



Asymmetric dispersal in freshwater species distribution models varies depending on the spatial pattern of fish species

A dispersão assimétrica em modelos de distribuição de espécies de água doce varia dependendo do padrão espacial das espécies de peixes

Micael Rosa Parreira^{1*}  and João Carlos Nabout² 

¹Universidade Federal de Goiás – UFG, Campus Samambaia, Av. Esperança, s/n, Chácaras de Recreio Samambaia, CEP 74690-900, Goiânia, GO, Brasil

²Universidade Estadual de Goiás – UEG, Campus Central, BR-153, 3105, Fazenda Barreiro do Meio, CEP 75132-903, Anápolis, GO, Brasil

*e-mail: micael_rp@hotmail.com

Cite as: Parreira, M.R. and Nabout, J.C. Asymmetric dispersal in freshwater species distribution models varies depending on the spatial pattern of fish species. *Acta Limnologica Brasiliensia*, 2025, vol. 37, e16. <https://doi.org/10.1590/S2179-975X8324>

Abstract: Aim: We estimated the relative contribution of environmental and dispersal predictors in distribution models of freshwater species with different distribution ranges (29 fish species). Specifically, we tested whether model performances vary depending on the fish species' range and distribution across sub-basins and whether the relationship with those predictors is stronger depending on the type of variables (environmental or asymmetrical dispersal) used for modeling. **Methods:** The study area used for modeling was the Tocantins-Araguaia River basin, encompassing the entire hydrographic network. We applied six niche modeling methods to project the geographic distributions of 29 fish species within the Tocantins-Araguaia River basin, using environmental and asymmetrical dispersal predictors. **Results:** Generally, the models built using dispersal predictors generated more accurate predictions than those using environmental variables. However, although we found no significant difference in the accuracy among models built using different variables, their accuracy metrics were correlated with the species range and distribution across sub-basins. Also, more restricted species (i.e., lower range and distribution limited to one sub-basin) showed a greater difference in model accuracy between models built using dispersal and environmental predictors, with more accurate models being generated for restricted species when modelled using dispersal-related predictors only. **Conclusions:** the use of asymmetric dispersal predictors in SDM, besides generating accurate models, avoids/reduces model overpredictions to geographically close and climatically suitable areas, especially for restricted species, by predicting the species' distribution based on their dispersal routes through the actual directional gradient of the basin's hydrographic network.

Keywords: asymmetric eigenvector maps; Tocantins; Araguaia; range size.

Resumo: Objetivo: Estimamos a contribuição relativa de preditores ambientais e de dispersão em modelos de distribuição de espécies de água doce com diferentes amplitudes de distribuição (um total de 29 espécies de peixes). Especificamente, testamos se o desempenho dos modelos varia em função da amplitude de distribuição das espécies de peixes e da distribuição delas entre as sub-bacias, e se essa relação com os preditores é mais forte considerando os tipos de variáveis (ambientais ou de dispersão assimétrica) usadas para modelagem. **Métodos:** A área de estudo utilizada para a modelagem foi a bacia do Rio Tocantins-Araguaia, abrangendo toda a rede hidrográfica. Aplicamos seis métodos de modelagem de nicho para projetar as distribuições geográficas de 29 espécies de peixes na bacia



do Rio Tocantins-Araguaia, utilizando preditores ambientais e de dispersão assimétrica. **Resultados:** Em geral, os modelos construídos usando preditores de dispersão geraram previsões mais precisas do que aqueles que usaram variáveis ambientais. No entanto, embora não tenhamos encontrado uma diferença significativa na precisão entre os modelos construídos com diferentes variáveis, as métricas de precisão estavam correlacionadas com a amplitude e a distribuição das espécies entre as sub-bacias. Além disso, espécies mais restritas (ou seja, com menor amplitude e distribuição limitada a uma sub-bacia) apresentaram uma maior diferença na precisão dos modelos entre aqueles construídos com preditores de dispersão e preditores ambientais, sendo que modelos mais precisos foram gerados para espécies restritas quando modelados usando apenas preditores relacionados à dispersão. **Conclusões:** O uso de preditores de dispersão assimétrica em modelos de distribuição de espécies (SDM), além de gerar modelos precisos, evita/reduz superestimativas de distribuição em áreas geograficamente próximas e climaticamente adequadas, especialmente para espécies restritas, ao prever a distribuição das espécies com base em suas rotas de dispersão através do gradiente direcional real da rede hidrográfica da bacia.

Palavras-chave: mapas de autovetores assimétricos; Tocantins; Araguaia; amplitude de distribuição.

1. Introduction

Dispersal plays a fundamental role in the survival and persistence of species, the structuring of populations and communities, and the geographic distribution of species, among others (Clobert et al., 2004). Additionally, dispersal is essential for species facing changes in habitats due to anthropogenic or natural pressures, where when they move to other environments, they counteract the effects of environmental changes (Chaine & Clobert, 2012). In freshwater environments, the physical organization of the basin's hydrographic network has a significant influence on the dispersal of organisms (Tonkin et al., 2018). Specifically for freshwater fish, the dendritic structure of rivers, besides being vital for the development of the life cycle of these organisms (e.g., spawning), is essential to determine their geographic distribution as they use river and stream channels as dispersal routes (Padial et al., 2014). In this sense, when evaluating the geographic distributions of freshwater species, such as for conservation approaches, the methods used for predicting potentially suitable areas should consider the species' dispersal routes.

Species Distribution Models (SDMs) are tools often used by biogeographers and ecologists to predict potentially suitable areas for species distribution (Peterson et al., 2011). There are different frameworks for building an SDM depending on the objective (Anderson, 2013; Soliman et al., 2012), which can generate inaccurate and biased predictions when their conceptual basis is ignored (Araújo & Peterson, 2012; Jiménez-Valverde et al., 2008; Peterson & Soberón, 2012). In this sense, there are three main components for the studied distribution area of species: the abiotic (A) factors the species exists within, the biotic (B) interactions, and the "movement" (M) defined by the reachable areas and limitations of species dispersal (Barve et al., 2011; Soberón,

2007; Soberon & Peterson, 2005). However, many studies using SDMs consider only the abiotic components as environmental variables in their predictions (Peterson et al., 2011). Therefore, species dispersal capacity (movement) is often ignored when modelling the species' distribution (Miller & Holloway, 2015), which can generate overpredictions to unreachable but climatically suitable areas for the species (Mendes et al., 2020).

