Freshwater Biology

Freshwater Biology (2014) **59,** 2303–2314 doi:10.1111/fwb.12432

Evaluating the use of macroscale variables as proxies for local aquatic variables and to model stream fish distributions

RENATA G. FREDERICO*,†, PAULO DE MARCO JR† AND JANSEN ZUANON‡

*Programa de Pós-Graduação em Ecologia, Instituto Nacional de Pesquisas da Amazônia (INPA), Manaus, AM, Brazil

†Laboratório de Ecologia Teórica e Síntese, Universidade Federal de Goiás (UFG), Goiânia, Brazil

‡Coordenação de Biodiversidade, Instituto Nacional de Pesquisas da Amazônia (INPA), Manaus, AM, Brazil

SUMMARY

1. The geographical ranges of species are influenced by three components: spatial distribution of environmental conditions, biotic interactions and the dispersal capacity of species. The scarcity of distributional records in vast regions such as the Amazon impedes understanding of fish distribution. Predictive distribution models have emerged as a better alternative to surpass this problem, but the absence of large-scale maps for aquatic variables has been suggested as an important limitation.
2. We aimed to evaluate the use of macroclimatic variables as surrogates for local limnological variables in the Brazilian Amazon. Ordinary least squares model were used to predict the local habitat variables from climatic and geomorphological information as macroscale variables. Models for six stream-dwelling fish were built in MaxEnt and validated using area under curve and true skill statistics (TSS).
3. All local variables were predicted successfully ($R^2 > 0.39$), and MaxEnt models had good suitability using the macroscale variables (TSS higher than 0.70). We conclude that macroscale variables can be effective surrogates for local habitat variables, at least for large-scale analyses on poorly sampled regions such as the Brazilian Amazon.

Keywords: Amazon, aquatic variables, macroscale variables, modelling species distribution, stream fish

Introduction

Current understanding of the factors that determine the geographical ranges of species is focused on three main factors, which determine the sites that are accessible for colonisation (Soberón, 2007): (i) the spatial distribution of environmental conditions that determine the survival of individuals and the persistence of populations; (ii) biotic interactions and the dynamics of resources; and (iii) the dispersal capacity of species. This general conceptual model has been used to predict species distributions for a variety of different problems, including the conservation biogeography of individuals or groups of species (Esselman & Allan, 2011; Nóbrega & De Marco Jr, 2011; Rodríguez-Soto *et al*., 2011), the identification of potential areas of invasion by non-native species (Mata *et al*., 2010) and the possible response of species to climate change (Araújo *et al*., 2005; Araújo, Thuiller & Pearson, 2006; Buisson *et al*., 2010). The methods developed under this approach have captured the attention of many researchers, especially in areas where large gaps in biogeographical information have hindered conservation efforts (Siqueira *et al*., 2009), such as in the Brazilian Amazon (Buermann *et al*., 2008).

Species distribution modelling (SDM) is based on the Grinnellian niche (Soberón, 2007), which is defined as the set of climatic variables at broad resolution scales (hereafter macroscale variables) that determine the distribution of organisms. This view suggests that the dominant filters determining potential areas of occurrence for a given species are macroscale variables, which may include topographical/geomorphological informa-

Correspondence: Renata G. Frederico, Programa de Pós-Graduação em Ecologia, Instituto Nacional de Pesquisas da Amazônia (INPA), Av. André Araújo, 2936 – Petrópolis, CEP – 69080-971, Manaus, AM, Brazil. E-mail: renatafrederico@gmail.com

tion. Otherwise, the availability of many macroscale variables at higher resolutions (e.g. digital maps with 1 km^2 to 90 m^2 pixel resolutions; www.dpi.inpe.br/Ambdata/; www.worldclim.org) has favoured the development of SDM using the ecological niche modelling approach (Peterson & Soberón, 2012). The past 10 years have experienced an increase in studies on predictive modelling for the distribution of fish and other aquatic organisms. There are studies focusing on priority areas for conservation (Argent *et al*., 2003; Filipe *et al*., 2004; Esselman & Allan, 2011), native species conservation (Domínguez-Domínguez *et al*., 2006), non-native species distribution (Kornis & Zanden, 2010) and the testing and comparison of modelling methods (Olden & Jackson, 2002; McNyset, 2005; Oakes *et al*., 2005). Usually, those models are based on macroscale variables, including mean annual temperature, annual precipitation and topographic variables such as altitude (Argent *et al*., 2003; McNyset, 2005; Oakes *et al*., 2005; Domínguez-Domínguez *et al*., 2006). However, several studies have shown that macroscale variables can affect fish species composition in local assemblages (Matthews & Matthews, 2000; Oberdorff *et al*., 2001; Kennard *et al*., 2007; Torrente-Vilara *et al*., 2011). Several studies suggest that models built using macroscale variables perform similar to models based on local variables for aquatic species distributions (Kruse, Hubert & Rahel, 1997; Watson & Hillman, 1997; Porter, Rosenfeld & Parkinson, 2000). For instance, Kruse *et al.* (1997) found that geomorphologic variables such as slope, altitude and stream width were good predictors of the distribution of trout in Yellowstone Park. Similarly, Watson & Hillman (1997) found landscape and catchment variables to be the most important predictors of the distribution and abundance of bull trout in the western U.S.A. The importance of those macroscale variables in determining species distribution in streams and rivers may arise from the basic arrangement of the landscapes as a continuum of heterogeneous and hierarchical networks shaped by topography, geology and climatic factors (Fausch *et al*., 2002). Thus, fish distribution and assemblage structure are driven by factors acting at multiple spatial scales (Kennard *et al*., 2007; Stewart-Koster *et al*., 2013).

Some studies have discussed the lack of local environmental variables as a significant drawback in SDM of aquatic organisms. The models for some species had lower performance, possibly because they are general when compared with macroscale (climatic) variables (Oakes *et al*., 2005). In such cases, the inclusion of information regarding the water systems at local scales, such as pH and conductivity, could decrease commission errors of the models (Domínguez-Domínguez *et al*., 2006). For instance, fish from central Amazonia are known to be affected by local variables such as stream dimension, pH, conductivity and dissolved oxygen at local scales (Mendonça, Magnusson & Zuanon, 2005). The use of local variables for SDM is hindered because large extension maps of local limnological variables, such as pH or conductivity, are rarely available. For instance, the available grid of monitored limnological variables for the Brazilian Amazon has extensive gaps in its spatial distribution (http://hidroweb.ana.gov.br), precluding the use of spatial interpolation to provide mapping of these variables. Nevertheless, some studies have also demonstrated the possibility of inferring local environmental conditions based on similar characteristics at the drainage basin scale (Davies, Norris & Thoms, 2000; Mugodo *et al*., 2006). Thus, the existence of a strong relationship between local variables and macroscale variables could allow the latter to be used as effective surrogates for the local limnological variables and the ecological processes that those variables may represent. If properly demonstrated, this may also improve the accuracy of SDM models, especially in areas where local environmental variables are unavailable. We therefore tested the assumption that macroscale variables are good proxies for local limnological conditions in one of the more heterogeneous aquatic systems of the world, the Brazilian Amazon basin. We first present support for macroscale variables as surrogates for local variables using a series of regression analyses. We then use those macroscale variables to model the distributions of six fish species and evaluate the accuracy of the resulting models.

