic guide. *Biodiv Inform.* 2020;15:
461 67–68. doi:10.17161/bi.v15i2.13376
- 462 10. Araújo MB, Peterson AT. Uses and misuses of bioclimatic envelope modeling. *Ecology.*
463 2012;93: 1527–1539. doi:10.1890/11-1930.1
- 464 11. Kriticos DJ, Jarošik V, Ota N. Extending the suite of bioclim variables: A proposed registry
465 system and case study using principal components analysis. *Methods Ecol Evol.* 2014;5:
466 956–960. doi:10.1111/2041-210X.12244
- 467 12. Cobos ME, Peterson AT, Barve N, Osorio-Olvera L. kuenm: An R package for detailed
468 development of ecological niche models using Maxent. *PeerJ.* 2019;7: e6281.
469 doi:10.7717/peerj.6281
- 470 13. O'Brien RM. A caution regarding rules of thumb for variance inflation factors. *Qual Quant.*
471 2007;41: 673–690. doi:10.1007/s11135-006-9018-6
- 472 14. Cobos ME, Peterson AT, Osorio-Olvera L, Jiménez-García D. An exhaustive analysis of
473 heuristic methods for variable selection in ecological niche modeling and species
474 distribution modeling. *Ecol Inform.* 2019;53: 100983. doi:10.1016/j.ecoinf.2019.100983
- 475 15. Petitpierre B, Broennimann O, Kueffer C, Daehler C, Guisan A. Selecting predictors to
476 maximize the transferability of species distribution models: Lessons from cross-continental
477 plant invasions. *Global Ecol Biogeogr.* 2017;26: 275–287. doi:10.1111/geb.12530
- 478 16. Fan JY, Zhao NX, Li M, Gao WF, Wang ML, Zhu GP. What are the best predictors for
479 invasive potential of weeds? Transferability evaluations of model predictions based on

- diverse environmental data sets for *Flaveria bidentis*. Weed Res. 2018;58: 141–149.
doi:10.1111/wre.12292
17. Feng X, Park DS, Liang Y, Pandey R, Papeş M. Collinearity in ecological niche modeling: Confusions and challenges. Ecol Evol. 2019;9: 10365–10376. doi:10.1002/ece3.5555
18. Barve N, Barve V, Jiménez-Valverde A, Lira-Noriega A, Maher SP, Peterson AT, et al. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. Ecol Modell. 2011;222: 1810–1819. doi:10.1016/j.ecolmodel.2011.02.011
19. Anderson RP, Raza A. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: Preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. J Biogeogr. 2010;37: 1378–1393. doi:10.1111/j.1365-2699.2010.02290.x
20. Harris RMB, Porfirio LL, Hugh S, Lee G, Bindoff NL, Mackey B, et al. To be or not to be? Variable selection can change the projected fate of a threatened species under future climate. Ecol Manag Restor. 2013;14: 230–234. doi:https://doi.org/10.1111/emr.12055
21. Nyström Sandman A, Wikström SA, Blomqvist M, Kautsky H, Isaeus M. Scale-dependent influence of environmental variables on species distribution: A case study on five coastal benthic species in the Baltic Sea. Ecography. 2013;36: 354–363. doi:10.1111/j.1600-0587.2012.07053.x
22. Bosch S, Tyberghein L, Deneudt K, Hernandez F, De Clerck O. In search of relevant predictors for marine species distribution modelling using the MarineSPEED benchmark dataset. Divers Distrib. 2018;24: 144–157. doi:10.1111/ddi.12668
23. Guo C, Lek S, Ye S, Li W, Liu J, Li Z. Uncertainty in ensemble modelling of large-scale species distribution: Effects from species characteristics and model techniques. Ecol Modell. 2015;306: 67–75. doi:10.1016/j.ecolmodel.2014.08.002
24. Zhu G, Gutierrez Illan J, Looney C, Crowder DW. Assessing the ecological niche and invasion potential of the Asian giant hornet. Proc Natl Acad Sci USA. 2020;117: 24646–

- 506 24648. doi:10.1073/pnas.2011441117
- 507 25. Jacobs DL. An ecological life-history of *Spirodela polyrhiza* (greater duckweed) with
508 emphasis on the turion phase. *Ecol Monogr.* 1947;17: 437–469. doi:10.2307/1948596
- 509 26. Landolt E. Lemnaceae. In: Kubitzki K, editor. Flowering Plants. Monocotyledons:
510 Alismatanae and Commelinanae (except Gramineae). Berlin, Heidelberg: Springer; 1998.
511 pp. 264–270. doi:10.1007/978-3-662-03531-3_28
- 512 27. Landolt E, Kandeler R. The family of Lemnaceae—A monographic study, Volume
513 2. Biosystematic investigation in the family of duckweeds, Lemnaceae. Zürich:
514 Geobotanischen Institutes der ETH; 1987.
- 515 28. Lemon GD, Posluszny U, Husband BC. Potential and realized rates of vegetative
516 reproduction in *Spirodela polyrhiza*, *Lemna minor*, and *Wolffia borealis*. *Aquat Bot.*
517 2001;70: 79–87. doi:10.1016/S0304-3770(00)00131-5
- 518 29. Tippery NP, Les DH. Tiny Plants with Enormous Potential: Phylogeny and Evolution of
519 Duckweeds. In: Cao XH, Fourounjian P, Wang W, editors. The Duckweed Genomes.
520 Cham: Springer International Publishing; 2020. pp. 19–38. doi:10.1007/978-3-030-11045-
521 1_15
- 522 30. Ziegler P, Sree KS, Appenroth K-J. Duckweed biomarkers for identifying toxic water
523 contaminants? *Environ Sci Pollut Res.* 2019;26: 14797–14822. doi:10.1007/s11356-018-
524 3427-7
- 525 31. Ma YB, Zhu M, Yu CJ, Wang Y, Liu Y, Li ML, et al. Large-scale screening and
526 characterisation of *Lemna aequinoctialis* and *Spirodela polyrhiza* strains for starch
527 production. *Plant Biol.* 2018;20: 357–364. doi:10.1111/plb.12679
- 528 32. Appenroth K-J, Sree KS, Bog M, Ecker J, Seeliger C, Böhm V, et al. Nutritional value of
529 the duckweed species of the genus *Wolffia* (Lemnaceae) as human food. *Front Chem.*
530 2018;6: 483. doi:10.3389/fchem.2018.00483
- 531 33. Appenroth K-J, Sree KS, Böhm V, Hammann S, Vetter W, Leiterer M, et al. Nutritional