Species movement can be inserted into models by including accessible areas through spatial barriers or the historical species dispersal capacity (Barve et al., 2011; Miller & Holloway, 2015). Specifically for freshwater environments, species movement can be incorporated by considering the hierarchical structure and flow directionality of rivers, which are dispersal routes for species to migrate (Perrin et al., 2020; Tonkin et al., 2018). A method for assessing dispersal in the model's predictions is by incorporating symmetric spatial filters as predictors in SDMs (Allouche et al., 2008; Marco Júnior et al., 2008), while for freshwater environments, it should consider the directionality and hierarchical structure of the hydrographic network. In this case, asymmetric spatial filters (AEM) are more appropriate since they consider the dispersal routes and directionality through rivers, which is assessed through an edge-by-site directional binary matrix simulating the connections of the basin's hydrographic network (Blanchet et al., 2008).

However, it is necessary to consider the dispersal capacity of fish species in the predictions of potential distribution. For terrestrial models, it is possible to evaluate different dispersal ranges via dispersal/migration rates using, for example, distance buffers around the current distribution or species-specific fixed dispersal rates, which can be included as dispersal-related predictors in SDMs (Holloway et al., 2016; Monsimet et al., 2020). This becomes important as

models can be affected by the size of the accessible area used for modelling where more restricted species (i.e., lower dispersal capacity) generate models with higher performances given their smaller historically accessible areas (Barve et al., 2011). Indeed, there are many factors, other than extension size, that may affect the performance of SDMs (Elith & Graham, 2009), such as spatial autocorrelation (Guélat & Kéry, 2018; Segurado et al., 2006), uncertainty (Buisson et al., 2010) and species-specific factors such as prevalence (i.e., the presence/absence ratio; [Jiménez-Valverde et al., 2009]), environmental tolerance (Hernandez et al., 2006), species rarity (Franklin et al., 2009) and the sample size/distribution range (Liu et al., 2019; Stockwell & Peterson, 2002; Wisz et al., 2008). Therefore, since species dispersal is a fundamental mechanism for the distribution of freshwater fishes and varies depending on species-specific characteristics (Tonkin et al., 2018), it becomes essential to evaluate how models incorporating dispersal-related predictors perform and are affected by species-specific characteristics in comparison to those using traditional climate variables.

In this sense, we estimate the relative contribution of environmental and dispersal predictors in distribution models of freshwater species with different distribution ranges. Specifically, we tested whether the models' performances vary depending on the species range and distribution among sub-basins

and if this relationship with those predictors is higher depending on the type of variables (environmental or asymmetrical dispersal) used for modelling. We expected a significant effect of species range and distribution among sub-basins on the models' performance, especially for models built using asymmetrical dispersal (AEM) since these models consider the species directionality throughout the basin. This prediction is based on the hypothesis that inserting asymmetrical dispersal into SDMs will generate more accurate and realistic models, especially in species with more restricted distribution (lower range and limited to one sub-basin) due to a higher dispersal cost because it adds to the prediction of a directional effect of dispersal through rivers, avoiding overprediction to climatically similar but non-accessible areas or with difficult access.

2. Material and Methods

2.1. Study area

The study area used for modelling was the Tocantins-Araguaia River basin, covering the entire hydrographic network. This basin has two major rivers (Tocantins and Araguaia), forming two sub-basins that merge at one point up in the north region, close to the mouth (Figure 1). Therefore, a species occurring exclusively in one sub-basin must

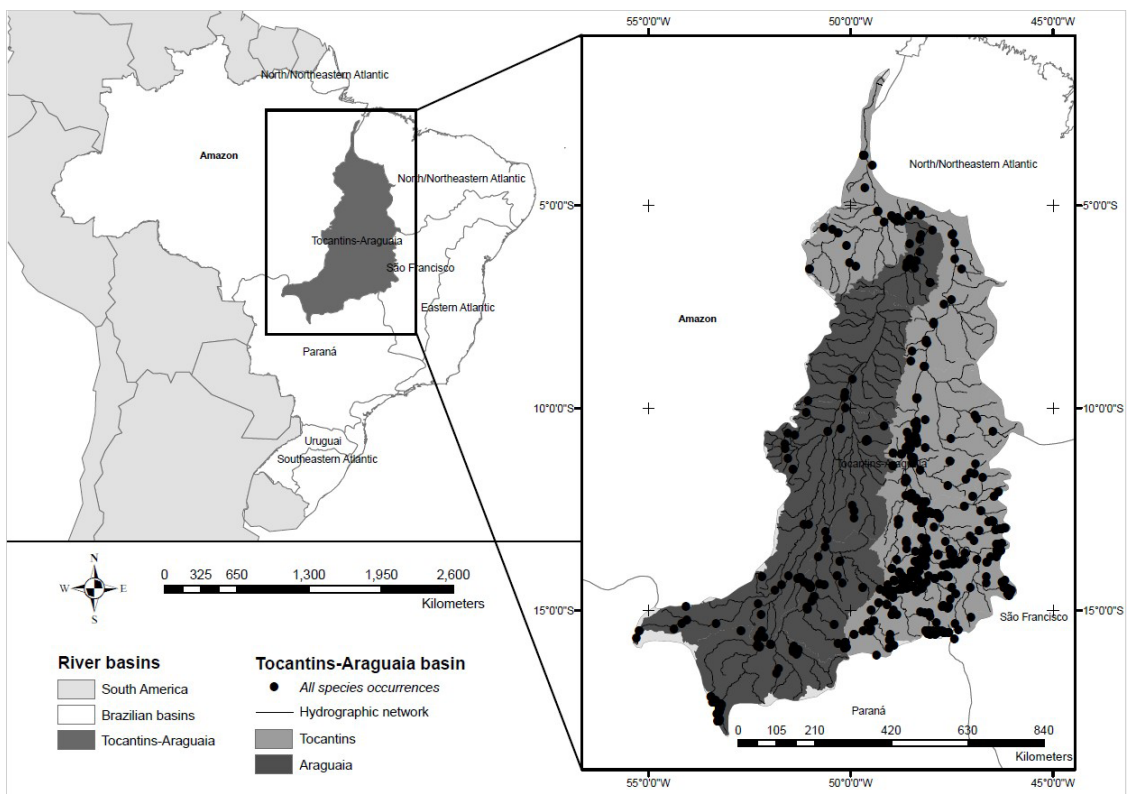


Figure 1. Study area of the Tocantins-Araguaia River basin with all species occurrences (black dots).