Methods

Environmental variables

The Amazon basin is the largest freshwater aquatic system in the world, occupying *c.* 7 million km^2 with *c.* 70% of this area located in Brazil. Pristine rainforests broadly cover one-third to one-half of the floodplains of rivers and streams (Goulding, Barthem & Ferreira, 2003), which are located under the shade of the forest canopy. The water types in the Amazon depend on the geological origins of the rivers and are generally defined as *clear waters*, found in water courses draining the old Guyana and Central Brazilian shields; *black waters*, originating in forested podzols and coloured by humic acids; and *white waters*, originating in the geologically recent Andean terrains and subjected to increasing intemper-

ism (Sioli, 1985). These main water types have different limnological characteristics and represent the broadest variation in the aquatic environments in the Amazon.

The local variables used in this study refer to limnological factors, including pH, conductivity, dissolved oxygen and water temperature. Data for these variables were obtained from the Brazil National Water Agency (ANA), the 'Brasil das Águas' Project (http://www.brasildasaguas.com.br/) and the Igarapés Project (J. Zuanon, unpubl. data; www.igarapes.bio.br). The data set is derived from 600 sampling locations from first-order streams to very large (14th order) rivers (Fig. 1). The ANA's database comprises the hydrological information system from hydrometeorological monitoring stations. In the Amazon, these data are restricted to 4th order and above rivers (http://hidroweb.ana.gov.br/). The sampling of 'Brasil das Águas' Project was conducted in 524 Brazilian rivers restricted to watercourses above the 3rd order, with more than one standardised sampling for several (but not all) rivers due to river size variation. The Igarapés Project has *c.* 380 standardised sampling sites in the Brazilian Amazon, from 1st to 3rd order streams. After combining the sampling points, we used Moran's I to test the spatial autocorrelation of variables (Table S3).

Macroscale variables at 1 km^2 spatial resolution were obtained from AMBDATA (Environmental variables for modelling species distribution), a database from the Brazil National Institute for Space Research (INPE) (www.dpi.inpe.br/Ambdata), which includes all of the Brazilian Amazon surfaces. Vegetation characteristics and soil type were obtained from maps of 1 : 250 000 and 1 : 5 000 000, respectively, and then converted to raster format with the same spatial resolution as the other variables. These differences in the scale of the latter variables did not cause any substantial effect on the analytical procedures. The conversion to 1 km^2 resolution will only produce a distribution of equal values for contiguous cells in the final raster for the variables with coarse resolution.

The variables were chosen by their potential to influence aquatic ecosystems and are summarised in Table 1. The annual mean precipitation and annual mean temperature were included to represent the variation in the hydrological periods in the Amazon. The percentage of vegetation cover and river order may show the variation in allochthonous and autochthonous inputs to the aquatic ecosystem.

The terrain slope, soil type and vegetation characteristics are correlated with the amount of nutrients and sediments in streams and rivers and determine the three main water types in the Amazon basin, as described above (Angermeier & Karr, 1983; Sioli, 1985; Goulding *et al.*, 2003).

Species distribution data

The Igarapés Project has *c.* 380 sampling locations distributed throughout the Brazilian Amazon, and *c.* 450 fish species were collected and sampled using standardised methodology. The fish were collected with hand nets and small seine nets in a 50-m stream stretch after blocking it with fine-mesh nets (see Mendonça *et al.*,

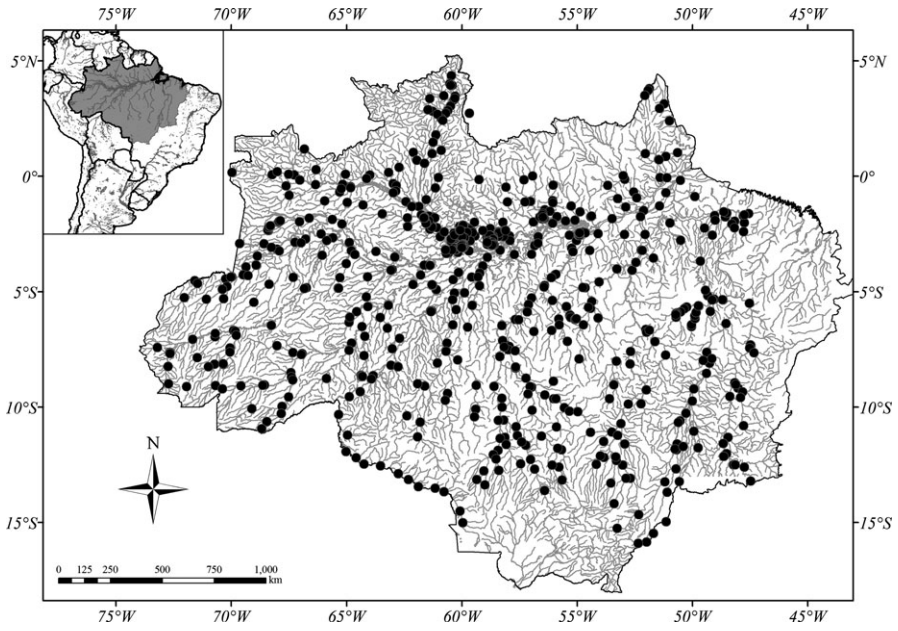

Fig. 1 Sampling points for local limnological variables in the Brazilian Amazon basin.

Table 1 Local and macroscale variables used in local variable model analysis

Macroscale variables	Local variables
Annual mean temperature (AMT)	pH
Annual mean precipitation (AMP)	Conductivity (Cond)
Terrain slope (SL)	Dissolved oxygen (OD)
Vegetation cover percentage (VC)	Water temperature (Temp)
Vegetation type (Veg)	
Soil type (Soil)	
River order (Ord)	

2005 for details). Six species were chosen from the Igarapés Project database due to the availability of additional biogeographical information from the Species Link website (http://splink.cria.org.br/). We chose only species that had more than 60 spatially unique points of occurrence. *Copella nattereri* and *Pyrrhulina* cf. *brevis* are small Lebiasinidae species very common and abundant in Amazonian streams. The distribution of *C. nattereri* is known to encompass a large portion of the Amazon basin, from the north-western portion to the lower Amazon, including the Negro River and the upper Orinoco River. *Pyrrhulina* cf. *brevis* is found throughout the Amazon basin and occurs with *Erythrinus erythrinus* and *Hoplias malabaricus* (both Erythrinidae). *Erythrinus erythrinus* is commonly found in small streams, whereas *H. malabaricus* predominates in lakes and large rivers (Oyakawa, 2003). *Helogenes marmoratus* (Cetopsidae) is a small catfish species usually found in streams distributed throughout the Amazon basin (Vari & Ferraris, 2003) and feeds on allochthonous insects. *Carnegiella strigata* (Gasteropelecidae) is known to occur throughout the Negro River as well as in small-sized tributaries in the upper and middle parts of the Amazon River basin (Weitzman & Palmer, 2003).

