

**A peer-reviewed version of this preprint was published in PeerJ on 6 December 2019.**

[View the peer-reviewed version](https://doi.org/10.7717/peerj.7771) (peerj.com/articles/7771), which is the preferred citable publication unless you specifically need to cite this preprint.

Zamorano D, Labra FA, Villarroel M, Lacy S, Mao L, Olivares MA, Peredo-Parada M. 2019. Assessing the effect of fish size on species distribution model performance in southern Chilean rivers. PeerJ 7:e7771  
<https://doi.org/10.7717/peerj.7771>

# Assessing the effect of fish size on species distribution model performance in southern Chilean rivers

**Daniel Zamorano**<sup>1,2</sup>, **Fabio Labra**<sup>1,3</sup>, **Marcelo Villarroel**<sup>2</sup>, **Luca Mao**<sup>4</sup>, **Shaw Lucy**<sup>4</sup>, **Marcelo Olivares**<sup>5,6</sup>, **Matias Peredo-Parada**<sup>Corresp. 2</sup>

<sup>1</sup> Centro de Investigación e innovación en Cambio Climático, Universidad Santo Tomás, Santiago, Chile

<sup>2</sup> Area de Ecohidráulica, Plataforma de Investigación en Ecohidrología y Ecohidráulica Ltda, Santiago, Chile

<sup>3</sup> Programa de Doctorado en Conservación y Gestión de la Biodiversidad, Facultad de Ciencias, Universidad Santo Tomás, Santiago, Chile

<sup>4</sup> Departamento de Ecosistemas y Medio Ambiente, Facultad de Agronomía, Pontificia Universidad Católica de Chile, Santiago, Chile

<sup>5</sup> Departamento de Ingeniería Civil, Universidad de Chile, Santiago, Chile

<sup>6</sup> Centro de Energía, Universidad de Chile, Santiago, Chile

Corresponding Author: Matias Peredo-Parada

Email address: [matias.peredo@ecohyd.com](mailto:matias.peredo@ecohyd.com)

Despite its theoretical relationship, the effect of body size on the performance of species distribution models (SDM) has only been assessed in a few studies of terrestrial taxa. We aim to assess the effect of body size on the performance of SDM in river fish. We study seven Chilean freshwater fish, using models trained with three different sets of predictor variables: ecological (Eco), anthropogenic (Antr) and both (*Eco+Antr*). Our results indicate that the performance of the *Eco+Antr* models improves with fish size. These results highlight the importance of two novel predictive layers: the source of river flow and the overproduction of biotopes by anthropogenic activities. We compare our work with previous studies that modeled river fish, and observe a similar relationship in most cases. We discuss the current challenges of the modeling of riverine species, and how our work helps suggest possible solutions.

# Assessing the effect of fish size on species distribution model performance in southern Chilean rivers

D. Zamorano<sup>1,2</sup>, F.A. Labra<sup>1,3</sup>, M. Villarroel<sup>2</sup>, L. Mao<sup>4</sup>, S.N. Lacy<sup>4</sup>, M. Olivares<sup>5,6</sup> & M. Peredo-Parada<sup>\*2</sup>

1.- Centro de Investigación e innovación para el Cambio Climático, Universidad de Santo Tomás

2.- Plataforma de Investigación en Ecohidrología y Ecohidráulica limitada, Santiago, Chile.

3.- Programa de Doctorado en Conservación y Gestión de la Biodiversidad, Facultad de Ciencias,

Universidad Santo Tomás, Santiago, Chile.

4.-Departamento de Ecosistemas y Medio Ambiente, Facultad de Agronomía e Ingeniería Forestal,

Facultad de Agronomía, Pontificia Universidad Católica de Chile, Santiago, Chile

5.- Departamento de Ingeniería Civil, Universidad de Chile, Santiago, Chile.

6.- Centro de Energía, Universidad de Chile, Santiago, Chile.

\* Corresponding Author (matias.peredo@ecohyd.com)

## KEYWORDS

Chilean fishes, Random Forest, Neural Networks, General Lineal Model, Species Distribution Model,

Anthropogenic variables.

## 18 ABSTRACT

19 Despite its theoretical relationship, the effect of body size on the performance of species distribution  
20 models (SDM) has only been assessed in a few studies of terrestrial taxa. We aim to assess the effect of  
21 body size on the performance of SDM in river fish. We study seven Chilean freshwater fish, using models  
22 trained with three different sets of predictor variables: ecological (Eco), anthropogenic (Antr) and both  
23 (Eco+Antr). Our results indicate that the performance of the *Eco+Antr* models improves with fish size.  
24 These results highlight the importance of two novel predictive layers: the source of river flow and the  
25 overproduction of biotopes by anthropogenic activities. We compare our work with previous studies that  
26 modeled river fish, and observe a similar relationship in most cases. We discuss the current challenges of  
27 the modeling of riverine species, and how our work helps suggest possible solutions.

## 28 INTRODUCTION

29 Species distribution models (SDM) provide an important management tool to support conservation  
30 planning. SDMs generate species distribution maps that allow for more efficient and effective field  
31 inventories, suggest sites of high potential occurrence of rare species for survey planning, and permit  
32 testing biogeographical, ecological and evolutionary hypotheses (Guisan & Thuiller, 2005). Given these  
33 advantages, different international organizations (e.g., UNEP, Conservation International, IUCN, WWF)  
34 have employed SDM to address key policy objectives at a global scale (Cayuela et al., 2009).

35 Different species traits have been shown to influence model performance (Brotons et al., 2004; Segurado  
36 & Araújo, 2004; McPherson & Jetz, 2007; França & Cabral, 2016). One important trait is body size  
37 (Radinger et al., 2017). Larger species detect less food but can tolerate lower resource concentrations  
38 within their food, while smaller species detect more food, but require higher resource concentrations  
39 within it (Ritchie, 1998; Ritchie & Olff, 1999). As a result, larger species have larger home ranges than  
40 smaller species (Calder & A, 2001; Woolnough, Downing & Newton, 2009).

41 Body size may affect the performance or accuracy of SDM in different ways (McPherson & Jetz, 2007).  
 42 First, species with larger home ranges may perceive the environment at coarser scales, improving the  
 43 performance of distribution models based on coarse-grained predictors (Suarez-Seoane, Osborne &  
 44 Alonso, 2002). Second, home-range extent may influence the amount of data available, as well as the  
 45 balance between presences and absences (McPherson, Jetz & Rogers, 2004). In addition, species with  
 46 local adaptations in habitat preferences may generate models that overestimate their ecological niches  
 47 (Stockwell & Peterson, 2002). To date, the effect of body size on distribution models has been tested in  
 48 different taxa with unclear results (e.g. M. McPherson & Jetz, 2007; França & Cabral, 2016; Morán-  
 49 Ordóñez et al., 2017; Radinger et al., 2017).