disperse through the main river course to occupy the other sub-basin. Thus, this basin is an interesting area for studies on dispersal limitation because it shows the species' effort to disperse between the two basins (Tocantins and Araguaia) via downstream connectivity. The entire hydrographic network was rasterized into grid-cells of 0.5° spatial resolution (latitude and longitude) totalling 282 cells for the Tocantins-Araguaia basin.

2.2. Species occurrences

We selected 29 freshwater species native to the Tocantins-Araguaia River basin. The fish occurrence records were obtained in the online databases *Species Link* (specieslink.net), *Gbif* (gbif.org) and *FishBase* (fishbase.org). The species have different distribution range sizes, where the species with the most restricted range is *Gymnotocinclus anosteos* and the widest range is *Hypostomus ericae* (Table 1). These species were chosen based on their distributions being restricted to the studied basin and having more than five unique occurrences (non-duplicated within grid-cells). In this sense, for each species we removed duplicated occurrences (more than one coordinate within a cell) by assigning a cell ID to each coordinate that fell within a cell, then removing the points with duplicated cell IDs, resulting in unique occurrences (only one occurrence per cell). These unique occurrence records (coordinates) for each species were later used to generate the models.

2.3. Environmental variables

The environmental variables used were derived from the 19 bioclimatic variables available in the Worldclim online database for the current scenario (worldclim.com). These variables were derived from monthly temperature and rainfall values to generate more biologically meaningful variables (Hijmans et al., 2005). These variables were rescaled to a spatial resolution of 0.5° (55.6 km at the equator) using the function *aggregate* of package *raster* (Hijmans, 2019) available in the R software (R Core Team, 2023). The non-collinear

environmental variables (ENV) used for modelling were selected using functions of the package *psych* (Revelle, 2019) available in the R software. The number of necessary non-orthogonal axes were determined using the *fa* function that compares the eigenvalues of factors (axes) in a scree plot, which retained four factors. Then, we selected the variables with the highest either positive or negative loadings in each of the four axes, and the following variables were selected: BIO1 = Annual Mean Temperature, BIO2 = Mean Diurnal Temperature Range, BIO13 = Precipitation of Wettest Month and BIO15 = Precipitation Seasonality (Figure S2).

2.4. Asymmetric dispersal variables

The asymmetric eigenvector map (AEM) variables (hereafter called asymmetric dispersal variables) were generated using the function *aem* of the *adespatial* package (Dray et al., 2019) available in the R software. The calculation of these variables in this function is based on the Singular Value Decomposition (SVD) analysis using an asymmetric site-by-edge binary matrix. In our study, this matrix is represented by a binary matrix of directional connectivity (asymmetric) using the grid cells as sites with nodes between grid cells and tributaries and the edge as the river path until it reaches a riverbed. The first edge (E1) represents the longest river path until it reaches the farthest riverbed. The subsequent edge is the same as the previous minus the first node (grid cell), changing only when it reaches a connection to a lower-order river tributary, which creates one or more edges (depending on tributary size and sub-connections) exclusive to this river branching. After finishing every tributary edge, the edges that count return to the main connection (river), receiving a 1 for the remaining grid cells (nodes) in the edge and a 0 for the nodes already included in previous edges, until it reaches the farthest node (grid cell with the farthest source) (see Parreira et al., 2023 for more explanation on the matrix's construction). The construction of this binary matrix is based on Blanchet et al. (2008), who modelled the directional spatial processes in freshwater environments.

Table 1. ANCOVA results for both response variables (AUC, Area Under Curve, and TSS, True Skill Statistic).

	AUC						TSS				
	Df	SS	MS	F	P		Df	SS	MS	F	P
Range	1	0.108	0.108	10.49	0.00*	Range	1	0.435	0.435	22.88	0.00*
Var_type	2	0.003	0.001	0.14	0.87	Var_type	2	0.020	0.010	0.59	0.56
Residuals	83	0.851	0.010			Residuals	83	1.578	0.019		

*Significance level at 5%. The range indicates the number of grid cells occupied by the species, and var_type indicates sets of variables used for modelling (i.e., AEM, asymmetric dispersal predictors, ENV, environmental variables, or both ENV_AEM).

Posteriorly, we used the broken stick method (Jackson, 1993) to determine the number of non-collinear asymmetric eigenvectors (AEM). These selected axes (eigenvectors) were used as asymmetric dispersal predictors for generating the species distribution models (see Figure S3 for mapped predictors).