Statistical analysis and modelling procedures

We performed the variance inflation factor (VIF) for all of the regression models to evaluate multicollinearity of all of the predictors. This test shows which correlated variable is inflating the model regression, with a VFI higher than five indicating collinearity (Zuur *et al.*, 2009). We removed one variable at a time and recalculated the VIF until values were <3 (Appendix S1, Table S1). After VIF analysis, we used a Spearman's rank correlation matrix to estimate macroscale variables to evaluate the collinearity among the predictors, allowing nonlinear relationships. We excluded variables that present rank correlations larger than 0.6 (Appendix S1, Table S2). To generate the predictions for limnological variables based on independent macroscale variables, we used the ordinary least squares model (OLS) in the Leaps package in R software (R Development Core Team 2012). We used a model selection approach based on the Bayesian information criterion (BIC) and an exhaustive procedure to select the best set of predictor variables. In addition, the accuracy of each best limnological model in predicting local variable values was measured by the regression R^2 values, with predicted values (model value results) versus observed values (limnological value results). It is important to note that our primary concern is the predictive power of the set of macroscale variables regarding limnological variables and not its explanatory power to select which single variable could explain the variation in water parameters. In fact, we consider macroscale variables to be only proxies of unobserved processes that affect local environmental water parameters.

We used the MaxEnt programme for niche modelling. MaxEnt works with presence (occurrence) data only and categorical information. This programme estimates the probability of distribution by fitting a function close to the uniform distribution (the probability of maximum entropy) under the restriction of environmental information associated with the occurrence points (Phillips, Anderson & Schapire, 2006). The methods are based on discriminating between the environmental variables of the presence data and the background variation in the environmental variables sampled from 10 000 random background (Phillips & Dudík, 2008). One important problem mentioned in recent studies is the low transferability in studies with a high number of occurrence points (Peterson, Papes & Eaton, 2007), explained as a consequence of the higher number of parameters in the resulting models. To circumvent this problem, we controlled the number of parameters in the resulting models by restricting the MaxEnt to linear and quadratic features, similar to Elith, Kearney & Phillips (2010).

Model evaluation involves predicting niche suitability and comparing the results with a subset of the observed occurrence points that were not included in model training (test subset). We randomly divided the occurrence points into two subsets (subset A and B) for an independent evaluation of the models. We produced a MaxEnt prediction of the distribution for one subset and used the other to estimate model evaluation measures. Model evaluation was performed with both threshold-independent (e.g. area under the curve – AUC) and threshold-dependent (e.g. true skill statistics – TSS) measures, as recently suggested (Liu, White & Newell, 2011). The AUC is obtained by plotting true positives (sensitive val-

ues) against false-positive values (1-specificity), which is known as the receiver operating characteristic (ROC) curve (Fielding & Bell, 1997). The AUC can then be interpreted as an average sensitivity of all possible values of specificity, producing a global measure of fit for the model. Nevertheless, this method may be subject to problems related to the use of prevalence in the model (Lobo, Jiménez-Valverde & Real, 2008). TSS is less sensitive to prevalence and represents the average rate of prediction success as values varying between −1 and 1 (Liu, White & Newell, 2009). As our approach is entirely based on the presence data, TSS and AUC are calculated using background data as pseudo-absence, a common procedure in recent SDM literature (see Liu *et al*., 2009).

Deriving presence and absence from these models requires determination of thresholds that are expected to minimise omission and commission errors. The threshold derived from the ROC is the point that minimises both errors and is considered an equilibrium choice, minimum difference threshold criterion (Jiménez-Valverde & Lobo, 2007). The alternative to ROC is the low presence training (LPT), which minimises only the omission errors and is used mostly for species with a small number of distribution records (Pearson *et al*., 2006). As observed above, the TSS is dependent on the threshold choice, and we used both measures to test the accuracy of the models.

Results

Environmental variables model

First- to third-order streams yielded low values for all local variables (water temperature, conductivity, dissolved oxygen and pH), as expected (Fig. 2). These lower order streams are not flood plains and have a higher slope and altitude and dense canopy cover.