- value of duckweeds (Lemnaceae) as human food. Food Chem. 2017;217: 266–273.
doi:10.1016/j.foodchem.2016.08.116
34. Appenroth K-J, Sree KS, Fakhoorian T, Lam E. Resurgence of duckweed research and applications: Report from the 3rd International Duckweed Conference. Plant Mol Biol. 2015;89: 647–654. doi:10.1007/s11103-015-0396-9
35. Laird RA, Barks PM. Skimming the surface: Duckweed as a model system in ecology and evolution. Am J Bot. 2018;105: 1962–1966. doi:10.1002/ajb2.1194
36. Appenroth K-J, Ziegler P, Sree S. Duckweed as a model organism for investigating plant-microbe interactions in an aquatic environment and its applications. Endocytobiosis Cell Res. 2016;27: 94–106.
37. Sree KS, Appenroth K-J. Worldwide Genetic Resources of Duckweed: Stock Collections. In: Cao XH, Fourounjian P, Wang W, editors. The Duckweed Genomes. Cham: Springer International Publishing; 2020. pp. 39–46. doi:10.1007/978-3-030-11045-1_15
38. Appenroth K-J, Teller S, Horn M. Photophysiology of turion formation and germination in *Spirodela polyrhiza*. Biol Plant. 1996;38: 95. doi:10.1007/BF02879642
39. Michael TP, Bryant D, Gutierrez R, Borisjuk N, Chu P, Zhang H, et al. Comprehensive definition of genome features in *Spirodela polyrhiza* by high-depth physical mapping and short-read DNA sequencing strategies. Plant J. 2017;89: 617–635. doi:10.1111/tpj.13400
40. Chen Y, Zhao X, Li G, Kumar S, Sun Z, Li Y, et al. Genome-wide identification of the Nramp gene family in *Spirodela polyrhiza* and expression analysis under cadmium stress. Int J Mol Sci. 2021;22: 6414. doi:10.3390/ijms22126414
41. Kuehdorf K, Jetschke G, Ballani L, Appenroth K-J. The clonal dependence of turion formation in the duckweed *Spirodela polyrhiza*—An ecogeographical approach. Physiol Plant. 2014;150: 46–54. doi:10.1111/ppl.12065
42. Les DH, Crawford DJ, Kimball RT, Moody ML, Landolt E. Biogeography of discontinuously distributed hydrophytes: A molecular appraisal of intercontinental disjunctions. Int J Plant

- Sci. 2003;164: 917–932. doi:10.1086/378650
43. Coughlan NE, Kelly TC, Jansen MAK. “Step by step”: High frequency short-distance epizoochorous dispersal of aquatic macrophytes. Biol Invasions. 2017;19: 625–634. doi:10.1007/s10530-016-1293-0
44. Coughlan NE, Kelly TC, Jansen M A. K. Mallard duck (*Anas platyrhynchos*)-mediated dispersal of Lemnaceae: A contributing factor in the spread of invasive *Lemna minuta*? Plant Biol. 2015;17: 108–114. doi:10.1111/plb.12182
45. Kimball RT, Crawford DJ, Les DH, Landolt E. Out of Africa: Molecular phylogenetics and biogeography of *Wolffiella* (Lemnaceae). Biol J Linn Soc. 2003;79: 565–576. doi:10.1046/j.1095-8312.2003.00210.x
46. Telenius A. Biodiversity information goes public: GBIF at your service. Nord J Bot. 2011;29: 378–381. doi:10.1111/j.1756-1051.2011.01167.x
47. Enquist BJ, Condit R, Peet RK, Schildhauer M, Thiers BM. Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. PeerJ Preprints. 2016; e2615v2. doi:https://doi.org/10.7287/peerj.preprints.2615v2
48. Cobos ME, Jiménez L, Nuñez-Penichet C, Romero-Alvarez D, Simões M. Sample data and training modules for cleaning biodiversity information. Biodiv Inform. 2018;13: 49–50. doi:10.17161/bi.v13i0.7600
49. Chamberlain S, Barve V, McGlinn D, Oldoni D, Geffert L, Ram K. rgbif: Interface to the Global Biodiversity Information Facility API. R package. 2018. Available: <https://CRAN.R-project.org/package=rgbif>
50. Chamberlain S. spocc: Interface to species occurrence data sources. R package. 2021. Available: <https://CRAN.R-project.org/package=spocc>
51. Maitner B, Brad B, Nathan C, Rick C, John D, M DS, et al. The BIEN R package: A tool to access the botanical information and ecology network (BIEN) database. Methods Ecol

- Evol. 2017;9: 373–379. doi:10.1111/2041-210X.12861
52. Cobos ME, Osorio-Olvera L, Soberón J, Peterson AT. ellipsenm: Ecological niche's characterizations using ellipsoids. R package. 2020. Available: <https://github.com/marlonecobos/ellipsenm>
53. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available: <https://www.R-project.org/>
54. Fick SE, Hijmans RJ. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. Int J Climatol. 2017;37: 4302–4315. doi:10.1002/joc.5086
55. Hengl T, Jesus JM de, Heuvelink GBM, Gonzalez MR, Kilibarda M, Blagotić A, et al. SoilGrids250m: Global gridded soil information based on machine learning. PLoS ONE. 2017;12: e0169748. doi:10.1371/journal.pone.0169748
56. Yang X, Post WM, Thornton PE, Jain A. The distribution of soil phosphorus for global biogeochemical modeling. Biogeosciences. 2013;10: 2525–2537. doi:10.5194/bg-10-2525-2013
57. Hijmans RJ. raster: Geographic data analysis and modeling. R package. 2019. Available: <https://CRAN.R-project.org/package=raster>
58. O'Brien J. gdalUtilities: Wrappers for “GDAL” utilities executables. R package. 2021. Available: <https://CRAN.R-project.org/package=gdalUtilities>
59. Machado-Stredel F, Cobos ME, Peterson AT. A simulation-based method for identifying accessible areas as calibration areas for ecological niche models and species distribution models. Front Biogeogr. 2021;13: e48814. doi:10.21425/F5FBG48814
60. Bivand R, Rundel C. rgeos: Interface to geometry engine-open source ('GEOS'). R package. 2019. Available: <https://CRAN.R-project.org/package=rgeos>
61. Bivand R, Keitt T, Rowlingson B. rgdal: Bindings for the “geospatial” data abstraction library. R package. 2019. Available: <https://CRAN.R-project.org/package=rgdal>
62. Phillips SJ, Anderson RP, Dudík M, Schapire RE, Blair ME. Opening the black box: An