2.5. Modelling procedures

Species distribution models (SDMs) were built using 75% of the presence and absence points for training (model's construction) and 25% for testing the models' performance (Guisan & Zimmermann, 2000) that were randomly selected from each species' unique presence and the 56 randomly generated pseudo-absences. The set of presence and absence points selected for training (constructing) were used together with modelling methods (algorithms) and sets of environmental and spatial variables. We used six algorithms: Bioclim (Nix, 1986), Domain (Carpenter et al., 1993), Support Vector Machines (SVM; Schölkopf et al., 2001), Generalized Linear Models (GLM; Nelder and Wedderburn, 1972), Maximum Entropy (MaxEnt; Phillips et al., 2006) and Random Forest (Breiman, 2001). The algorithms used are available in the R package *dismo* (Hijmans et al., 2016) and were chosen because they consider different statistical methods (climatic envelopes, environmental distances, machine-learning, regressions) (Rangel & Loyola, 2012). In addition to algorithms, the models were built using different types of variables (predictors): environmental variables (ENV), asymmetric dispersal (AEM) and the combination of environmental and dispersal variables (ENV+AEM). The model's training and validation occurred 20 times through cross-validation (using the selected presences and pseudo-absences) for all combinations of algorithms and types of variables for each species. In this sense, we built 360 models (20 repetitions x 6 algorithms x 3 types of variables) for each of the 29 species, totalling 10,440 models.

All models generated were evaluated using the Area Under the "receiver operating characteristic" Curve (AUC; Swets, 1988), which is a threshold-independent evaluation metric (the limit for determination of presences/absences) that compares predicted with observed values and the True Skill Statistic (TSS; Allouche et al., 2006), which compares the number of correct predictions minus those attributable to random guessing not affected by the prevalence and size of the validation set. The

final map of potential distribution is a consensus map (ensemble) built using the mean suitability values of all models with AUC > 0.7 weighted by the AUC values for each species. The models were generated and evaluated using functions from the *dismo* package (Hijmans et al., 2016) available in the R software (R Core Team, 2023).

2.6. Calculation of variables

During model evaluation in each run, we obtained the evaluation metrics (AUC and TSS) for each model. The metrics (accuracy) were then summarized for each species, where the overall mean AUC and TSS values for each species were calculated using the mean accuracy metric values of models built using the six different algorithms for each of the three different sets of variables (AEM, ENV, AEM+ENV). Thus, we ended up with overall mean AUC and TSS values for each species for each set of variables, which are hereafter used as response variables for the analysis.

The distribution range of species was calculated using the Minimum Convex Polygon (MCP) approach, which is the smallest possible polygon with straight lines that contains all occurrence records. We calculated the MCP in Km² using the function *Minimum Bounding Geometry*, which is available in the ArcGIS software version 10.5. We also calculated the number of non-duplicated records for each species to be used as the species range. However, since this variable was highly correlated with MCP values ($r = 0.87$), we kept only MCP since it better represents the distribution of poorly surveyed species (e.g., few occurrences widely distributed throughout the basin).

Finally, through the distribution maps of each species, we assessed whether the occurrence records fall within only one sub-basin (Tocantins or Araguaia) or two sub-basins (Tocantins and Araguaia). Also, for species with distributions in only one sub-basin, we determined which sub-basin it was. These variables were mainly used to control the effect of accuracy on models using species with a small range size but distributed across both sub-basins (e.g., *Acestrocephalus maculosus* – Figure S1.1 in Supplementary Material) and a large range but within only one sub-basin (e.g., *Moenkhausia tergimacula* – Figure S1.22).

2.7. Data analysis

First, we calculated an Analysis of Variance (ANOVA) to test the difference between mean AUC and TSS values of each species among the

type of variables. We used an Analysis of Covariance (ANCOVA) to test the relationship between species range and the model's evaluation metrics (AUC and TSS) and if this relationship varies according to the variables used (ENV, AEM, ENV+AEM). For these analyses, AUC and TSS were the dependent variables (continuous, varying from 0 to 1), species range was the continuous independent variable and the type of variables used for modelling was the covariate (nominal categorical) with three levels: ENV (environmental variables), AEM (asymmetric dispersal variables) and ENV+AEM (environment and asymmetric dispersal variables) (Table S1).

Subsequently, we tested whether the accuracy metrics were significantly different between AEM-based and ENV-based models by calculating delta AUC and TSS. The delta accuracy here was the mean accuracy values from the AEM models subtracted from mean accuracy values from the ENV models ($\Delta = \text{AEM} - \text{ENV}$). In this sense, AUC and TSS positive values represent models that, for those specific species, AEM generated models with higher performance in comparison to models built using ENV. On the contrary, AUC and TSS negative values show for which species the models built using ENV had a higher performance in comparison to AEM-based models. Additionally, we took those delta accuracy values (response variables) and calculated an ANOVA to test a relationship between Delta accuracy and the Number of basins each species occurs within ($\Delta_{\text{AUC}} \sim \text{Nbasins}$ and $\Delta_{\text{TSS}} \sim \text{Nbasins}$). We also calculated a linear regression between Delta accuracy and the distribution range of species. For all statistical tests, we considered p-values less than 0.05 as significant.

3. Results

Among the 29 species modelled, most species (18 species or 62%) had models with the best performance (both AUC and TSS) when modelled using AEM predictors, while 21% (six species) had better performance using ENV predictors and 17% (five species) using the combination of AEM and ENV predictors. In general, the models built for each species had good performances (Figure 2). Among the models built using only asymmetric dispersal (AEM), 24 (83%) species had models considered adequate ($\text{AUC} \geq 0.75$, $\text{TSS} \geq 0.5$). For models using environmental variables (ENV), 21 (72%) had performances above the acceptance threshold, and for models using both ENV and AEM, 23 (79%) species had models with performances above the threshold. However, the

overall difference among types of variables using the mean accuracy values of each species model was not statistically significant (AUC: $p = 0.87$; TSS: $p = 0.62$), indicating that AEM and ENV models presented similar performances (Figure 2).

The ANCOVA analyses for both evaluation metrics (AUC and TSS) showed no significant difference between types of variables (Table 1). However, the species range showed a significant relationship with both accuracy values. Therefore, models built for species with more restricted distributions showed better performance, i.e., with more realistic predictions, regardless of the type of variables used for modelling (Table 1 and Figure 3). There are no significant differences in the performance of models using different types of variables and species ranges.