50 In the case of fish, Radinger *et al.* (2017) indicate that smaller-body fishes are less sensitive to  
 51 anthropogenic intervention in the river network, due to their smaller home ranges. However, their study  
 52 did not explicitly test variation in model performance in response to fish size. Recent research on fish  
 53 species distribution models has shown that species with different body sizes are impacted differently by  
 54 the same sets of environmental features derived from anthropogenic activities (Perry *et al.*, 2005;  
 55 Radinger *et al.*, 2017). A relevant research question is whether SDM performance or accuracy for  
 56 different body-sized fishes vary in the same manner when considering different predictor variable sets,  
 57 such as i) ecological predictors, ii) anthropogenic predictors, and iii) ecological and anthropogenic  
 58 predictors.

59 The ichthyofauna in Chile comprises a total of 44 species, including two lampreys (Habit, Dyer & Vila,  
 60 2006), and is characterized as being highly endemic, adapted to high slope rivers, and with small body  
 61 sizes (Vila, Fuentes & Contreras, 1999; Vila et al., 2006; Habit, Dyer & Vila, 2006). Despite its high  
 62 biogeographic value, the Chilean ichthyofauna is broadly endangered, with only two species (*Cheirodon*  
 63 *australe* y *Mugil cephalus*) classified as non-endangered. In Chile, anthropogenic variables represent the  
 64 main group of threats to river fishes (Habit et al., 2002, 2006). Therefore, understanding the potential

65 impact of body size on SDM performance is highly relevant for conservation and management planning  
66 efforts.

67 In this study, we quantify anthropogenic variables (*Antr*) and ecological variables (*Eco*) at the scale of the  
68 river segment, and we generate SDMs for seven native freshwater species. We focus on two well studied  
69 southern Chilean basins: Bueno and Valdivia. Our specific objectives are: (1) to assess the effect of fish  
70 size on species distribution model performance and variable participation by model, fitted using three  
71 sets of environmental features: i) models trained with ecological predictors (*Eco*), ii) models trained  
72 models with anthropogenic predictors (*Antr*), and iii) models trained with ecological and anthropogenic  
73 predictors (*Eco+Antr*); (2) to examine the predicted biotopes generated by each model for different  
74 species studied; and (3) to compare our results with model performances in previous studies.

## 75 **METHODS**

### 76 **Study area and modeled species**

77 The study area covers the Valdivia and the Bueno river basins, located in the central-southern zone of  
78 Chile, between the parallels 39.33° and 41.08° S (Figure 1). The Valdivia River basin has a pluvial  
79 hydrological regime, and it is characterized by having a chain of interconnected lakes at higher altitudes.  
80 The upper section of the Bueno River basin has a pluvial-nival regime, while the middle and lower part of  
81 the basin is governed by a pluvial regime (Errázuriz K. *et al.*, 1998).

82 To characterize a set of hydrological variables for the study area, we used the national official drainage  
83 network generated by the Military Geographic Institute (Instituto Geográfico Militar, Government of  
84 Chile). This drainage network was divided in segments to build the SDM. We considered river segment  
85 between 2 and 10 km of length, having homogeneous hydromorphological conditions with no significant  
86 confluences. This definition was generated using cartographic information, Google Earth (Google inc,  
87 2009), and Arc GIS version 9.2 (ESRI, 2010).

Our study included seven freshwater fish species (Table 1): *Aplochiton taeniatus* (Jenyns, 1842), *Aplochiton zebra* (Jenyns, 1842), *Brachygalaxias bullocki* (Regan, 1908), *Cheirodon australe* (Eigenmann, 1928), *Odontesthes mauleanum* (Steindachner, 1896), *Percilia gillissi* (Girard, 1855), and *Trichomycterus areolatus* (Valenciennes, 1846). Statistical analysis of the effect of body size was carried out using theoretical species maximum length, which is available for all these species. Most maximum length estimates were obtained from official species descriptions provided in each species conservation assessment developed by the Chilean Ministry of the Environment (Table 1). The only exception was *B. bullocki*, which had not been assessed by the Ministry of Environment, and whose maximum length was obtained from Fishbase (Froese & Pauly, 2017) (Table S1).

This species was selected because represent a good size gradient (between 5.5 cm and 30 cm) (Table 1) to Chilean species case, and particularly all this species encompasses completely the latitudinal range of both basins (Table 1), situation that allow comparing model performance without the distribution range by specie affect in the predicted distribution.

## Modeling methods

### *Species occurrence data*

Historical records of the presences of the study species were obtained from the Ministry of the Environment's (Ministerio del Medio Ambiente, Government of Chile) database on freshwater organisms. This database was generated by collecting published databases of scientific samples in the study area (Ministerio de Energía - División de Desarrollo Sustentable, 2016).

In addition, a field sampling campaign was conducted in the study area to complement existing information in the government database. The sampling was done between December 2015 and January 2016, using electrofishing equipment (SAMUS, model 745G). We collected all fish along a 100-meter transect, with sampling times of 45 to 60 min, depending of the hydromorphological features of the site.

All collected fish were identified *in situ*, using a field identification manual (Habit *et al.* 2006). The electrofishing was approved by National Fisheries Service permit number 630.

Each presence record was associated with the closest river segment in the GIS, thus building a presence database for species distribution modeling. Overall, 118 river segments had at least one presence record (Fig. 1). The number of presences modeled for each species ranged between 9 and 39 (Table 1). We considered other records (118 - n) as true absences in each generated model.

### *Predictor variables*

The predictor variables or features considered were: accumulated rainfall, catchment area, source-of-flow, altitude, slope, channel width, riparian vegetation percent, land-use, cross-channel construction, and within-channel construction. All variables were grouped according to their origin (ecological variables and anthropogenic variables) and their spatial scale (inter-basin, basin or segment) (Table 2).