- open-source release of Maxent. *Ecography*. 2017;40: 887–893. doi:10.1111/ecog.03049
63. Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. *Ecol Modell*. 2006;190: 231–259. doi:10.1016/j.ecolmodel.2005.03.026
64. Elith J, Graham CH, Anderson RP, Dudík M, Ferrier S, Guisan A, et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*. 2006;29: 129–151. doi:10.1111/j.2006.0906-7590.04596.x
65. Guisan A, Edwards TC, Hastie T. Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecol Modell*. 2002;157: 89–100. doi:10.1016/S0304-3800(02)00204-1
66. Fithian W, Hastie T. Finite-sample equivalence in statistical models for presence-only data. *Ann Appl Stat*. 2013;7: 1917–1939. doi:10.1214/13-AOAS667
67. Owens HL, Campbell LP, Dornak LL, Saupe EE, Barve N, Soberón J, et al. Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. *Ecol Modell*. 2013;263: 10–18. doi:10.1016/j.ecolmodel.2013.04.011
68. Peterson AT, Papeş M, Soberón J. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecol Modell*. 2008;213: 63–72. doi:10.1016/j.ecolmodel.2007.11.008
69. Anderson RP, Lew D, Peterson AT. Evaluating predictive models of species' distributions: Criteria for selecting optimal models. *Ecol Modell*. 2003;162: 211–232. doi:10.1016/S0304-3800(02)00349-6
70. Warren DL, Seifert SN. Ecological niche modeling in Maxent: The importance of model complexity and the performance of model selection criteria. *Ecol Appl*. 2011;21: 335–342. doi:<doi:10.1016/j.ecolmodel.2011.02.011>
71. Peterson AT, Cohoon KP. Sensitivity of distributional prediction algorithms to geographic data completeness. *Ecol Modell*. 1999;117: 159–164. doi:10.1016/S0304-3800(99)00023-X

72. Colwell RK, Rangel TF. Hutchinson's duality: The once and future niche. *Proc Natl Acad Sci USA*. 2009;106: 19651–19658. doi:10.1073/pnas.0901650106
73. Radosavljevic A, Anderson RP. Making better Maxent models of species distributions: Complexity, overfitting and evaluation. *J Biogeogr*. 2014;41: 629–643. doi:10.1111/jbi.12227
74. Yates KL, Bouchet PJ, Caley MJ, Mengersen K, Randin CF, Parnell S, et al. Outstanding challenges in the transferability of ecological models. *Trends Ecol Evolut*. 2018;33: 790–802. doi:10.1016/j.tree.2018.08.001
75. Suárez-Seoane S, Virgós E, Terroba O, Pardavila X, Barea-Azcón JM. Scaling of species distribution models across spatial resolutions and extents along a biogeographic gradient. The case of the Iberian mole *Talpa occidentalis*. *Ecography*. 2014;37: 279–292. doi:10.1111/j.1600-0587.2013.00077.x
76. Sari BG, Lúcio AD, Olivoto T, Kryszun DK, Tischler AL, Drebes L. Interference of sample size on multicollinearity diagnosis in path analysis. *Pesq Agropec Bras*. 2018;53: 769–773. doi:10.1590/S0100-204X2018000600014
77. Qiao H, Soberón J, Peterson AT. No silver bullets in correlative ecological niche modelling: Insights from testing among many potential algorithms for niche estimation. *Methods Ecol Evol*. 2015;6: 1126–1136. doi:10.1111/2041-210X.12397
78. Elith J, Kearney M, Phillips S. The art of modelling range-shifting species. *Methods Ecol Evol*. 2010;1: 330–342. doi:10.1111/j.2041-210X.2010.00036.x
79. Hitsman HW, Simons AM. Latitudinal variation in norms of reaction of phenology in the greater duckweed *Spirodela polyrhiza*. *J Evol Biol*. 2020;33: 1405–1416. doi:10.1111/jeb.13678
80. Woleck J. Assessment of the possibility of exoornithochory of duckweeds (Lemnaceae) in the light of researches into the resistance of these plants to desiccation. *Ekol Pol*. 1982;29: 405–419.

81. Austin MP. Spatial prediction of species distribution: An interface between ecological theory and statistical modelling. *Ecol Modell.* 2002;157: 101–118. doi:10.1016/S0304-3800(02)00205-3

Supporting information

S1 Fig. Histograms of environmental variable values in calibration areas and occurrence records. Results for variables at 10' resolution and buffer calibration areas are shown.

S2 Fig. Histograms of environmental variable values in calibration areas and occurrence records. Results for variables at 10' resolution and concave calibration areas are shown.

S3 Fig. Histograms of environmental variable values in calibration areas and occurrence records. Results for variables at 10' resolution and calibration areas resulting from ecoregions are shown.

S4 Fig. Histograms of environmental variable values in calibration areas and occurrence records. Results for variables at 30' resolution and buffer calibration areas are shown.

S5 Fig. Histograms of environmental variable values in calibration areas and occurrence records. Results for variables at 30' resolution and concave calibration areas are shown.

S6 Fig. Histograms of environmental variable values in calibration areas and occurrence records. Results for variables at 30' resolution and calibration areas resulting from ecoregions are shown.

S7 Fig. Histograms of environmental variable values in calibration areas and occurrence records. Results for variables at 30' resolution and calibration areas resulting from intersection are shown.

S8 Fig. Results from linear correlation tests for initial variables. Values of correlation above |0.8| are magnified threefold. Results for variables at 30' resolution are shown.

S9 Fig. Predictor contribution to Maxent models created with variables and parameter settings selected after model calibration. Results for variables at 10' resolutions are shown.

S10 Fig. Predictor contribution to Maxent models created with variables and parameter settings selected after model calibration. Results for variables at 30' resolutions are shown.

S11 Fig. Geographic projections of suitability values deriving from final models created with the selected variables. Results for variables at 30' resolution are shown.

S12 Fig. Projections of suitability values in a three-dimensional environmental space. Values of suitability derive from final models created with selected variables and parameters. GLM results for variables at 10' resolution and calibration areas resulting from buffers are shown.