The comparison of delta accuracy values (AUC and TSS) between the AEM and ENV models for each species had different but complementary results among the number of sub-basins and distribution ranges for each species. We found no relationship between range and accuracy metrics between the AEM and ENV models for both AUC ($R^2 = 0.04$, $p = 0.29$) and TSS ($R^2 = 0.04$, $p = 0.26$) metrics. However, the delta accuracy was significantly different between models for species distributed only in one sub-basin and distributed in both sub-basins for both AUC ($F_{(1, 27)} = 9.18$, $p < 0.01$) and TSS ($F_{(1, 27)} = 8.50$, $p < 0.01$) metrics (Figure 4), meaning that species distributed in only one basin modelled using asymmetric dispersal (AEM) generally had higher performance than using environmental variables (ENV), and this positive delta was significantly different from models for species distributed in both basins. That difference is especially higher for more restricted species

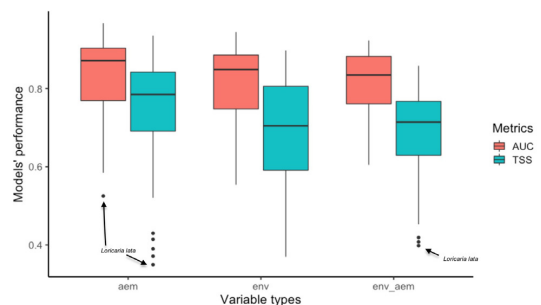


Figure 2. Difference in the models' performances using AUC (Area Under Curve) and TSS (True Skill Statistic) evaluation metrics among models built using different variables for all species (aem, asymmetric dispersal predictors, env, environmental variables, or both env_aem).

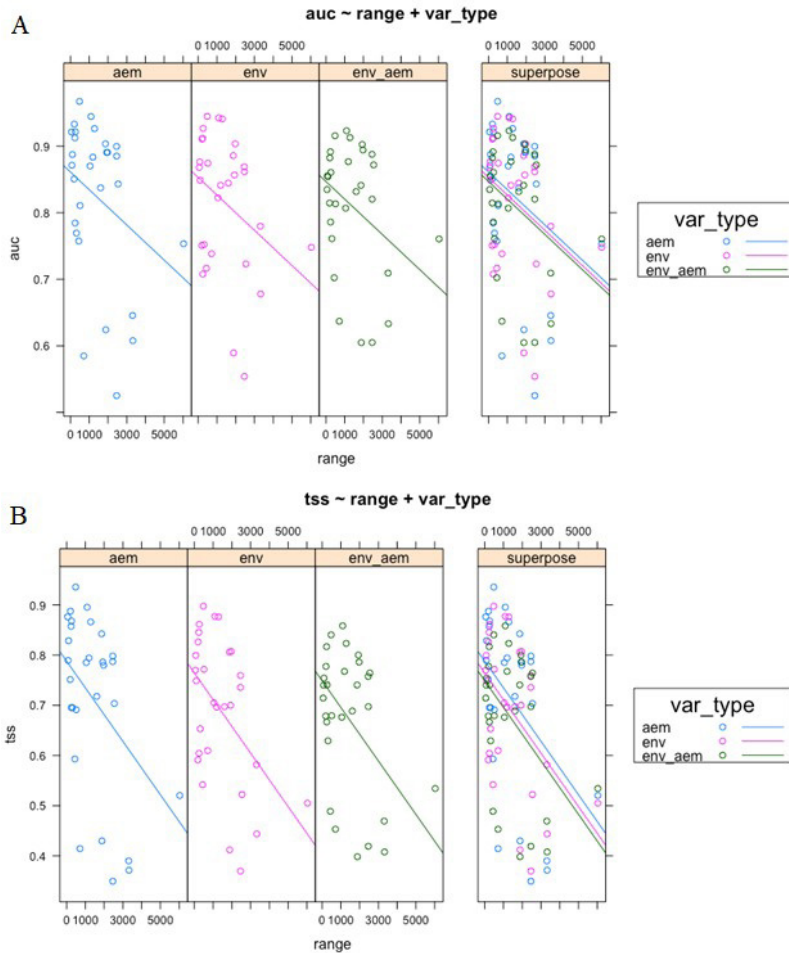


Figure 3. Relationship between A) AUC, Area Under Curve; and B) TSS True Skill Statistic; and the species range (Km²/100) among the models built using different types of variables (aem, asymmetric dispersal predictors, env, environmental variables, or both env_aem).

distributed within the same sub-basin (Figure 4) with the exception of species widely distributed latitudinally but still within the same sub-basin, such as *Moenkhausia tergimacula* (Figure S1.21), which also had a higher accuracy when modelled using AEM than ENV. Nevertheless, ENV models for each species had a generally higher performance for highly dispersed species in both basins (regardless of distribution range). For example, *Moenkhausia pyrophthalma* and *Ammoglanis diaphanus* have quite different distribution range sizes but both had a higher performance when modelled using ENV than AEM.

4. Discussion

Although not significantly different, in our study, the models' accuracies were generally higher for AEM models than ENV models. Moreover, this difference of accuracy among the built SDMs varies depending on the species range size combined

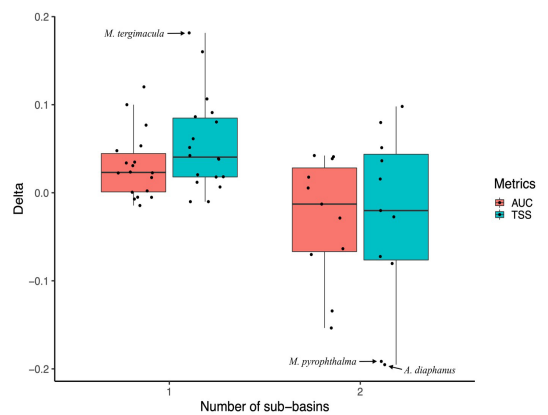


Figure 4. Delta values (Δ =AEM-ENV) of accuracy metrics (AUC, Area Under Curve, and TSS, True Skill Statistic) of all species modelled using AEM (asymmetric dispersal predictors) and ENV (environmental variables) distributed within one or two sub-basins

with the species distribution within the main sub-basins where restricted species, when modelled

using AEM, had models with higher accuracy and less overprediction to areas climatically suitable but longitudinally disconnected. Therefore, the insertion of asymmetric dispersal (AEM) predictors into freshwater SDMs, besides generating models with overall high accuracy for the fish species evaluated (higher for most species), avoids or reduces the overprediction to climatically close but, in many cases, disconnected areas that occur when using only traditional environmental variables.