Accumulated rainfall was obtained by relating the isolines of annual rainfall published by the Water General Directorate (Dirección General de Aguas, Government of Chile), accumulated over the basin. Catchment area was calculated with a DEM image of 1km × 1km (Landsat 7 images from 2015, <https://landsat.usgs.gov/>) using the Hydrology package in of ArcGIS. Source-of-flow was obtained from the published REC-Chile classification (Peredo-Parada *et al.*, 2011). Altitude and slope were estimated using the altitudes of the ends of each river segment, based on the DEM. Channel width, riparian vegetation coverage, land-use, cross-channel construction, and within-channel construction were estimated by visual analysis of Google Earth imagery. Channel width was calculated as the mean of three points along the section. Riparian vegetation coverage was considered up to 50m distant from the stream, with sections and land use percent considered up to 200m. Within-channel constructions includes road parallel to the river, bank reinforcement, maintenance river channel, channelization, among others. Cross-channel constructions include bridges, dams and intake structure.

# 134 *Model training and evaluation*

135 We used three algorithms to estimate SDM for all seven species: random forest (RF) (Breiman, 2001),  
136 neural network (NNET) (Stern, 1996), and general lineal model (GLM) (McCullagh, 1984). These methods  
137 were chosen based on their good performance with presence and absence or pseudoabsences for  
138 species-distribution data (Mastorillo *et al.*, 1997; Cutler *et al.*, 2007; Elith & Leathwick, 2009). RF uses a  
139 learning strategy, based on the generation of many classification trees, then aggregating their results in  
140 the final output (Breiman, 2001). NNET is derived from a simple model that mimics of the structure and  
141 function of the brain, and maximizes the prediction during the model-training phase by comparing actual  
142 outputs with desired outputs (Manel, Dias & Ormerod, 1999). GLM is a statistical model that predict  
143 values determined by discrete and continuous predictor variables and by the link function (e.g. logistic  
144 regression, Poisson regression) (Bolker *et al.*, 2009). Using these different models allowed us to compare  
145 the performance of anthropogenic predictors in algorithms with different interpretation methods.  
146 Analysis was performed in R (v 2.3.3), using the Caret package (Kuhn, 2008).

147 For all the algorithms and species, we first trained the models using 70% of the dataset randomly  
148 selected, and evaluated SDM final performance with the remaining 30%. Each model was trained using a  
149 5-fold cross-validation scheme, except for the *O. mauleanum* (9 presences), where we used  
150 bootstrapping, due its low presences. During the training, imbalanced classes were corrected selecting a  
151 random sample (with replacement) of the minority class to be the same size as the majority class. For  
152 each specie, we trained 10 models with presences/absences resample of 70% of the original dataset. The  
153 final model was designated as the consensus of these 10 models based on the area under the curve  
154 (AUC) of the receiver operator characteristic (ROC). In order to assess model performance for each  
155 algorithm and species, we calculated the mean and confidence intervals of AUC (Fielding & Bell, 1997)  
156 and true skill statistic (TSS) (Allouche, Tsoar & Kadmon, 2006) using the 30% of observations separated at  
157 the beginning. TSS compares the number of correct forecasts, minus those attributable to random  
158 guessing, to that of a hypothetical set of perfect forecasts. In comparison, AUC is a single threshold-

independent measure for model performance obtained from ROC curves. These curves are constructed using all possible thresholds to classify the scores into confusion matrices (Allouche, Tsoar & Kadmon, 2006).

For RF and NNET, the Caret R package was used to fitting and tuning. Many predictive and machine learning models have structural or hyperparameters that cannot be directly estimated from the data. For example, in the case of RF models, the classification trees may be built using a given number of randomly selected predictors, which are named “*mtry*” (Kuhn & Johnson, 2013a). A hyperparameter such as *mtry* is usually fixed at a given value when training and calibrating an RF model, which is an iterative optimization process itself. Hyperparameter tuning of an RF model refers to the grid search procedure that allows the algorithm to find the best value of *mtry* to obtain the best model performance (given a set of calibration and validation data points). In our implementation of RF models, the search for an optimal *mtry* value spanned the space between 2 and 10 variables. Thus, the tuning process allowed us to explore a range of values for the RF hyperparameter, further improving model performance. This generated a final model with the best hyperparameter value for a given search grid (Kuhn & Johnson, 2013b). For NNET models, two tuning hyperparameters were used. These were the weight decay for successive neural layers (“*decay*”) and the number of hidden units (“*size*”). The grid search procedure examined weight decay values ranging between 4 and 6, while the number of hidden units was allowed to vary between 0.05 and 0.9. Both hyperparameter range are calibrated by trial and error process, optimizing the model performance. GLM algorithms were optimized using a stepwise procedure for variable selection (Zhang, 2016), implemented with the “*stepAIC*” function (R MASS package in R v 2.3.3) (R Core Team, 2017).

Occurrence probabilities were categorized to presence/absence for all models. Thresholds were determined so as to maximize the sum of sensitivity and specificity (MaxSens+Spec; PresenceAbsence package in R v 2.3.3) (R Core Team, 2017). This criterion is independent of the theoretical prevalence (Manel, Dias & Ormerod, 1999; Allouche, Tsoar & Kadmon, 2006), causing the distribution of rare species to be overpredicted. In our particular case, the theoretical prevalence in the study area for all the species

184 is close to 0.5, but presences of our species are low, requiring a relaxation of this criterion when defining  
185 the threshold that allows for the definition of each of the species distribution across the studied  
186 watersheds.

187 In order to examine the predicted distribution for each species across the study area, each river segment  
188 was categorized into eight classes, according of the presence/absence results of each model: 1) no model  
189 selection as presence, presence determined by 2) only *Eco*, 3) only *Antr*, 4) *Eco+Antr*, 5) *Eco* and *Antr*, 6)  
190 *Eco* and *Eco+Antr*, 7) *Antr* and *Eco+Antr*, and 8) all models.

## 191 Relationship between fish size and models

192 In order to examine statistical effect of body size,  $\log_{10}$ -transformations of maximum length (*max. length*)  
193 were calculated for each species. *Max. length* was related with *TSS* and *AUC*. Also, *max. length* and the  
194 predictors variables for all (*Eco*, *Antr* and *Eco+Antr*) models was related by correlating its participation For  
195 each models (*Eco*, *Antr* and *Eco + Antr*) the level of participation of their predictor variables was  
196 correlated to the *max. length*. This relationship was corrected by the permutation procedure (Legendre &  
197 Legendre, 1998).

## 198 Biotope comparisons

199 We compared *Eco*, *Antr*, and *Eco+Antr* biotopes generated for each species by using Venn diagrams.  
200 Overlap of the ellipses in the Venn diagrams let us determine whether these models predicted the same  
201 observed river sections as shown by presence records. Non-overlapping of *Antr*, *Eco*, and *Eco+Antr*  
202 ellipses meant that at least one model predicted a different pattern of river segment occupation.  
203 Geographic information was processed in QGIS software v 2.18.10 (QGIS, 2015). Models were executed  
204 and evaluated in R v 3.3.2 (R Core Team, 2017).