695 **S13 Fig. Projections of suitability values in a three-dimensional environmental space.**
696 Values of suitability derive from final models created with selected variables and parameters.
697 GLM results for variables at 10' resolution and calibration areas resulting from concave hulls are
698 shown.

699 **S14 Fig. Projections of suitability values in a three-dimensional environmental space.**
700 Values of suitability derive from final models created with selected variables and parameters.
701 GLM results for variables at 10' resolution and calibration areas resulting from ecoregions are
702 shown.

703 **S15 Fig. Projections of suitability values in a three-dimensional environmental space.**
704 Values of suitability derive from final models created with selected variables and parameters.
705 GLM results for variables at 30' resolution and calibration areas resulting from buffers are shown.

706 **S16 Fig. Projections of suitability values in a three-dimensional environmental space.**
707 Values of suitability derive from final models created with selected variables and parameters.
708 GLM results for variables at 30' resolution and calibration areas resulting from concave hulls are
709 shown.

710 **S17 Fig. Projections of suitability values in a three-dimensional environmental space.**
711 Values of suitability derive from final models created with selected variables and parameters.
712 GLM results for variables at 30' resolution and calibration areas resulting from ecoregions are
713 shown.

714 **S18 Fig. Projections of suitability values in a three-dimensional environmental space.**
715 Values of suitability derive from final models created with selected variables and parameters.
716 GLM results for variables at 30' resolution and calibration areas resulting from intersection are
717 shown.

718 **S19 Fig. Projections of suitability values in a three-dimensional environmental space.**
719 Values of suitability derive from final models created with selected variables and parameters.
720 Maxent results for variables at 10' resolution and calibration areas resulting from buffers are
721 shown.

722 **S20 Fig. Projections of suitability values in a three-dimensional environmental space.**
723 Values of suitability derive from final models created with selected variables and parameters.
724 Maxent results for variables at 10' resolution and calibration areas resulting from concave hulls
725 are shown.

726 **S21 Fig. Projections of suitability values in a three-dimensional environmental space.**
727 Values of suitability derive from final models created with selected variables and parameters.
728 Maxent results for variables at 10' resolution and calibration areas resulting from ecoregions are
729 shown.

730 **S22 Fig. Projections of suitability values in a three-dimensional environmental space.**
731 Values of suitability derive from final models created with selected variables and parameters.
732 Maxent results for variables at 10' resolution and calibration areas resulting from intersection are
733 shown.

734 **S23 Fig. Projections of suitability values in a three-dimensional environmental space.**
735 Values of suitability derive from final models created with selected variables and parameters.
736 Maxent results for variables at 30' resolution and calibration areas resulting from buffers are
737 shown.

738 **S24 Fig. Projections of suitability values in a three-dimensional environmental space.**
739 Values of suitability derive from final models created with selected variables and parameters.
740 Maxent results for variables at 30' resolution and calibration areas resulting from concave hulls
741 are shown.

742 **S25 Fig. Projections of suitability values in a three-dimensional environmental space.**
743 Values of suitability derive from final models created with selected variables and parameters.
744 Maxent results for variables at 30' resolution and calibration areas resulting from ecoregions are
745 shown.

746 **S26 Fig. Projections of suitability values in a three-dimensional environmental space.**
747 Values of suitability derive from final models created with selected variables and parameters.
748 Maxent results for variables at 30' resolution and calibration areas resulting from intersection are
749 shown.

750

751 **S1 Table. Selected parameter settings and variables after model calibration for analyses**
752 **with variables at 10' resolution.** AIC/AICc values are not comparable across distinct calibration
753 areas.

754 **S2 Table. Selected parameter settings and variables after model calibration for analyses**
755 **with variables at 30' resolution.** AIC/AICc values are not comparable across distinct calibration
756 areas.

757 **S3 Table. Effects of predictors on GLMs produced using variables and parameter settings**
758 **selected after model calibration.** Results for models created with variables at 10' resolution,
759 using buffer calibration areas are shown. Quadratic = " 2 "; Product = ":".

760 **S4 Table. Effects of predictors on GLMs produced using variables and parameters settings**
761 **selected after model calibration.** Results for models created with variables at 10' resolution,
762 using concave calibration areas are shown. Quadratic = " 2 "; Product = ":".

763 **S5 Table. Effects of predictors on GLMs produced using variables and parameters settings**
764 **selected after model calibration.** Results for models created with variables at 10' resolution,
765 using calibration areas from ecoregions are shown. Quadratic = " 2 "; Product = ":".

766 **S6 Table. Effects of predictors on GLMs produced using variables and parameters settings**
767 **selected after model calibration.** Results for models created with variables at 10' resolution,
768 using calibration areas from intersection are shown. Quadratic = " 2 "; Product = ":".

769 **S7 Table. Effects of predictors on GLMs produced using variables and parameters settings**
770 **selected after model calibration.** Results for models created with variables at 30' resolution,
771 using buffer calibration areas are shown. Quadratic = " 2 "; Product = ":".

772 **S8 Table. Effects of predictors on GLMs produced using variables and parameters settings**
773 **selected after model calibration.** Results for models created with variables at 30' resolution,
774 using concave calibration areas are shown. Quadratic = " 2 "; Product = ":".

775 **S9 Table. Effects of predictors on GLMs produced using variables and parameters settings**
776 **selected after model calibration.** Results for models created with variables at 30' resolution,
777 using calibration areas from ecoregions are shown. Quadratic = " 2 "; Product = ":".