In general, models built using asymmetric dispersal predictors (AEM) had higher performances than those built using only environmental predictors (ENV) and the combination of AEM and ENV predictors. This trend continues specifically for most species, having models with higher performance and above the acceptable threshold using AEM predictors. Even though this effect could be due to environmental or physical variables not measured here (e.g., water flow velocity), the inclusion of spatial predictors in species distribution models can generate accurate models by including the spatial structure (variation) of the studied area (e.g., coordinates [x and y gradients] and large- and fine-spatial filters) (Bahn & McGill, 2007; Stockwell & Peterson, 2002), spatial autocorrelation (Guélat & Kéry, 2018) and species dispersal, either symmetric through dispersal rates (Holloway et al., 2016; Monsimet et al., 2020) or asymmetric (Parreira et al., unpublished data). In this sense, the inclusion of species dispersal (movement) through asymmetric dispersal (AEM) predictors can generate more reliable and accurate models by reducing overprediction to close, and thus climatically similar but disconnected areas (e.g., rivers separated by physical barriers [natural: elevation; artificial: dams]) by including a directional gradient into the predictors that matches the hierarchy and flow direction of river basins (Blanchet et al., 2008).

Species more restricted (i.e., small range) tend to have models with higher accuracy (Stockwell & Peterson, 2002) with no overall differences in this relationship depending on the predictors used for modelling. However, the difference among predictors arises when considering the basin limitation (limited accessible areas) for species distribution. In our study, species with more restricted distributions due to their small range and distribution limited to a sub-basin had models with higher performances (accuracy). These restricted species can generate models with higher performances (accuracy) due to their historically limited accessible areas (Barve et al., 2011) with prediction accuracy decreasing as the range size increases (Stockwell & Peterson, 2002).

Furthermore, by correlating these predictors with the model's delta accuracy for each species, we found a higher and significant difference in performance between models built using only AEM and only ENV, with models built using AEM having higher accuracy, especially for restricted species. AEM-based models increase accuracy since they include the longitudinal disconnection between the main sub-basins (Araguaia and Tocantins) which act as a dispersal barrier between sub-basins; thus, including the directionality for the species dispersal into models. Hence, the insertion of asymmetric dispersal into freshwater SDMs, besides generating more accurate predictions, more realistically represent the actual distribution of continental fish through rivers, increasing the ecological reality of predictions.

One of the issues of using only environmental predictors when modelling is not considering dispersal movement (Miller & Holloway, 2015), either a priori during the modelling process (e.g., as explanatory variables) or a posteriori (e.g., by overlapping accessible and suitable areas), generating overpredictions (Mendes et al., 2020). In this study, the species most affected by overprediction when modelled using only environmental variables are species with restricted or wide distributions throughout one sub-basin but with few occurrence points in the other sub-basin geographically close to the species' core distribution. For example, both *Moenkhausia pyrophthalma* (wide distribution) and *Ammoglanis diaphanus* (restricted distribution) had quite higher performances when modelled using ENV than AEM predictors due to not considering the species dispersal route throughout the rivers to reach the geographically close and climatically suitable areas, but disconnected in the upper portion of the basin, generating overpredicted distribution when modelling these species with only environmental variables.

Therefore, the models built using environmental and dispersal-related variables show different accuracies (positive or negative deltas) depending on the species' range size and distribution across sub-basins. However, those differences in accuracy for models built using those variables for each species are small and not statistically significant for the overall difference among models. Besides, some species have quite small sample sizes due to their narrow distribution in large grid cells (i.e., small unique presences and low prevalence), which in turn generate models with poor accuracy, results already expected for accuracy in models built with small samples (Liu et al., 2019) and low prevalence (Jiménez-Valverde et al., 2009). Despite

that, AEM-based models were able to address this problem, avoiding overprediction to areas outside the restricted species core distribution, generating overall higher performance for those species with limited distribution and sample size.

We conclude that since the predicted suitability distributions and accuracy between AEM and ENV models did not vary much or were slightly higher for restricted species using asymmetric dispersal (AEM) predictors, dispersal predictors can be used as a surrogate of climate-based environmental variables or as complementary variables to insert dispersal restrictions into SDMs, aiming to reduce model overprediction. Furthermore, the use of asymmetric dispersal in SDMs should be explored further in larger-scale studies (e.g., assessing dispersal restrictions among basins) and in freshwater studies modelling species distributions in areas with anthropogenic dispersal restrictions. For example, building SDMs with asymmetric dispersal predictors enables the insertion of river disconnections due to anthropic actions, such as hydroelectric power plants (HPP) or water dams, that can be inserted as disconnections (0s in the asymmetric binary matrix) in grid cells that have built or planned HPPs, assessing this effect on the species' potential distribution. Therefore, using dispersal-related predictors in SDMs allows a better representation of the species habitat and movement, avoiding/reducing overprediction to unreachable climatically suitable areas.

Acknowledgements

We thank Fabrício Barreto Teresa for the help in obtaining the species distribution data. MRP thanks the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES – finance code 001) and FAPEG for the PhD scholarship received. Our work on SDMs has been continuously supported by different grants: CNPq, Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG/process 201710267000519), and Instituto Nacional de Ciências e Tecnologia (INCT) em Ecologia, Evolução e Conservação da Biodiversidade (EECBio) (MCTI/CNPq/FAPEG/465610/2014-5). The biodiversity analyses for Araguaia River are developed in the context of an agreement between the Tropical Water Research Alliance (TWRA) and FAPEG (proc. 202210267000536), PPBio Araguaia project supported by CNPq (proc. 441114/2023-7) and PELD Planície de Inundação do Rio Araguaia (FAPEG/CNPq). JCN was supported by the CNPq productivity fellowship (303181/2022-2).