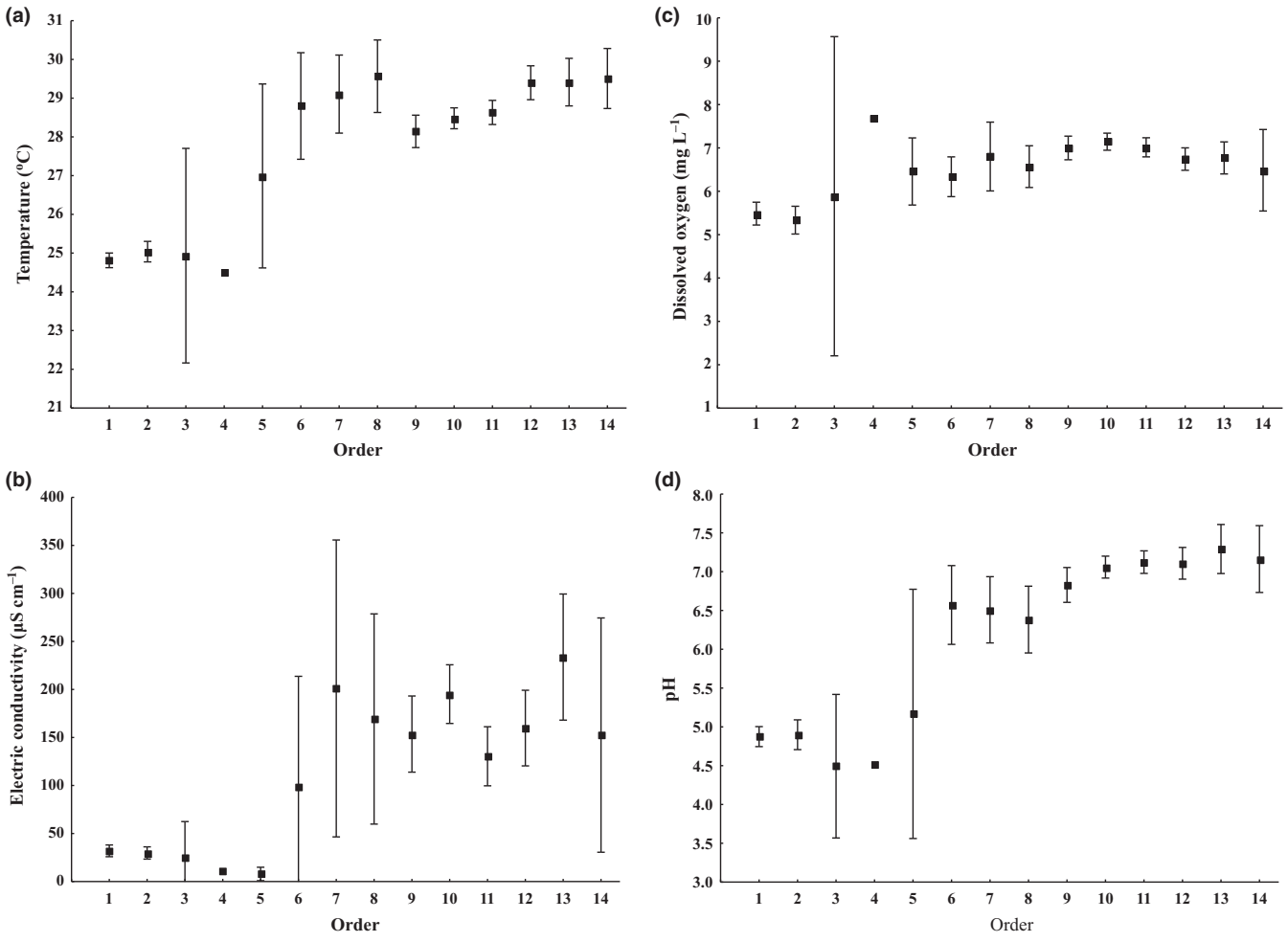

Fig. 2 Means (±standard deviation) in relation to river order for local limnological variables (a) temperature, (b) conductivity, (c) dissolved oxygen and d) pH.

Third-order streams showed the highest variance, most probably due to the few available sampling points; 4th-order streams are represented by a single sampling point. Streams from the 5th to 8th orders also showed high variance, indicating large variation in limnological characteristics (e.g. differences in water types) among samples. Water temperature was high in streams and rivers of the 5th order and higher, as was conductivity (although with a higher variability) in the 6th order and higher rivers (Fig. 2a,b, respectively). Dissolved oxygen values were almost uniformly high among the river orders, with a pronounced variance in 3rd-order streams (Fig. 2c). The pH showed similar patterns as temperature and conductivity, with a marked increase above the 5th order (Fig. 2d). In general, rivers above the 6th order became limnologically more homogeneous because the increase in width, discharge and runoff lead to greater buffering of the aquatic system in relation to the variability of local factors influencing the catchment areas of headwater streams (Fig. 2).

Our general statement that macroscale variables could predict local environmental variables is supported by the good general fit of the models (R^2 values ranging from 0.53 to 0.79, Table 2). Among the modelled local variables, pH presents the best relationship between the observed and predicted values, with 79% of the variation explained by the model (Table 2, Fig. 3a). For this variable, the best model (BIC = −650) included annual mean precipitation, river order, soil type and vegetation type as component variables (Table 2). The second-best relationship was obtained for water temperature (R^2 = 0.76 and BIC = −590, Table 2, Fig. 3b); in this case, the variables included in the best model were river order, soil type and vegetation type. Conductivity was explained by stream order, soil type and vegetation type (R^2 = 0.60 and BIC = −310, Table 2, Fig. 3c). The model with the lowest relationship between the observed and predicted values was dissolved oxygen (BIC = −240 and R^2 = 0.53, Table 2, Fig. 3d). It is important to note that river order, soil and vegetation types were present in all models and annual mean precipitation in only one model.

Modelling the distribution of Amazonian stream fish

The AUC values for the model subsets A and B for each species were very similar. TSS had more variation between species (0.68–0.87). Despite these differences, both indices indicate that the models produced good predictions (Table 3). The species whose occurrence points were concentrated in the central Amazon (*C. nattereri*, *P. cf. brevis* and *H. marmoratus*) showed high values for AUC and TSS (Table 3, Fig. 4), and *C. nattereri* had very similar models for the subsets A and B, for both the ROC and LPT thresholds (Fig. 4b). The models also indicated potential areas of occurrence outside the central Amazon. *Carnegiella strigata* had occurrence points that were more concentrated in streams of the western Brazilian Amazon but demonstrated results generally similar to the previous species with high AUC and TSS values and similar models for the LPT and ROC thresholds in model A (Fig. 4a). *Erythrinus erythrinus* and *H. malabaricus*, which had broad distributions in the Amazon, had the lowest AUC and TSS values. The models (especially for the LPT threshold) indicate a very large potential distribution area for those species (Fig. 4e,f, respectively). Considering the six species, the ability of the models to predict distribution points among data subsets was high (0.53–1.0).

Discussion

Macroscale variables as surrogates for local aquatic conditions

The local variables studied here are all affected by the historical formation of the Amazon basin and are also strongly dependent on its surrounding forest. The soil in the Amazon is overall of low fertility, yet its vegetation cover is the main source of energy and nutrients for the

Table 2 Best models and macroscale predictors for each local variable

	Annual mean precipitation	Annual mean temperature	Vegetation cover	Slope	Soil	Vegetation type	River order	BIC	R^2
pH	X				X	X	X	**−650**	**0.79**
Water temperature					X	X	X	**−590**	**0.76**
Conductivity					X	X	X	**−310**	**0.60**
Dissolved oxygen					X	X	X	**−240**	**0.53**

The Bayesian information criterion (BIC) and R^2 for the best models are shown in bold.