778 **S10 Table. Effects of predictors on GLMs produced using variables and parameters**
 779 **settings selected after model calibration.** Results for models created with variables at 30'
 780 resolution, using calibration areas from intersection are shown. Quadratic = “^2”; Product = “.”.
 781
 782 **S1 File. References for occurrence data obtained from the GBIF and BIEN databases.**

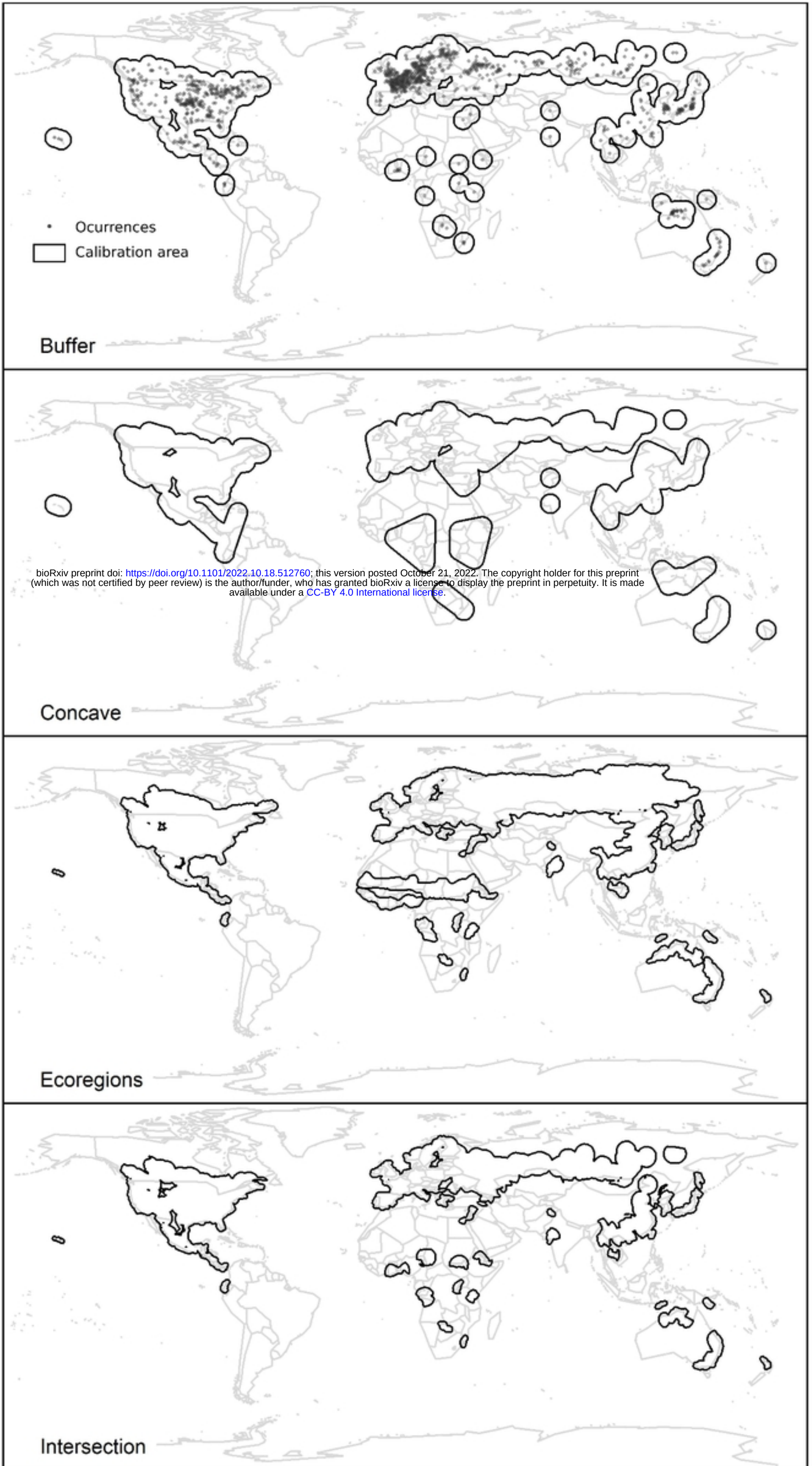


Fig 1

Variable selection process

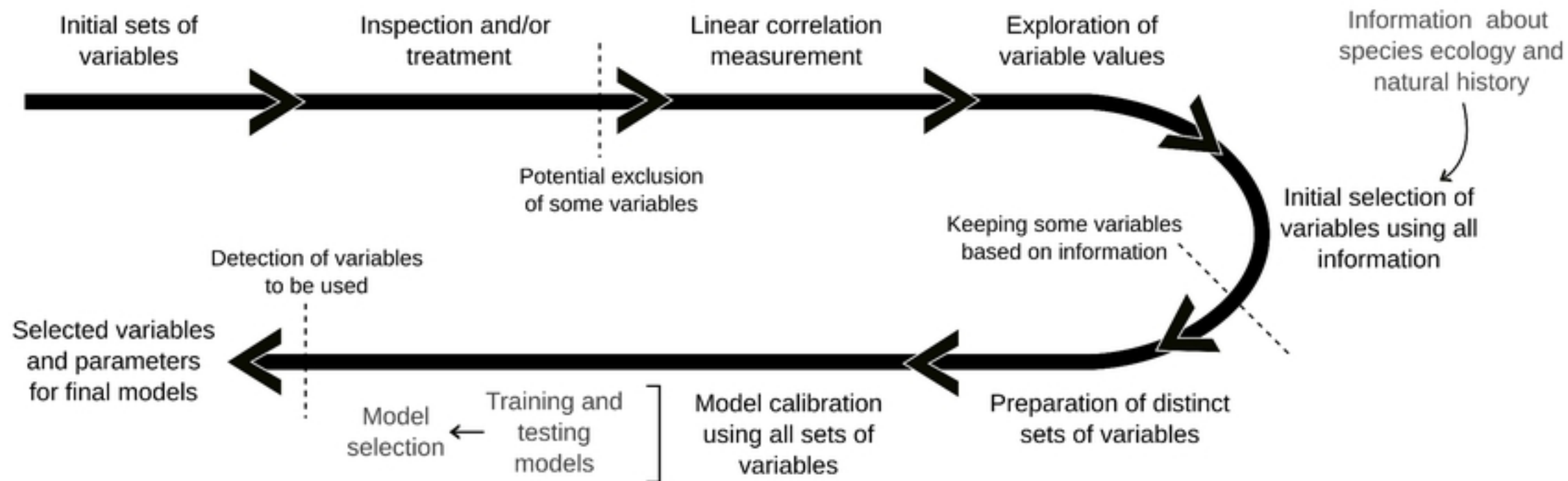


Fig 2

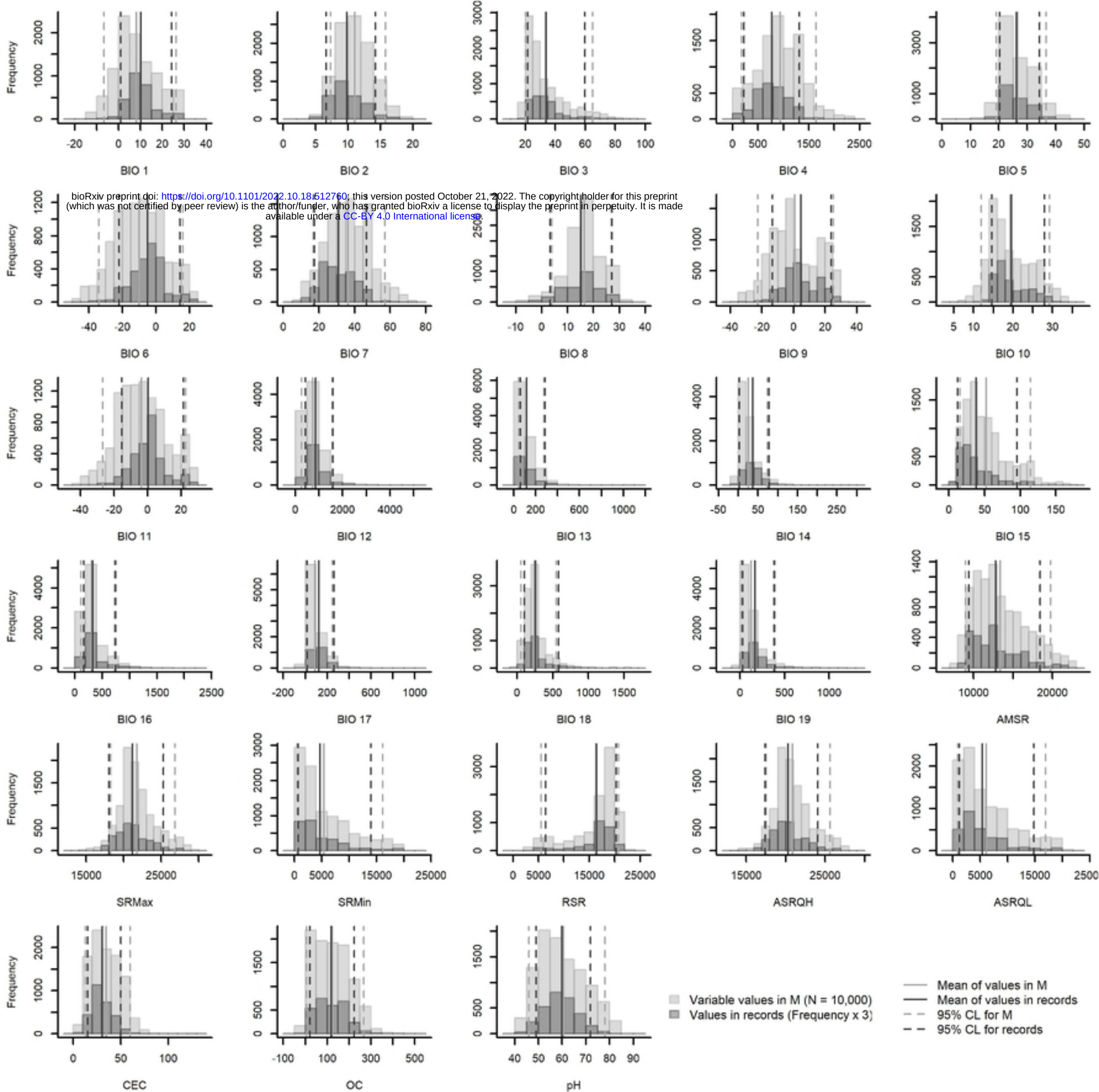


Fig 3