## Comparison with prior research

We conducted a bibliographic review of research that used SDMs for assessing riverine fish, considering three characteristics: 1) modeled groups of fish ( $n > 5$ ); 2) used ecological and anthropogenic predictor variables; and 3) had a river-segment-scale model grain. To compare results, we obtain the maximum length of each modeled fish from Fishbase (Froese & Pauly, 2017).

Filipe, Cowx & Collares-Pereira (2002) indicated that percent of total correctly classified; percent of presences correctly classified; and percent of absences correctly classified functioned as measurements of fit. These were transformed to TSS for results comparisons.

In order to compare our results with previous studies, it was necessary to perform two statistical analyzes. The first analysis was compared result with TSS, while the second was compared result with AUC. In both, we use an ANCOVA (Heiberger & Holland, 2013) with  $\log_{10}$ -transformed maximum length for each species as covariable and the fit metric as the response variable. In the AUC test, the response variable used a Box-Cox transformation (Box & Cox, 1964) to obtain normal residuals. Finally, for comparing the number of records used per species, we used a Kruskal-Wallis test (Hollander & Wolfe, 1999) to compare all the papers at the same time. All analyses were done using R (R Core Team, 2017).

## RESULTS

### TSS relates positively with size fishes

Results show that only in four species (*A. zebra*, *A. teniatus*, *P. Gillisi*, and *C. australe*) have good model performances with AUC values greater than 0.75 (Table 1).

The TSS of the *Eco+Antr* models are related positively and marginally significant with fish sizes ( $R = 0.73$ ,  $p$  value = 0.06,  $p$  perm = 0.07). For *Antr* and *Eco* models, the relationships with fish size were not significant, but there was a negative relationship between fish size, TSS and AUC values in the *Eco* models, and a positive relationship between fish size, TSS and AUC values in the *Antr* models.

Only “altitude” (*Eco+Antr* model:  $R = -0.72$ ,  $p$  value = 0.06,  $p$  perm = 0.08; and *Eco* model:  $R = -0.71$ ,  $p$  value = 0.07,  $p$  perm = 0.09) and “slope” (*Eco* model:  $R = -0.74$ ,  $p$  value = 0.06,  $p$  perm = 0.09) showed a marginally significant and negative relationship with fish size (Table S3).

## Variable scale determining its participation

Regarding variable participation, in the *Eco+Antr* and *Eco* models, the “accumulated rain” (regional scale) had the biggest average percent participation (82% and 77%, respectively), followed by “source-of-flow” (64% and 65%, respectively) and “catchment” (60% in both models). Source-of-flow and catchment were considered at the basin scale. Anthropogenic variables (segment scale) did not show important participation, except in the *Eco+Antr* models of *O. mauleanum*, *B. bullocki*, and *P. gillisi*. In these cases, land-use was the most important variable. In the other species, *Eco* and *Eco+Antr* models held the same important predictor variables (Figure 3). In *Antr* models, mean variable participation was: 77% to “cross-channel construction”, 75% to “land-use” and 62% to “within-channel construction;” all variables at the segment scale.

In all models, except in *B. bullocki*, the *Antr* models represented over the 40% of the all biotopes predicted by all the models. In these cases, *Eco* and *Eco+Antr* models coincided in the most segments predicted in common by both models. All species models predicted presences over more than 50% of total river-distance (Figure 4).

## Similar results to prior research

Our main results were compared with three previous pieces of research: Filipe, Cowx & Collares-Pereira (2002) (Sample unit = river lineal segment, Fit metric = TSS); Markovic, Freyhof & Wolter (2012) (Sample unit = pixel; Fit metric = AUC); and Radinger et al. (2017) (Sample unit = pixel; Fit metric = AUC).

In the TSS test, only the fish size covariate shows a significant relation, interacting positively (Appendix S1). In the AUC test, there is a significant difference in AUC values between both studies. In Radinger et al. (2017), the relationship between AUC and fish size is negative, opposite to what was shown by

Markovic, Freyhof & Wolter (2012) (Appendix S2). When we compared species presence numbers between papers, the Kruskal-Wallis test reported significant differences (chi-squared = 54.52, df = 3, p-value < 0.001). Markovic, Freyhof & Wolter (2012) worked with a greater number of presences (Markovic, Freyhof & Wolter (2012),  $\bar{x}$  = 932.32 presences per species; the others papers,  $\bar{x}$  = 43.05 presences per species) (Appendix S3).

## DISCUSSION

### Fish size and model fit

The relationship between fish sizes and model performance can be summarized as follows. First, *Eco+Antr* models showed the best performance in larger fish species, while *Antr* models show a marginally significant trend. Secondly, SDM fitted for smaller fish species did not achieve good fits, regardless of hyperparameter grid search procedure used to optimize the machine learning algorithms or the stepwise procedure used for GLM. A third emerging pattern is that performance to smaller fish species in *Eco* models improves slightly, without achieving good fit. As mentioned earlier, these body-size effects on SDM performance have been demonstrated in a few previous studies, despite the expected theoretical relationship (McPherson & Jetz, 2007). For example, Morán-Ordóñez et al. (2017) found no relationship between body size and model performance for trees and birds. França & Cabral (2016) successfully related model performance to species feeding mode and estuarine functional groups, with little involvement of body size in the relationship. In studies aimed at river fish, both Radinger et al. (2017) and Filipe, Cowx & Collares-Pereira (2002) found that fish size increased model performance, which coincides with our main results. However, Markovic, Freyhof & Wolter (2012) did not find this pattern. The main difference between those studies is the number of presences used in each model. The observed correlation between fish size and the model fits might be explained by this difference. Identifying pattern distributions for small fishes is more difficult due to small homes range and other considerations (McPherson & Jetz, 2007), but we could get better model results for small fishes when we

increased the number of presences for model calibration and validation, as suggested by Stockwell & Peterson (2002) and done by Markovic, Freyhof & Wolter (2012) .

Our results are even more relevant in regions where the entire fish community is particularly small, like in Chile (Vila et al., 2006). Moreover, in Chile there are no SDM reports for fish implemented with more than 100 presences as in Markovic, Freyhof & Wolter (2012). In that case, one option is to obtain predictor variables at a lower spatial resolution. In general, Radlinger et al. (2017) achieved good performances with a pixel resolution of 250m, using secondary variables as predictors in a 1,094 km long basin. So, methodology of Radlinger et al. (2017) could apply in Chile. Since Chilean rivers are 150 km long aprox, selection of predictor variables of Radlinger et al. (2017) should be adjust to small basins, as our case.