Data availability

All material with table and figures is available in <https://doi.org/10.48331/scielodata.QSLQKM>. Access is free.

References

- Allouche, O., Steinitz, O., Rotem, D., Rosenfeld, A., & Kadmon, R., 2008. Incorporating distance constraints into species distribution models. *J. Appl. Ecol.* 45(2), 599-609. <http://doi.org/10.1111/j.1365-2664.2007.01445.x>.
- 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(6), 1223-1232. <http://doi.org/10.1111/j.1365-2664.2006.01214.x>.
- Anderson, R.P., 2013. A framework for using niche models to estimate impacts of climate change on species distributions. *Ann. N. Y. Acad. Sci.* 1297(1), 8-28. PMID:25098379. <http://doi.org/10.1111/nyas.12264>.
- Araújo, M.B., & Peterson, A.T., 2012. Uses and misuses of bioclimatic envelope modeling. *Ecology* 93(7), 1527-1539. PMID:22919900. <http://doi.org/10.1890/11-1930.1>.
- Bahn, V., & McGill, B.J., 2007. Can niche-based distribution models outperform spatial interpolation? *Glob. Ecol. Biogeogr.* 16(6), 733-742. <http://doi.org/10.1111/j.1466-8238.2007.00331.x>.
- 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. Modell.* 222(11), 1810-1819. <http://doi.org/10.1016/j.ecolmodel.2011.02.011>.
- Blanchet, F.G., Legendre, P., & Borcard, D., 2008. Modelling directional spatial processes in ecological data. *Ecol. Modell.* 215(4), 325-336. <http://doi.org/10.1016/j.ecolmodel.2008.04.001>.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45(1), 5-32. <http://dx.doi.org/10.1023/A:1010933404324>.
- Buisson, L., Thuiller, W., Casajus, N., Lek, S., & Grenouillet, G., 2010. Uncertainty in ensemble forecasting of species distribution. *Glob. Change Biol.* 16(4), 1145-1157. <http://doi.org/10.1111/j.1365-2486.2009.02000.x>.
- Carpenter, G., Gillison, A.N., & Winter, J., 1993. Domain - a flexible modeling procedure for mapping potential distributions of plants and animals. *Biodivers. Conserv.* 2(6), 667-680. <http://dx.doi.org/10.1007/BF00051966>.
- Chaine, A.S. & Clobert, J., 2012. Dispersal. In: Candolin, U., & Wong, B.B.M., eds. Behavioural responses to a changing world. Oxford: Oxford

- University Press, 63-79. <http://doi.org/10.1093/acprof:osobl/9780199602568.003.0005>.
- Clobert, J., Ims, R.A. & Rousset, F., 2004. Causes, mechanisms and consequences of dispersal. In: Hanski, I., ed. *Ecology, genetics and evolution of metapopulations*. Amsterdam: Elsevier, 307-335. <http://doi.org/10.1016/B978-012323448-3/50015-5>.
- Dray, S., Bauman, D., Blanchet, G., Borcard, D., Clappe, S., Guenard, G., Jombart, T., Larocque, G., Legendre, P., Madi, N. & Wagner, H.H., 2019. *adespatial: Multivariate Multiscale Spatial Analysis [software]*. Vienna: R Foundation for Statistical Computing.
- Elith, J., & Graham, C.H., 2009. Do they? How do they? WHY do they differ? on finding reasons for differing performances of species distribution models. *Ecography* 32(1), 66-77. <http://doi.org/10.1111/j.1600-0587.2008.05505.x>.
- Franklin, J., Wejnert, K.E., Hathaway, S.A., Rochester, C.J., & Fisher, R.N., 2009. Effect of species rarity on the accuracy of species distribution models for reptiles and amphibians in southern California. *Divers. Distrib.* 15(1), 167-177. <http://doi.org/10.1111/j.1472-4642.2008.00536.x>.
- Guélat, J., & Kéry, M., 2018. Effects of spatial autocorrelation and imperfect detection on species distribution models. *Methods Ecol. Evol.* 9(6), 1614-1625. <http://doi.org/10.1111/2041-210X.12983>.
- Guisan, A., & Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecol. Modell.* 135(2-3), 147-186. [http://doi.org/10.1016/S0304-3800\(00\)00354-9](http://doi.org/10.1016/S0304-3800(00)00354-9).
- Hernandez, P.A., Graham, C.H., Master, L.L., & Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography* 29(5), 773-785. <http://doi.org/10.1111/j.0906-7590.2006.04700.x>.
- Hijmans, R.J. 2019. *raster: geographic data analysis and modeling [software]*. Vienna: R Foundation for Statistical Computing.
- Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., & Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* 25(15), 1965-1978. <http://doi.org/10.1002/joc.1276>.
- Hijmans, R.J., Phillips, S.J., Leathwick, J.R. & Hortal, J., 2016. *dismo: Species Distribution Modeling. R package version 1.1-1 [software]*. Vienna: R Foundation for Statistical Computing.
- Holloway, P., Miller, J.A., & Gillings, S., 2016. Incorporating movement in species distribution models: how do simulations of dispersal affect the accuracy and uncertainty of projections? *Int. J. Geogr. Inf. Sci.* 30, 1-25. <http://doi.org/10.1080/13658816.2016.1158823>.
- Jackson, D.A., 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* 74(8), 2204-2214. <http://doi.org/10.2307/1939574>.
- Jiménez-Valverde, A., Lobo, J., & Hortal, J., 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. *Community Ecol.* 10(2), 196-205. <http://doi.org/10.1556/ComEc.10.2009.2.9>.
- Jiménez-Valverde, A., Lobo, J.M., & Hortal, J., 2008. Not as good as they seem: the importance of concepts in species distribution modelling. *Divers. Distrib.* 14(6), 885-890. <http://doi.org/10.1111/j.1472-4642.2008.00496.x>.
- Liu, C., Newell, G., & White, M., 2019. The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites. *Ecography* 42(3), 535-548. <http://doi.org/10.1111/ecog.03188>.
- Marco Júnior, P., Diniz-Filho, J.A.F., & Bini, L.M., 2008. Spatial analysis improves species distribution modelling during range expansion. *Biol. Lett.* 4(5), 577-580. PMID:18664417. <http://doi.org/10.1098/rsbl.2008.0210>.
- Mendes, P., Velazco, S.J.E., Andrade, A.F.A., & Marco Júnior, P., 2020. Dealing with overprediction in species distribution models: how adding distance constraints can improve model accuracy. *Ecol. Modell.* 431, 109180. <http://doi.org/10.1016/j.ecolmodel.2020.109180>.
- Miller, J.A., & Holloway, P., 2015. Incorporating movement in species distribution models. *Prog. Phys. Geogr.* 39(6), 837-849. <http://doi.org/10.1177/0309133315580890>.
- Monsimet, J., Devineau, O., Pétillon, J., & Lafage, D., 2020. Explicit integration of dispersal-related metrics improves predictions of SDM in predatory arthropods. *Sci. Rep.* 10(1), 16668. PMID:33028838. <http://doi.org/10.1038/s41598-020-73262-2>.
- Nelder, J.A., & Wedderburn, R.W.M., 1972. Generalized linear models. *J. R. Stat. Soc. Ser. A Stat. Soc.* 135(3), 370-384. <http://dx.doi.org/10.2307/2344614>.
- Nix, H.A., 1986. A biogeographic analysis of Australian elapid snakes. In: Longmore, R., ed. *Atlas of elapid snakes of Australia*. Canberra: Australian Government Publishing Service, 4-15.
- Padial, A.A., Ceschin, F., Declerck, S.A.J., De Meester, L., Bonecker, C.C., Lansac-Tôha, F.A., Rodrigues, L., Rodrigues, L.C., Train, S., Velho, L.F.M., & Bini, L.M., 2014. Dispersal ability determines the role of environmental, spatial and temporal drivers of metacommunity structure. *PLoS One* 9(10), e111227. PMID:25340577. <http://doi.org/10.1371/journal.pone.0111227>.
- Parreira, M.R., Tessarolo, G., & Nabout, J.C., 2023. Incorporating symmetrical and asymmetrical dispersal into Ecological Niche Models in freshwater