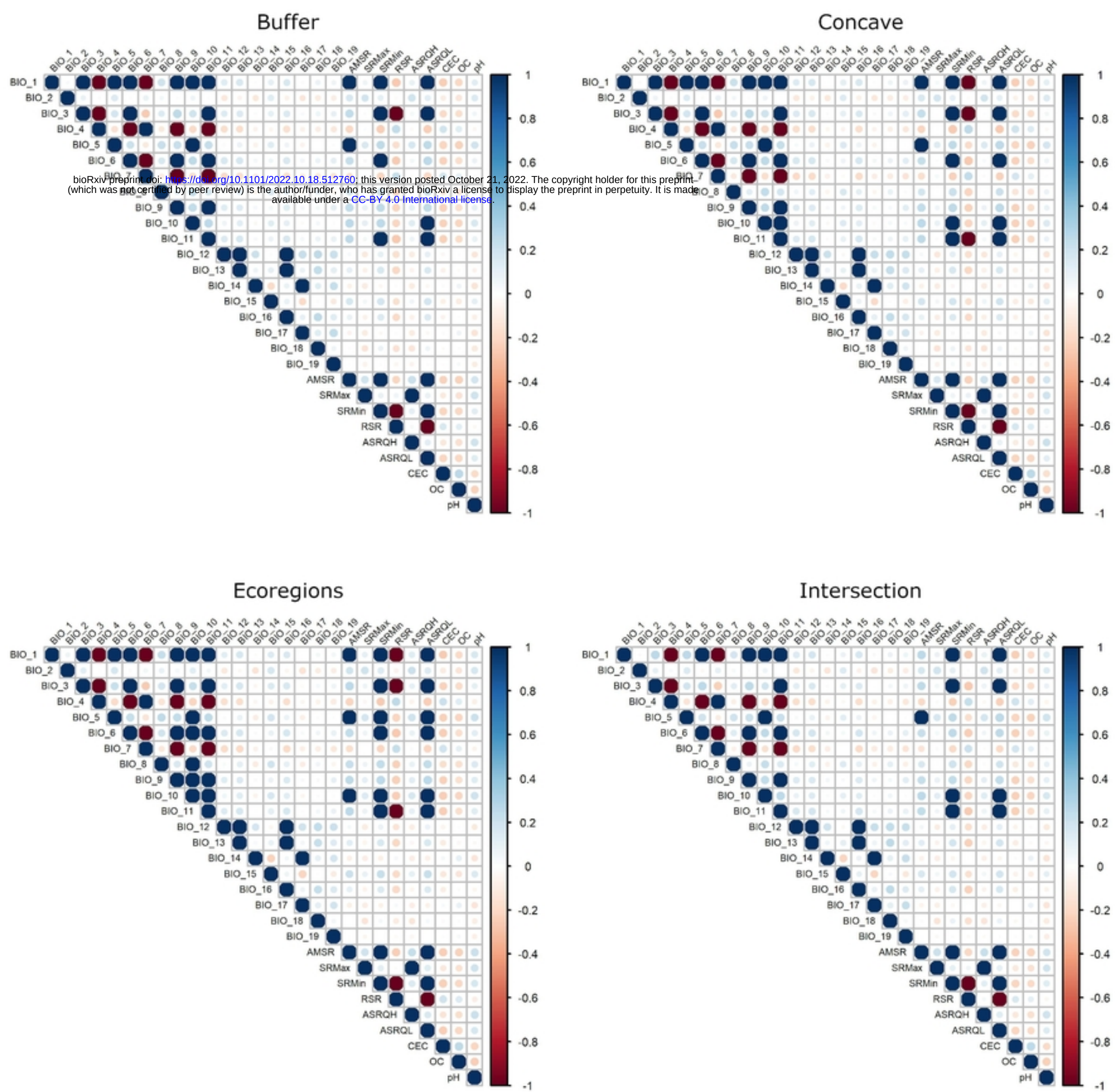


Fig 4

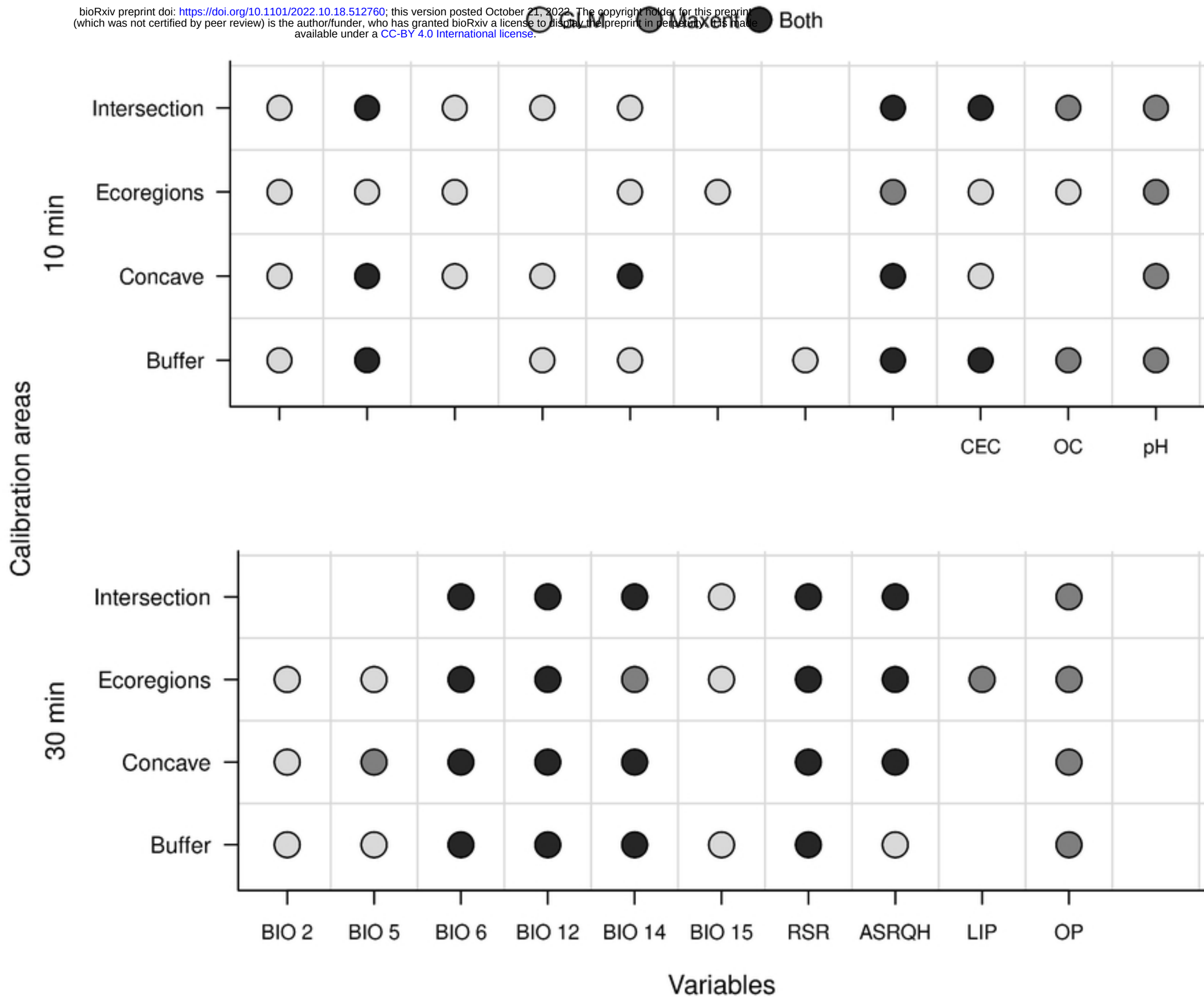


Fig 5

GLM projections

Maxent projections

bioRxiv preprint doi: <https://doi.org/10.1101/2022.10.18.512760>; this version posted October 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](#).