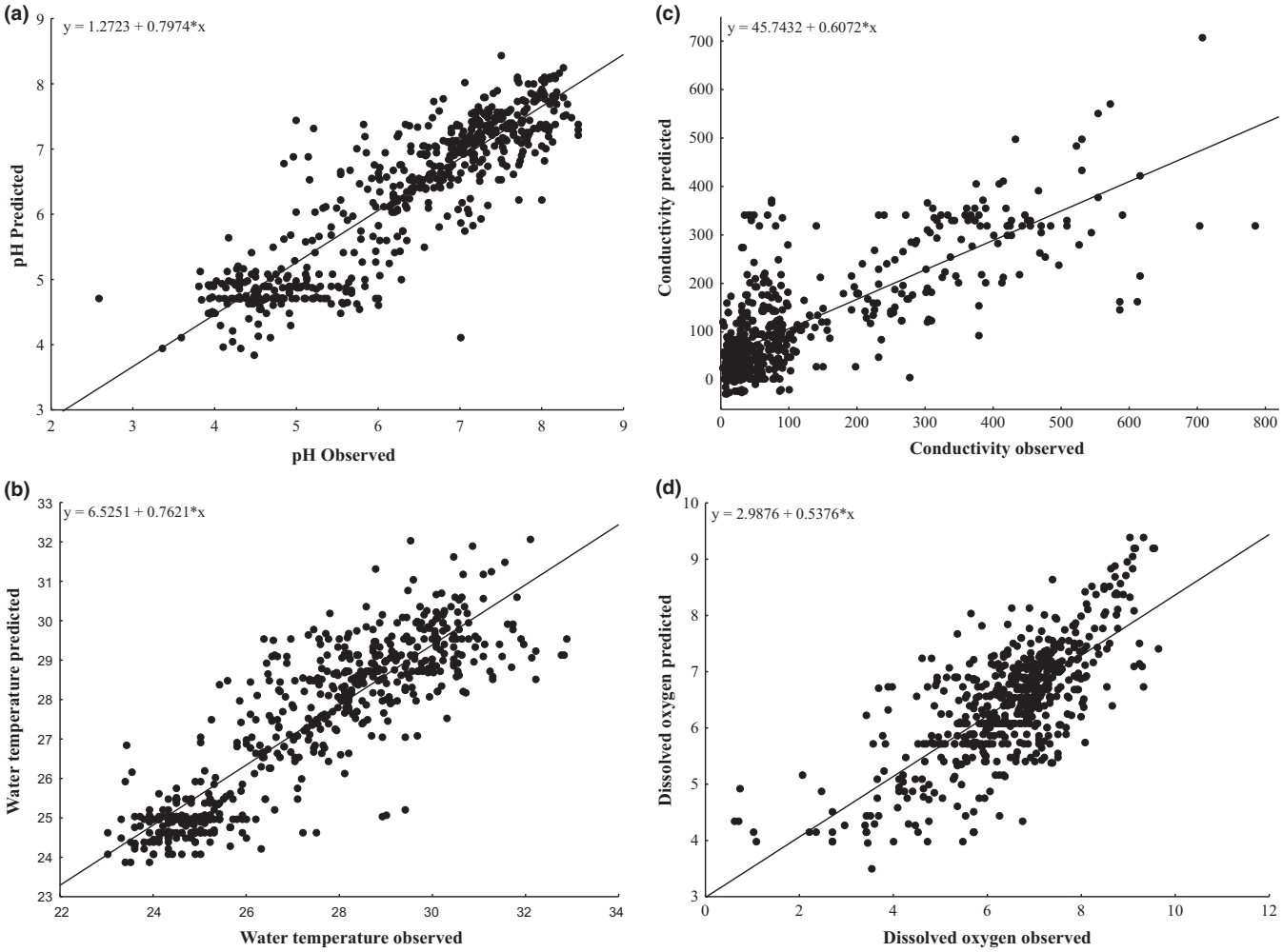

Fig. 3 Relationship between predicted (best model) and observed local variable values for (a) pH, (b) water temperature, (c) conductivity and (d) dissolved oxygen.

Table 3 Values of AUC and TSS for models A and B of each species and cross-validation with LPT and ROC

Species	AUC		TSS		A–B ROC/LPT	B–A ROC/LPT
	Model A	Model B	Model A	Model B		
Carnegiella strigata	0.943	0.944	0.803	0.803	0.53/0.78	0.78/0.78
Copella nattereri	0.956	0.958	0.843	0.851	1.00/1.00	1.00/1.00
Erythrinus erythrinus	0.910	0.924	0.680	0.732	0.71/0.95	0.85/0.95
Helogenes marmoratus	0.948	0.964	0.824	0.813	0.76/0.88	0.91/0.91
Hoplias malabaricus	0.920	0.899	0.693	0.701	0.53/0.83	0.61/0.85
Pyrrhulina cf. *brevis*	0.962	0.972	0.838	0.873	0.83/0.90	0.89/0.97

AUC, area under the curve; LPT, low presence training; TSS, true skill statistics.
A–B, ROC and LPT for the subset points A and in model B; B–A, ROC and LPT for the subset points B and in model A.

aquatic environment (Sioli, 1985; Goulding *et al.*, 2003). The organic matter inputs to rivers and streams originate from upper and lateral stream catchments and depend on the riparian vegetation and the connection to floodplain areas. The characteristics of the riparian vegetation can differ along the river and catchment area, which could also affect water quality (Kawaguchi, Taniguchi & Nakano, 2003; Dudgeon, 2008). Many observed differences between the limnological conditions of different rivers are associated with catchment