- environments. *Acta Limnol. Bras.* 35, e16. <http://doi.org/10.1590/s2179-975x2723>.
- Perrin, S.W., Englund, G., Blumentrath, S., Brian O'Hara, R., Amundsen, P., & Gravbrøt Finstad, A., 2020. Integrating dispersal along freshwater ecosystems into species distribution models. *Divers. Distrib.* 26(11), 1598-1611. <http://doi.org/10.1111/ddi.13112>.
- Peterson, A.T., & Soberón, J., 2012. Species distribution modeling and ecological niche modeling: getting the concepts right. *Nat. Conserv.* 10(2), 102-107. <http://doi.org/10.4322/natcon.2012.019>.
- 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: Princeton University Press. <http://doi.org/10.23943/princeton/9780691136868.001.0001>.
- Phillips, S.J., Anderson, R.P., & Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. *Ecol. Modell.* 190(3-4), 231-259. <http://dx.doi.org/10.1016/j.ecolmodel.2005.03.026>.
- R Core Team, 2023. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
- Rangel, T.F., & Loyola, R.D., 2012. Labeling Ecological Niche Models. *Nat. Conserv.* 10(2), 119-126. <http://dx.doi.org/10.4322/natcon.2012.030>.
- Revelle, W. 2019. *psych: Procedures for Personality and Psychological Research [software]*. Vienna: R Foundation for Statistical Computing.
- Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., & Williamson, R.C., 2001. Estimating the support of a high-dimensional distribution. *Neural Comput.* 13(7), 1443-1471. PMID:11440593. <http://dx.doi.org/10.1162/089976601750264965>.
- Segurado, P., Araújo, M.B., & Kunin, W.E., 2006. Consequences of spatial autocorrelation for niche-based models. *J. Appl. Ecol.* 43(3), 433-444. <http://doi.org/10.1111/j.1365-2664.2006.01162.x>.
- Soberón, J., 2007. Grinnellian and Eltonian niches and geographic distributions of species. *Ecol. Lett.* 10(12), 1115-1123. PMID:17850335. <http://doi.org/10.1111/j.1461-0248.2007.01107.x>.
- Soberon, J., & Peterson, A.T., 2005. Interpretation of models of fundamental ecological niches and species'. *Biodivers. Inform.* 2(0), 1-10. <http://doi.org/10.17161/bi.v2i0.4>.
- Soliman, T., Mourits, M.C.M., van der Werf, W., Hengeveld, G.M., Robinet, C., & Lansink, A.G.J.M.O., 2012. Framework for modelling economic impacts of invasive species, applied to pine wood nematode in Europe. *PLoS One* 7(9), e45505. PMID:23029059. <http://doi.org/10.1371/journal.pone.0045505>.
- Stockwell, D.R.B., & Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. *Ecol. Modell.* 148(1), 1-13. [http://doi.org/10.1016/S0304-3800\(01\)00388-X](http://doi.org/10.1016/S0304-3800(01)00388-X).
- Swets, J.A., 1988. Measuring the accuracy of diagnostic systems. *Science* 240(4857), 1285-1293. PMID:3287615. <http://dx.doi.org/10.1126/science.3287615>.
- Tonkin, J.D., Altermatt, F., Finn, D.S., Heino, J., Olden, J.D., Pauls, S.U., & Lytle, D.A., 2018. The role of dispersal in river network metacommunities: patterns, processes, and pathways. *Freshw. Biol.* 63(1), 141-163. <http://doi.org/10.1111/fwb.13037>.
- Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, 2008. Effects of sample size on the performance of species distribution models. *Divers. Distrib.* 14(5), 763-773. <http://doi.org/10.1111/j.1472-4642.2008.00482.x>.

Received: 20 September 2024

Accepted: 20 June 2025

Associate Editor: Ronaldo Angelini.