Buffer

Concave

Ecoregions

Intersection

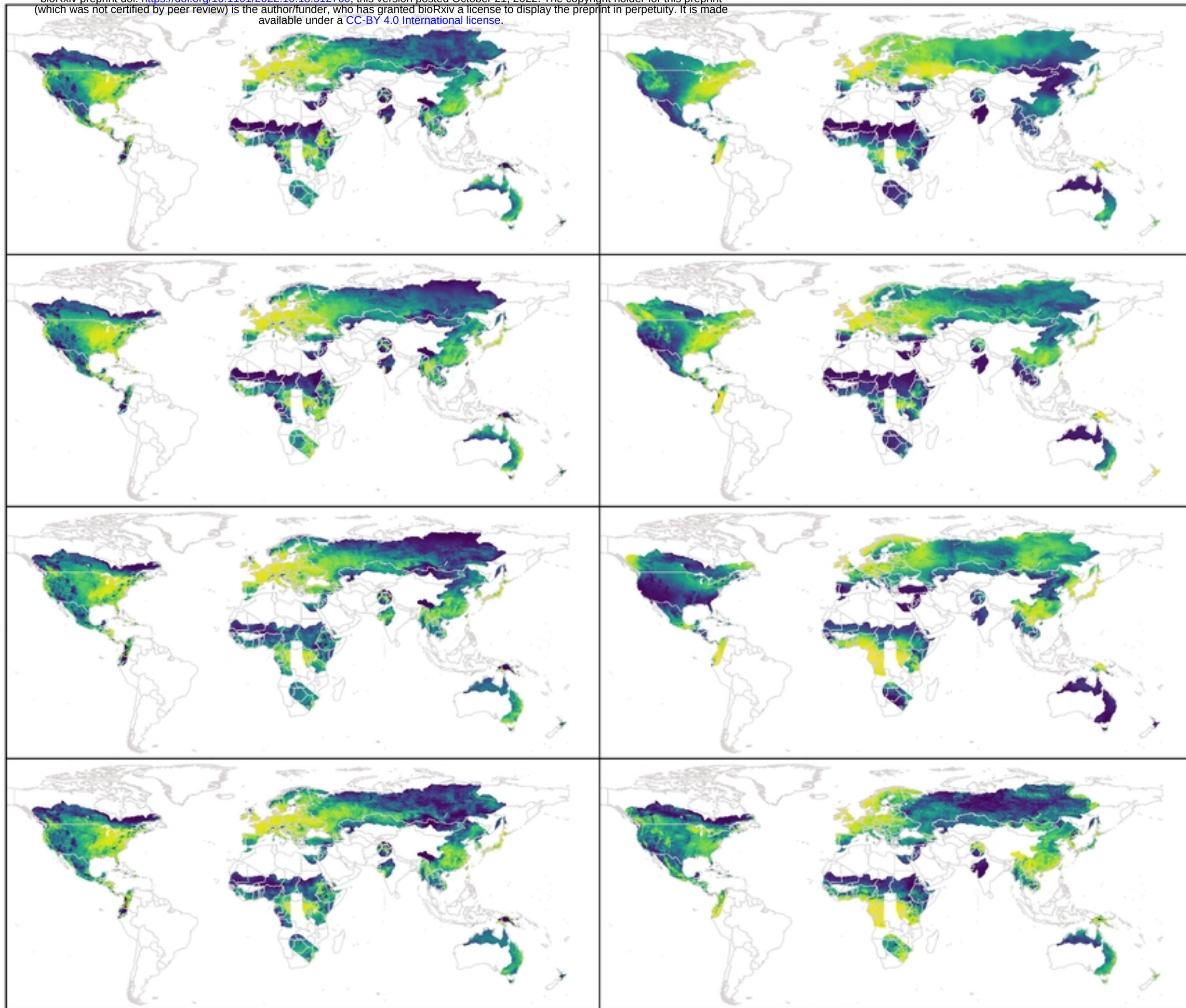


Fig 6

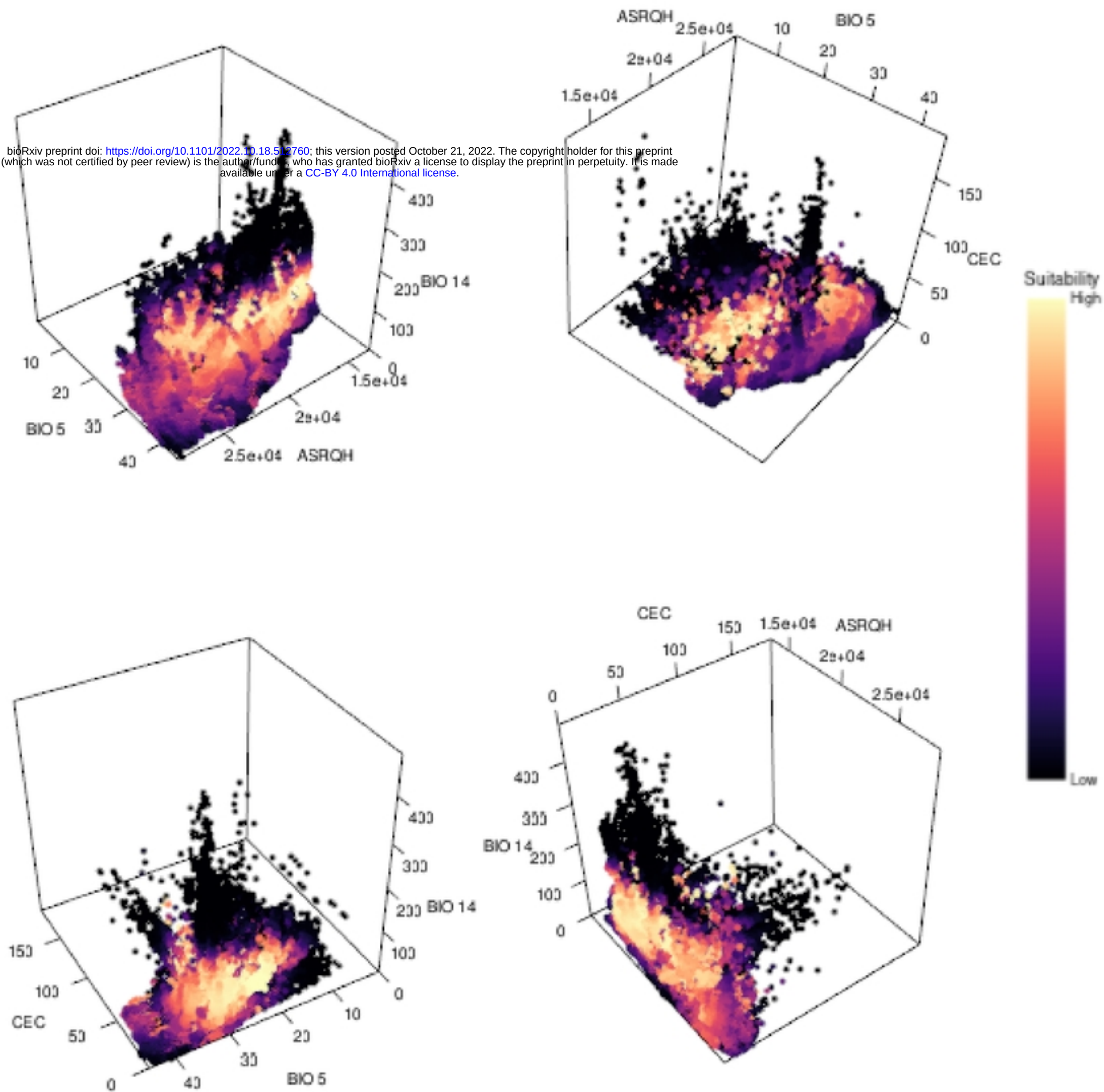


Fig 7