To develop models that perform well in small basins, besides incorporating predictor variables at different spatial scales, as we did, in further research we recommend incorporating different hydromorphological features to our variable set at the reach scale, such as sediment type or morphological classification and anthropogenic variables related to industrial development, like pollution or water extraction, that would play significant roles in riverine ecology according to the literature (Torgersen et al., 1999; Lange et al., 2014; Ramezani et al., 2016). For example, *T. areolatus* shows preferences for river bedrock, so we would expect that the incorporation of the “sediment type” variable would improve the model performance.

We found two important results, but non-significant tendencies: Smaller fishes have better fit in *Eco* models, and larger fishes have better fit in *Antr* models. This pattern can be explained by fish home ranges. Larger fishes are expected to be substantially restricted by movement barriers, given an ability to disperse farther than small fishes (Radlinger & Wolter, 2015; Radlinger et al., 2017), and so they respond better to *Antr* variables. Conversely, the lower dispersal ability of smaller fishes implies a slower response to anthropogenic drivers (Radlinger & Wolter, 2015; Radlinger et al., 2017), so it is better modeled with

300 *Eco* variables, since these variables project a potential distribution without anthropogenic interventions.  
301 These results are coherent with current literature and shows the differential relationship between  
302 anthropogenic pressure and fish size (Radinger et al., 2017).  
303 Altitude (in *Eco+Antr* and *Eco* models) and slope (in *Eco+Antr* models) were the predictor variables  
304 associated to body size: for large fishes, altitude and slope weren't important for model fit. This  
305 relationship between altitude and body size was also reported by Markovic, Freyhof & Wolter (2012) ( $r =$   
306  $-0.48$ ,  $p$  value =  $0.03$ ). We associate this result to river turbulence, since large fishes better resist  
307 turbulence (Lupandin, 2005), which is often greater at higher altitudes and higher slopes (Elliott, 2010).  
308 This resistance would indicate that altitude and slope are not relevant environmental filters in habitat  
309 selection among larger fishes, decreasing its participation in these SDMs.

## 310 **Participation by predictor variable**

311 On the *Eco+Antr* and *Eco* models, regardless fish size, the relevance by predictor variable for all models  
312 responded to the hierarchical framework of stream habitat proposed in literature (Frissell et al., 1986;  
313 Snelder & Biggs, 2002; Creque, Rutherford & Zorn, 2005; Steen et al., 2008; Peredo-Parada et al., 2011),  
314 and while predictor variables (or landscape filters) at bigger scales have more participation in the models,  
315 as the geographical scale of the variables decreases, so does its participation in the model, and their  
316 importance is resolved species by species.

317 While accumulated rain structures the landscape from east to west (from mountain to ocean), and from  
318 north to south (greater precipitation to the South), source-of-flow represents territorial particularities,  
319 like glaciers, lakes, and valleys. In this way, both variables summarize much of the spatial variability of  
320 both basins, having more participation in the majority of *Eco+Antr* and *Eco* models.

321 We want to highlight the use in our study of source-of-flow as a predictor variable, which is not found in  
322 any research of river species modelling (Filipe, Cowx & Collares-Pereira, 2002; Chu, Mandrak & Minns,  
323 2005; Steen et al., 2008; Markovic, Freyhof & Wolter, 2012; Jähnig et al., 2012; Domisch et al., 2013;

Elliott et al., 2015; Pletterbauer, Graf & Schmutz, 2016; Radinger et al., 2017; Taylor, Papeş & Long, 2017), especially in torrential basins like those found in Chile, which have short runs, with relatively large lakes, glaciers, or salt pans that significantly affect hydrological and hydraulic conditions. Source-of-flow variable is implement in river of New Zealand (Snelder & Biggs, 2002) and Chile (Peredo-Parada et al., 2011) what would facilitate its use in SDMs.

In *Antr* models, that within-channel construction generating a direct impact in the reach, Land-Use and cross-channel construction were the anthropogenic variables with the most participation in the *Antr* and *Eco+Antr* models. We relate this result to impact scale of within-channel construction. This variable frequently represents a proxy of intervention at reach scale, and since the model grain was the segment scale, the model resolution probably was unable to completely capture the impacts to reach scale.

In the current context of river species modeling, there is no broad agreement on predictor variables for modeling, unlike terrestrial species modeling, where Bioclim is the most used spatially database for predictor variables (Booth et al., 2014). In case of river species models, consideration of hierarchical, longitudinal, lateral and vertical river links to select predictors (Domisch et al., 2015) is necessary, and the most of riverine predictor variables are correlated theoretical and statistically (Leopold, 1969; Elliott, 2010). This makes the number of potential predictors of a riverine freshwater SDM very high, and allow that many of these variables can be change by proxies, diversifying hugely the predictor selected between papers. For example, many authors have used common proxies for river temperature, discharge and turbulence, like altitude, flow accumulation, slope, catchment, among others (Filipe, Cowx & Collares-Pereira, 2002; Markovic, Freyhof & Wolter, 2012; Elliott et al., 2015; Pletterbauer, Graf & Schmutz, 2016; Radinger et al., 2017; Taylor, Papeş & Long, 2017), but there is no general agreement in the literature as to which proxy to use. This ecological context hindering to systematic use of any variable as predictor, and the lack of an agreement between researchers difficult to compare results between them. This problem should be resolve in a future, increasing consensus in terms of predictor variables selection.

## Spatial patterns of distribution

We interpret the biotope overprediction of *Antr* models for six from seven species models as a consequence of statistic structure of the *Antr* variable predictors. Land-use is categorical variable, and the others are discrete variable with low variability (maximum number of within-channel construction: 8, maximum number of cross-channel construction: 10), and the segment percent without interventions (exclusively natural land use, and without any intervention that cross the river or within the river) is 51%. Thus, when the models relate the presence of any species with little disturbed segment, the number of river segment matching this condition is very high, increasing the biotope in comparison to the *Eco+Antr* and *Eco* models. Regrettably, we do not find other research where modeled only with anthropogenic variables, so we cannot compare our results with current literature.

The great coincidence between the biotopes generated by the *Eco+Antr* and *Eco* models was unexpected, since *Eco* biotopes were expected to be bigger than *Eco+Antr* biotopes, as reported in Taylor, Papeş & Long (2017), since the *Eco* models estimate potential niche, while *Eco+Antr* models estimate realized niche, with the former always larger than the latter (Jackson & Overpeck, 2009). While, the Bueno and Valdivia River basins have significant levels of anthropogenic pressures, these are apparently insufficient for to change the projected biotopes in under the *Eco+Antr* models. This result provides an optimistic view of the environmental conditions for the presence of threatened fishes in the Valdivia and Bueno River basins.