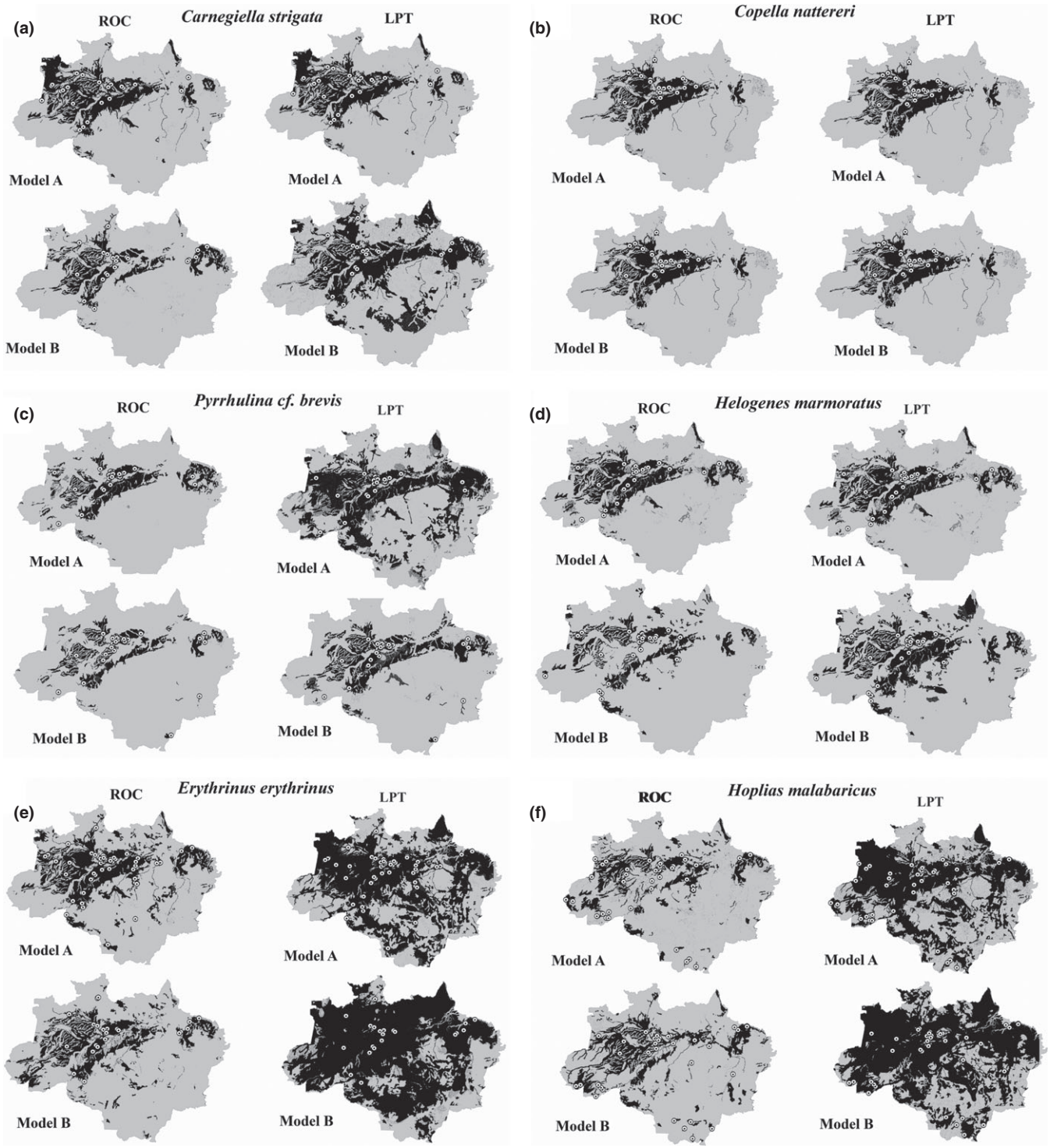

Fig. 4 Predictive models for the distribution of fish species using MaxEnt. Models A and B refer to the partition of training data for modelling (subset points A and B). ROC and low presence training (LPT) were the thresholds used for the most suitable areas. The black areas were niche areas predicted for the model, and the points are the occurrence points used to build the models.

geomorphology. The soil type could be a surrogate for geomorphological conditions and an important variable to distinguish between subregions in the Amazon, as was evident from our results. Factors such as soil type, geological formation and vegetation type could change the structure, function and quality of these waterbodies,

which explains the observed relationship with local variables at broad scales. Limnological condition in the Amazon basin is strongly related to its geological formation, resulting in the different water types. Finally, river size (represented by its hierarchical order) is another variable that reflects the structure of the aquatic ecosystem; as order increases, the system changes from almost exclusively allochthonous sources of energy and nutrients to autochthonous sources of energy, with an increase in the importance of aquatic primary productivity to the structure of its food chains (Vannote *et al.*, 1980). First-order forest rivers are allochthonous systems, closely dependent on the surrounding vegetation as a source of energy and nutrients (Kawaguchi *et al.*, 2003). Otherwise, as river width increased with increasing order, the importance of direct sunlight to the ecosystem also increases, and the system becomes autochthonous (Vannote *et al.*, 1980). However, local aquatic conditions are also affected by many other factors acting at small spatial scales, such as soil deposition, canopy cover and stream width and depth, especially in small streams. Therefore, our results suggest that macroscale variables could capture the effect of local aquatic variables and thus allow the application of the current general understanding of the Grinnellian niche (Soberón, 2007) to estimate distribution maps for Amazonian stream fishes. It is important to highlight that most variables selected by the OLS models are categorical. This could cause limitations in the choice of the SDM technique because only MaxEnt (Phillips & Dudík, 2008) and GLM based-techniques (Elith & Graham, 2009; Syphard & Franklin, 2009) were specially designed to handle categorical variables.

In a study conducted in the upper Murrumbidgee River in Australia, local stream characteristics were successfully predicted using macroscale information on the geology, alkalinity and catchment area (Davies *et al.*, 2000). A similar study performed on three hydrographic basins in southwest Queensland, Australia (Mugodo *et al.*, 2006), using river order, longitude, source distance and altitude, predicted five local habitat variables with R^2 values >0.2 between the observed and predicted variables. Our results showed that macroscale variables could provide even higher predictive power for limnological parameters (R^2 always higher than 0.53) in the Amazon. Previous studies have discussed the importance of geomorphologic and landscape variables that reflect the catchment structure in species distribution of fish (Kruse *et al.*, 1997; Watson & Hillman, 1997). Comparing the SDM of fish species using solely large-scale variables versus large-scale variables and local variables, Porter *et al.* (2000) showed that large-scale variables adequately predict fish species distributions in the Blackwater Drainage in Canada.

Species distribution studies of fish or aquatic organisms are usually conducted at geographical scales significantly smaller than the Amazonian approach employed here (Oakes *et al.*, 2005; Domínguez-Domínguez *et al.*, 2006; Buisson, Blanc & Grenouillet, 2008; Esselman & Allan, 2011; Grenouillet *et al.*, 2011). Due to its vast size and remoteness, the Amazon basin has large gaps in information describing species distributions, biological data and hydrological data. According to Abell *et al.* (2008), the Amazon basin is one of the richest basins of the world, with more than 200 endemic freshwater fish, the majority of which have low or fragmented biogeographical information. Thus, the use of broad-scale environmental variables may allow us to generate species distribution models for those species to prioritise inventory studies and to determine new conservation areas to protect riverine fish species. The SDM can also provide information on habitat requirements, hotspot biodiversity, inventory and the management of vulnerable areas of habitat change on a large scale (Porter *et al.*, 2000). Once we have large gaps of information in the Amazon basin, we will be better able to predict accurate models to contribute for the conservation of fish species and the Amazon ecosystem.

Modelling stream fish species across the Amazon basin

Terrestrial macroclimatic and topographic variables have been used for modelling fish species distributions (Argent *et al.*, 2003; McNyset, 2005; Oakes *et al.*, 2005; Domínguez-Domínguez *et al.*, 2006). However, few of these studies have discussed how well these variables might reflect local aquatic conditions. Since our macroscale variables appropriately reflect the local limnological conditions, their use reflects factors related to local conditions and allows us to generate better fish species distributions models.

The hatchet fish *C. strigata*, the pencil fishes *C. nattereri* and *P.* cf. *brevis*, and the leaf catfish *Helogenes marmoratus* demonstrated the best models. These species occur almost exclusively in pristine forest streams and are considered habitat specialists. The model subsets A and B were compared for each species, and both the ROC and LPT were similar in terms of range and area. Their original occurrence points are concentrated in the central-western Amazon, but all subset models showed other suitable areas for these species in the eastern Amazon. The areas that lack records of occurrence for the analy-

sed species may represent simple sampling gaps but could also harbour sister species (e.g. Raxworthy *et al.*, 2003). For instance, *C. nattereri* seems to be replaced by *Copella arnoldi* in the Brazilian Amazon east of the Tapajós River (J. Zuanon, pers. obs.), and the absence of data regarding species of *Pyrrhulina* and *Carnegiella* in the small streams of the eastern Amazon suggests that similar replacements (or maybe undisclosed sister species) may be expected.

The Erythrinidae contains species that occur in many different habitats, such as lakes and small and large rivers (Oyakawa, 2003). *Hoplias malabaricus* has the widest distribution of this family, occurring in streams and large rivers throughout the Amazon basin (as well as in most tropical and subtropical South American basins). The projected distribution maps for *H. malabaricus* and *E. erythrinus* showed wide distributions in the Amazon basin, but their models had low performances. These results could be expected as these species have a broad tolerance to changing environmental conditions and have low habitat specificity. In this case, models do not have enough accuracy to capture the requirements of species, and these intrinsic aspects of widespread species may generate models with lower TSS and AUC values, as observed in other studies (Stockwell & Peterson, 2002; Lobo *et al.*, 2008).

In conclusion, although two species were not represented by strongly fitted models, in general, we can use macroscale variables to build species distribution models and generate information about the biogeographic distribution for Amazon stream fishes.

Acknowledgments

We are grateful to Brasil das Águas Project, ANA and Igarapés Project for developing and maintain its database; Denis Nogueira and Thiago Bernardi for helping with statistical procedures in R software environment; Dr. Luis Mauricio Bini, Dr. Julian D. Olden and Murilo S. Dias for comments that greatly improved this manuscript; CNPq and CAPES for financial support. J. Zuanon and P. De Marco Jr. receive productivity grants from CNPq (#307464/2009-1 and #305542/2010-9, respectively). This is contribution number 35 of the Igarapés Project.

References

Abell R., Thieme M.L., Revenga C., Bryer M., Kottelat M., Bogutskaya N. *et al.* (2008) Freshwater ecoregions of the world: a new map of biogeographic units for freshwater biodiversity conservation. *BioScience*, **58**, 403–414.

Angermeier P.L. & Karr J.R. (1983) Fish communities along environmental gradients in a system of tropical streams. *Environmental Biology of Fishes*, **9**, 117–135.

Araújo M.B., Pearson R.G., Thuiller W. & Erhard M. (2005) Validation of species – climate impact models under climate change. *Global Change Biology*, **11**, 1504–1513.

Araújo M.B., Thuiller W. & Pearson R.G. (2006) Climate warming and the decline of amphibians and reptiles in Europe. *Journal of Biogeography*, **33**, 1712–1728.

Argent D.G., Bishop J.A., Stauffer J.R., Carline R.F. & Myers W.L. (2003) Predicting freshwater fish distributions using landscape-level variables. *Fisheries Research*, **60**, 17–32.

Buermann W., Saatchi S., Smith T.B., Zutta B.R., Chaves J.A., Milá B. *et al.* (2008) Predicting species distributions across the Amazonian and Andean regions using remote sensing data. *Journal of Biogeography*, **35**, 1160–1176.

Buisson L., Blanc L. & Grenouillet G. (2008) Modelling stream fish species distribution in a river network: the relative effects of temperature versus physical factors. *Ecology of Freshwater Fish*, **17**, 244–257.

Buisson L., Grenouillet G., Casajus N. & Lek S. (2010) Predicting the potential impacts of climate change on stream fish assemblages. *American Fisheries Society Symposium*, **73**, 327–346.

Davies N.M., Norris R.H. & Thoms M.C. (2000) Prediction and assessment of local stream habitat features using large-scale catchment characteristics. *Freshwater Biology*, **45**, 343–369.

Domínguez-Domínguez O., Martínez-Meyer E., Zambrano L. & León G.P.P. (2006) Using ecological-niche modeling as a conservation tool for freshwater species: live-bearing fishes in central Mexico. *Conservation Biology*, **20**, 1730–1739.

Dudgeon D., Ed. (2008) *Tropical Stream Ecology*, 1st edn. Elsevier, London.

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**, 66–77.

Elith J., Kearney M. & Phillips S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 330–342.

Esselman P.C. & Allan J.D. (2011) Application of species distribution models and conservation planning software to the design of a reserve network for the riverine fishes of northeastern Mesoamerica. *Freshwater Biology*, **56**, 71–88.

Fausch K.D., Torgersen C.E., Baxter C.V. & Li H.W. (2002) Landscapes to riverscapes: bridging the gap between research and conservation of stream fishes. *BioScience*, **52**, 483–498.

Fielding A.H. & Bell J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Filipe A.F., Marques T.A., Seabra S., Tiago P., Ribeiro F., Moreira da Costa L. *et al.* (2004) Selection of priority

areas for fish conservation in Guadiana River Basin, Iberian Peninsula. *Conservation Biology*, **18**, 189–200.

Goulding M., Barthem R. & Ferreira E., Eds. (2003) *The Smithsonian Atlas of the Amazon*. Smithsonian Institution, Washington.

Grenouillet G., Buisson L., Casajus N. & Lek S. (2011) Ensemble modelling of species distribution: the effects of geographical and environmental ranges. *Ecography*, **34**, 9–17.

Jiménez-Valverde A. & Lobo J.M. (2007) Threshold criteria for conversion of probability of species presence to either–or presence–absence. *Acta Oecologica*, **31**, 361–369.

Kawaguchi Y., Taniguchi Y. & Nakano S. (2003) Terrestrial invertebrate inputs determine the local abundance of stream fishes in a forested stream. *Ecology*, **84**, 701–708.

Kennard M.J., Olden J.D., Arthington A.H., Pusey B.J. & Poff N.L. (2007) Multiscale effects of flow regime and habitat and their interaction on fish assemblage structure in eastern Australia. *Canadian Journal of Fisheries and Aquatic Sciences*, **64**, 1346–1359.

Kornis M.S. & Zanden M.J.V. (2010) Forecasting the distribution of the invasive round goby (*Neogobius melanostomus*) in Wisconsin tributaries to Lake Michigan. *Canadian Journal of Fisheries and Aquatic Sciences*, **67**, 553–562.

Kruse C.G., Hubert W.A. & Rahel F.J. (1997) Geomorphic influences on the distribution of Yellowstone cutthroat trout in the Absaroka Mountains, Wyoming. *Transactions of the American Fisheries Society*, **126**, 418–427.

Liu C., White M. & Newell G. (2009) Measuring the accuracy of species distribution models: a review. In: *18th World IMACS/MODSIM Congress*, pp. 4241–4247. Cairns, Qld, Australia.