## CONCLUSIONS

SDM performance for small fish was found to be less accurate due to modelling grain of variable predictors, but would this effect can be alleviated by increasing the number of presences. When ecological and anthropogenic variables were considered together, ecological variables at the higher spatial scale were more relevant than predictor variables at the lower spatial scale, adjusting them to the hierarchical stream framework of Frissell et. al. (1986). Source-of-flow was found as a novel predictor

variable at the basin scale, with an important participation in the models of different sized fishes. High coincidence between the biotopes generated by *Eco+Antr* and *Eco* models, suggest that Bueno and Valdivia River basins have low anthropogenic interventions. We found evidence of how physiological characteristics determine SDM performance. This research serves as a base for future studies of river fish modelling in a particular ecological context, with relatively small fishes in moderately intervened, relatively short, torrential river basins.

## REFERENCES

- Allouche O., Tsoar A., Kadmon R. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* 43:1223–1232. DOI: 10.1111/j.1365-2664.2006.01214.x.
- Bolker BM., Brooks ME., Clark CJ., Geange SW., Poulsen JR., Stevens MHH., White J-SS. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* 24:127–135. DOI: 10.1016/J.TREE.2008.10.008.
- Booth TH., Nix HA., Busby JR., Hutchinson MF. 2014. bioclim : the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Diversity and Distributions* 20:1–9. DOI: 10.1111/ddi.12144.
- Box GEP., Cox DR. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)* 26:211–252. DOI: 10.2307/2984418.
- Breiman L. 2001. Random Forests. *Machine Learning* 45:5–32. DOI: 10.1023/A:1010933404324.
- Brotons L., Thuiller W., Araújo MB., Hirzel AH. 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography* 27:437–448. DOI: 10.1111/j.0906-7590.2004.03764.x.
- Calder WA., A W. 2001. Ecological Consequences of Body Size. In: *Encyclopedia of Life Sciences*. Chichester, UK: John Wiley & Sons, Ltd,. DOI: 10.1038/npg.els.0003208.
- Cayuela L., Golicher D., Newton A., Kolb H., de Albuquerque FS., Arets EJM., Alkemade JRM., Pérez AM. 2009. Species distribution modeling in the tropics: problems, potentialities, and the role of biological data for effective species conservation. *Tropical Conservation Science* 2:319–352.
- Chu C., Mandrak NE., Minns CK. 2005. Potential impacts of climate change on the distributions of several common and rare freshwater fishes in Canada. *Diversity and Distributions* 11:299–310. DOI: 10.1111/j.1366-9516.2005.00153.x.
- Creque SM., Rutherford ES., Zorn TG. 2005. Use of GIS-Derived Landscape-Scale Habitat Features to Explain Spatial Patterns of Fish Density in Michigan Rivers. *North American Journal of Fisheries Management* 25:1411–1425. DOI: 10.1577/M04-121.1.

- 406 Cutler DR., Edwards TC., Beard KH., Cutler A., Hess KT., Gibson J., Lawler JJ. 2007. Random Forests for  
407 Classification in Ecology. *Ecology* 88:2783–2792. DOI: 10.1890/07-0539.1.
- 408 Domisch S., Araújo MB., Bonada N., Pauls SU., Jähnig SC., Haase P. 2013. Modelling distribution in  
409 European stream macroinvertebrates under future climates. *Global Change Biology* 19:752–762.  
410 DOI: 10.1111/gcb.12107.
- 411 Domisch S., Jähnig SC., Simaika JP., Kuemmerlen M., Stoll S. 2015. Application of species distribution  
412 models in stream ecosystems: the challenges of spatial and temporal scale, environmental  
413 predictors and species occurrence data. *Fundamental and Applied Limnology / Archiv für*  
414 *Hydrobiologie* 186:45–61. DOI: 10.1127/fal/2015/0627.
- 415 Elith J., Leathwick JR. 2009. Species Distribution Models: Ecological Explanation and Prediction Across  
416 Space and Time. *Annual Review of Ecology, Evolution, and Systematics* 40:677–697. DOI:  
417 10.1146/annurev.ecolsys.110308.120159.
- 418 Elliott S. 2010. El río y la Forma. Introducción a la Geomorfología Fluvial. *RiL Editores. Chile*.
- 419 Elliott JA., Henrys P., Tanguy M., Cooper J., Maberly SC. 2015. Predicting the habitat expansion of the  
420 invasive roach *Rutilus rutilus* (Actinopterygii, Cyprinidae), in Great Britain. *Hydrobiologia* 751:127–  
421 134. DOI: 10.1007/s10750-015-2181-9.
- 422 Errázuriz K. AM., Cereceda T. P., Gonzalez L. JI., Gonzalez L. M., Henriquez R. M., Rioseco H. R. 1998.  
423 *Manual de geografía de Chile*. Editorial Andrés Bello.
- 424 ESRI. 2010. ArcGIS Desktop.
- 425 Fielding AH., Bell JF. 1997. A review of methods for the assessment of prediction errors in conservation  
426 presence/absence models. *Environmental Conservation* 24:38–49. DOI: DOI: undefined.
- 427 Filipe AF., Cowx IG., Collares-Pereira MJ. 2002. Spatial modelling of freshwater fish in semi-arid river  
428 systems: a tool for conservation. *River Research and Applications* 18:123–136. DOI:  
429 10.1002/rra.638.
- 430 França S., Cabral HN. 2016. Predicting fish species distribution in estuaries: Influence of species' ecology  
431 in model accuracy. *Estuarine, Coastal and Shelf Science* 180:11–20. DOI:  
432 10.1016/j.ecss.2016.06.010.
- 433 Frissell CA., Liss WJ., Warren CE., Hurley MD. 1986. A hierarchical framework for stream habitat  
434 classification: Viewing streams in a watershed context. *Environmental Management* 10:199–214.  
435 DOI: 10.1007/BF01867358.
- 436 Froese R., Pauly D. 2017. FishBase. Available at <http://www.fishbase.org/> (accessed January 22, 2018).
- 437 Google inc. 2009. Google Earth.
- 438 Guisan A., Thuiller W. 2005. Predicting species distribution: offering more than simple habitat models.  
439 *Ecology Letters* 8:993–1009. DOI: 10.1111/j.1461-0248.2005.00792.x.
- 440 Habit E., Dyer B., Vila I. 2006. Estado de conocimiento de los peces dulceacuícolas de Chile. *Gayana*  
441 *(Concepción)* 70:100–113. DOI: 10.4067/S0717-65382006000100016.
- 442 Heiberger RM., Holland B. 2013. *Statistical analysis and data display : an intermediate course with*  
443 *examples in S-plus, R, and SAS*.
- 444 Hollander M., Wolfe DA. 1999. *Nonparametric statistical methods*. Wiley.