Liu C., White M. & Newell G. (2011) Measuring and comparing the accuracy of species distribution models with presence-absence data. *Ecography*, **34**, 232–243.

Lobo J.M., Jiménez-Valverde A. & Real R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Mata R.A., Tidon R., Côrtes L.G., De Marco P. Jr & Diniz-Filho J.A.F. (2010) Invasive and flexible: niche shift in the drosophilid *Zaprionus indianus* (Insecta, Diptera). *Biological Invasions*, **12**, 1231–1241.

Matthews E.M. & Matthews W. (2000) Geographic, terrestrial and aquatic factors: which most influence the structure of stream fish assemblages in the midwestern United. *Ecology of Freshwater Fish*, **9**, 9–21.

McNyset K.M. (2005) Use of ecological niche modelling to predict distributions of freshwater fish species in Kansas. *Ecology of Freshwater Fish*, **14**, 243–255.

Mendonça F.P., Magnusson W.E. & Zuanon J. (2005) Relationships between habitat characteristics and fish assemblages in small streams of Central Amazonia. *Copeia*, **4**, 751–764.

Mugodo J., Kennard M., Liston P., Nichols S., Linke S., Norris R.H. *et al.* (2006) Local stream habitat variables predicted from catchment scale characteristics are useful for predicting fish distribution. *Hydrobiologia*, **572**, 59–70.

Nóbrega C.C. & De Marco Jr P. (2011) Unprotecting the rare species: a niche-based gap analysis for odonates in a core Cerrado area. *Diversity and Distributions*, **17**, 491–505.

Oakes R.M., Gido K.B., Falke J.A., Olden J.D. & Brock B.L. (2005) Modelling of stream fishes in the Great Plains, USA. *Ecology of Freshwater Fish*, **14**, 361–374.

Oberdorff T., Pont D., Hugueny B. & Chessel D. (2001) A probabilistic model characterizing fish assemblages of French rivers: a framework for environmental. *Freshwater Biology*, **46**, 399–415.

Olden J.D. & Jackson D.A. (2002) A comparison of statistical approaches for modelling fish species distributions. *Freshwater Biology*, **47**, 1976–1995.

Oyakawa O.T. (2003) Family Erythrinidae (Trahiras). In: *Check List of the Freshwater Fishes of South and Central America*, 1st edn (Eds R.E. Reis, S.O. Kullander & C.J. Jr Ferraris), pp. 238–240. EDIPUCRS, Porto Alegre.

Pearson R.G., Raxworthy C.J., Nakamura M. & Peterson A.T. (2006) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, **34**, 102–117.

Peterson A.T., Papes M. & Eaton M. (2007) Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. *Ecography*, **30**, 550–560.

Peterson A.T. & Soberón J. (2012) Species distribution modeling and ecological niche modeling: getting the concepts right. *Natureza e Conservação*, **10**, 102–107.

Phillips S.J., Anderson R.P. & Schapire R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips S.J. & Dudík M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.

Porter M.S., Rosenfeld J. & Parkinson E.A. (2000) Predictive models of fish species distribution in the Blackwater drainage, British Columbia. *North American Journal of Fisheries Management*, **20**, 349–359.

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Raxworthy C.J., Martinez-Meyer E., Horning N., Nussbaum R.A., Schneider G.E., Ortega-Huerta M.A. *et al.* (2003) Predicting distributions of known and unknown reptile species in Madagascar. *Nature*, **33**, 837–841.

Rodríguez-Soto C., Monroy-Vilchis O., Maiorano L., Boitani L., Faller J.C., Briones M.Á. *et al.* (2011) Predicting potential distribution of the jaguar (*Panthera onca*) in Mexico: identification of priority areas for conservation. *Diversity and Distributions*, **17**, 350–361.

Sioli H. ed. (1985) *Amazônia: Fundamentos da ecologia da maior região de florestas tropicais*, 1st edn. Vozes, Patrópolis.

Siqueira M.F., Durigan G., De Marco Jr P. & Peterson A.T. (2009) Something from nothing: using landscape similarity and ecological niche modeling to find rare plant species. *Journal for Nature Conservation*, **17**, 25–32.

Soberón J. (2007) Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters*, **10**, 1–9.

Stewart-Koster B., Boone E.L., Kennard M.J., Sheldon F., Bunn S.E. & Olden J.D. (2013) Incorporating ecological principles into statistical models for the prediction of species' distribution and abundance. *Ecography*, **36**, 342–353.

Stockwell D.R.B. & Peterson A.T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**, 1–13.

Syphard A.D. & Franklin J. (2009) Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. *Ecography*, **32**, 907–918.

Torrente-Vilara G., Zuanon J., Leprieur F., Oberdorff T. & Tedesco P.A. (2011) Effects of natural rapids and waterfalls on fish assemblage structure in the Madeira River (Amazon Basin). *Ecology of Freshwater Fish*, **20**, 588–597.

Vannote R.L., Minshall W.G., Cummins W., Sedell J.R. & Cushing C.E. (1980) The river continuum concept. *Canadian Journal of Fisheries and Aquatic Sciences*, **37**, 130–137.

Vari R.P. & Ferraris C.J. Jr (2003) Family Cetopsidae (Whale catfishes). In: *Check List of the Freshwater Fishes of South and Central America* (Eds R.E. Reis, S.O. Kullander & C.J. Jr Ferraris), pp. 257–260. EDIPUCRS, Porto Alegre.

Watson G. & Hillman T.W. (1997) Factors affecting the distribution and abundance of bull trout: an investigation at hierarchical scales. *Noth American Journal of Fisheries Management*, **2**, 237–252.

Weitzman S.H. & Palmer L. (2003) Family Gasteropelecidae (Freshwater hatchetfishes). In: *Check List of the Freshwater Fishes of South and Central America* (Eds R.E. Reis, S.O. Kullander & C.J. Jr Ferraris), pp. 101–103. EDIPUCRS, Porto Alegre.

Zuur F.A., Ieno N.E., Walker J.N., Saveliev A.A. & Smith M.G. (2009) Negative Binomial GAM and GAMM to Analyse Amphibian Roadkills.: In: *Mixed Effects Models and Extensions in Ecology with R* (Eds M. Gail, K. Krickeberg, J. Samet, A. Tsiatis & W. Wong), pp. 380–406. Springer Science+Business Media, New York.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. The variance inflation index (VIF) for macroscale variables and the standard coefficient of linear regression.

Table S2. Spearman analysis of collinearity among quantitative macroscale variables.

Table S3. Moran's I from residuals values from best model for each local variable, significant spatial autocorrelation $P < 0.005$.

Appendix S1. Collinearity and spatial autocorrelation among variables.

(Manuscript accepted 23 July 2014)