- 445 Jackson ST., Overpeck JT. 2009. Responses of Plant Populations and Communities to Environmental  
446 Changes of the Late Quaternary. *Paleobiology* 26:194–220. DOI: 10.2307/1571658.
- 447 Jähnig SC., Kuemmerlen M., Kiesel J., Domisch S., Cai Q., Schmalz B., Fohrer N. 2012. Modelling of  
448 riverine ecosystems by integrating models: conceptual approach, a case study and research agenda.  
449 *Journal of Biogeography* 39:2253–2263. DOI: 10.1111/jbi.12009.
- 450 Kuhn M. 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*  
451 28:1–26. DOI: 10.18637/jss.v028.i05.
- 452 Kuhn M., Johnson K. 2013a. *Applied Predictive Modeling*. New York, NY: Springer New York. DOI:  
453 10.1007/978-1-4614-6849-3.
- 454 Kuhn M., Johnson K. 2013b. *Applied Predictive Modeling*. New York, NY: Springer New York. DOI:  
455 10.1007/978-1-4614-6849-3.
- 456 Lange K., Townsend CR., Gabriellson R., Chanut PCM., Matthaei CD. 2014. Responses of stream fish  
457 populations to farming intensity and water abstraction in an agricultural catchment. *Freshwater*  
458 *Biology* 59:286–299. DOI: 10.1111/fwb.12264.
- 459 Legendre P., Legendre L. 1998. *Numerical ecology*. Elsevier.
- 460 Leopold LB. 1969. *Quantitative comparison of some aesthetic factors among rivers*. US Geological Survey.
- 461 Lupandin AI. 2005. Effect of Flow Turbulence on Swimming Speed of Fish. *Biology Bulletin* 32:461–466.  
462 DOI: 10.1007/s10525-005-0125-z.
- 463 Manel S., Dias JM., Ormerod SJ. 1999. Comparing discriminant analysis, neural networks and logistic  
464 regression for predicting species distributions: A case study with a Himalayan river bird. *Ecological*  
465 *Modelling* 120:337–347. DOI: 10.1016/S0304-3800(99)00113-1.
- 466 Markovic D., Freyhof J., Wolter C. 2012. Where Are All the Fish: Potential of Biogeographical Maps to  
467 Project Current and Future Distribution Patterns of Freshwater Species. *PLoS ONE* 7:e40530. DOI:  
468 10.1371/journal.pone.0040530.
- 469 Mastrorillo S., Lek S., Dauba F., Belaud A. 1997. The use of artificial neural networks to predict the  
470 presence of small-bodied fish in a river. *Freshwater Biology* 38:237–246. DOI: 10.1046/j.1365-  
471 2427.1997.00209.x.
- 472 McCullagh P. 1984. Generalized linear models. *European Journal of Operational Research* 16:285–292.  
473 DOI: 10.1016/0377-2217(84)90282-0.
- 474 McPherson J., Jetz W. 2007. Effects of species' ecology on the accuracy of distribution models. *Ecography*  
475 30:135–151. DOI: 10.1111/j.0906-7590.2007.04823.x.
- 476 McPherson JM., Jetz W., Rogers DJ. 2004. The Effects of Species' Range Sizes on the Accuracy of  
477 Distribution Models: Ecological Phenomenon or Statistical Artefact? *Journal of Applied Ecology*  
478 41:811–823. DOI: 10.2307/3505798.
- 479 Ministerio de Energía -División de Desarrollo Sustentable. 2016. *Estudio de Cuencas. Análisis de las*  
480 *Condicionantes para el Desarrollo Hidroeléctrico en las Cuencas del Maule, Biobío, Toltén, Valdivia,*  
481 *Bueno, Puelo, Yelcho, Palena, Cisnes, Aysén, Baker y Pascua*. Santiago: Gobierno de Chile.
- 482 Ministerio del Medio Ambiente. 2008a.Ficha de Antecedentes por Especie ID: 165. Available at  
483 [http://www.mma.gob.cl/clasificacionespecies/Anexo\\_tercer\\_proceso/especies\\_actualizadas/Brachy](http://www.mma.gob.cl/clasificacionespecies/Anexo_tercer_proceso/especies_actualizadas/Brachygalaxias_bullocki_PO3R3_RCE_CORREGIDO.doc)  
484 [galaxias\\_bullocki\\_PO3R3\\_RCE\\_CORREGIDO.doc](http://www.mma.gob.cl/clasificacionespecies/Anexo_tercer_proceso/especies_actualizadas/Brachygalaxias_bullocki_PO3R3_RCE_CORREGIDO.doc) (accessed January 23, 2018).

- 485 Ministerio del Medio Ambiente. 2008b.Ficha de Antecedentes por Especie ID: 167. Available at  
486 [http://www.mma.gob.cl/clasificacionespecies/Anexo\\_tercer\\_proceso/especies\\_actualizadas/Cheiro](http://www.mma.gob.cl/clasificacionespecies/Anexo_tercer_proceso/especies_actualizadas/Cheiro)  
487 [don\\_austrele\\_PO3R3\\_RCE\\_CORREGIDO.doc](http://www.mma.gob.cl/clasificacionespecies/Anexo_tercer_proceso/especies_actualizadas/Cheiro) (accessed January 23, 2018).
- 488 Ministerio del Medio Ambiente. 2008c.Ficha de Antecedentes por Especie ID: 183. Available at  
489 [http://www.mma.gob.cl/clasificacionespecies/Anexo\\_tercer\\_proceso/especies\\_actualizadas/Odont](http://www.mma.gob.cl/clasificacionespecies/Anexo_tercer_proceso/especies_actualizadas/Odont)  
490 [esthes\\_mauleanum\\_PO3R4\\_RCE\\_CORREGIDO.doc](http://www.mma.gob.cl/clasificacionespecies/Anexo_tercer_proceso/especies_actualizadas/Odont) (accessed January 23, 2018).
- 491 Ministerio del Medio Ambiente. 2008d.Ficha de Antecedentes por Especie ID: 192. Available at  
492 [http://www.mma.gob.cl/clasificacionespecies/Anexo\\_tercer\\_proceso/especies\\_actualizadas/Tricho](http://www.mma.gob.cl/clasificacionespecies/Anexo_tercer_proceso/especies_actualizadas/Tricho)  
493 [mycterus\\_areolatus\\_PO3R2\\_RCE\\_CORREGIDO.doc](http://www.mma.gob.cl/clasificacionespecies/Anexo_tercer_proceso/especies_actualizadas/Tricho) (accessed January 23, 2018).
- 494 Ministerio del Medio Ambiente. 2011a.Ficha de Antecedentes por Especie ID: 818. Available at  
495 [http://www.mma.gob.cl/clasificacionespecies/ficha5proceso/fichas\\_actualizadas/Aplochiton\\_zebra](http://www.mma.gob.cl/clasificacionespecies/ficha5proceso/fichas_actualizadas/Aplochiton_zebra)  
496 [\\_PO5R7-9\\_RCE.doc](http://www.mma.gob.cl/clasificacionespecies/ficha5proceso/fichas_actualizadas/Aplochiton_zebra) (accessed January 23, 2018).
- 497 Ministerio del Medio Ambiente. 2011b.Ficha de Antecedentes por Especie ID: 825. Available at  
498 [http://www.mma.gob.cl/clasificacionespecies/ficha5proceso/fichas\\_actualizadas/Aplochiton\\_tae](http://www.mma.gob.cl/clasificacionespecies/ficha5proceso/fichas_actualizadas/Aplochiton_tae)  
499 [atus\\_PO5-9\\_RCE.doc](http://www.mma.gob.cl/clasificacionespecies/ficha5proceso/fichas_actualizadas/Aplochiton_tae) (accessed January 23, 2018).
- 500 Morán-Ordóñez A., Lahoz-Monfort JJ., Elith J., Wintle BA. 2017. Evaluating 318 continental-scale species  
501 distribution models over a 60-year prediction horizon: what factors influence the reliability of  
502 predictions? *Global Ecology and Biogeography* 26:371–384. DOI: 10.1111/geb.12545.
- 503 Peredo-Parada M., Martínez-Capel F., Quevedo DI., Hernández-Mascarell AB. 2011. Implementación de  
504 una clasificación eco-hidrológica para los ríos de Chile. *Gayana (Concepción)* 75:26–38. DOI:  
505 10.4067/S0717-65382011000100003.
- 506 Perry AL., Low PJ., Ellis JR., Reynolds JD. 2005. Climate change and distribution shifts in marine fishes.  
507 *Science (New York, N.Y.)* 308:1912–5. DOI: 10.1126/science.1111322.
- 508 Pletterbauer F., Graf W., Schmutz S. 2016. Effect of biotic dependencies in species distribution models:  
509 The future distribution of *Thymallus thymallus* under consideration of *Allogamus auricollis*.  
510 *Ecological Modelling* 327:95–104. DOI: 10.1016/J.ECOLMODEL.2016.01.010.
- 511 QGIS DT. 2015. QGIS geographic information System. Open source geospatial Foundation project.
- 512 R Core Team. 2017. R: A language and environment for statistical computing. Vienna, Austria: R  
513 Foundation for Statistical Computing; 2014.
- 514 Radinger J., Essl F., Hölker F., Horký P., Slavík O., Wolter C. 2017. The future distribution of river fish: The  
515 complex interplay of climate and land use changes, species dispersal and movement barriers.  
516 *Global Change Biology* 23:4970–4986. DOI: 10.1111/gcb.13760.
- 517 Radinger J., Wolter C. 2015. Disentangling the effects of habitat suitability, dispersal, and fragmentation  
518 on the distribution of river fishes. *Ecological Applications* 25:914–927. DOI: 10.1890/14-0422.1.
- 519 Ramezani J., Akbaripasand A., Closs GP., Matthaei CD. 2016. In-stream water quality, invertebrate and fish  
520 community health across a gradient of dairy farming prevalence in a New Zealand river catchment.  
521 *Limnologia - Ecology and Management of Inland Waters* 61:14–28. DOI:  
522 10.1016/J.LIMNO.2016.09.002.
- 523 Ritchie ME. 1998. Scale-dependent foraging and patch choice in fractal environments. *Evolutionary*  
524 *Ecology* 12:309–330. DOI: 10.1023/A:1006552200746.
- 525 Ritchie ME., Olff H. 1999. Spatial scaling laws yield a synthetic theory of biodiversity. *Nature* 400:557–

- 526 560. DOI: 10.1038/23010.
- 527 Segurado P., Araújo MB. 2004. An evaluation of methods for modelling species distributions. *Journal of*  
528 *Biogeography* 31:1555–1568. DOI: 10.1111/j.1365-2699.2004.01076.x.
- 529 Snelder TH., Biggs BJF. 2002. MULTISCALE RIVER ENVIRONMENT CLASSIFICATION FOR WATER RESOURCES  
530 MANAGEMENT. *Journal of the American Water Resources Association* 38:1225–1239. DOI:  
531 10.1111/j.1752-1688.2002.tb04344.x.
- 532 Steen PJ., Zorn TG., Seelbach PW., Schaeffer JS. 2008. Classification Tree Models for Predicting  
533 Distributions of Michigan Stream Fish from Landscape Variables. *Transactions of the American*  
534 *Fisheries Society* 137:976–996. DOI: 10.1577/T07-119.1.
- 535 Stern HS. 1996. Neural Networks in Applied Statistics. *Technometrics* 38:205–214. DOI:  
536 10.1080/00401706.1996.10484497.
- 537 Stockwell DR., Peterson AT. 2002. Effects of sample size on accuracy of species distribution models.  
538 *Ecological Modelling* 148:1–13. DOI: 10.1016/S0304-3800(01)00388-X.
- 539 Suarez-Seoane S., Osborne PE., Alonso JC. 2002. Large-scale habitat selection by agricultural steppe birds  
540 in Spain: identifying species-habitat responses using generalized additive models. *Journal of Applied*  
541 *Ecology* 39:755–771. DOI: 10.1046/j.1365-2664.2002.00751.x.
- 542 Taylor AT., Papeş M., Long JM. 2017. Incorporating fragmentation and non-native species into distribution  
543 models to inform fluvial fish conservation. *Conservation Biology*. DOI: 10.1111/cobi.13024.
- 544 Torgersen CE., Price DM., Li HW., McIntosh BA. 1999. MULTISCALE THERMAL REFUGIA AND STREAM  
545 HABITAT ASSOCIATIONS OF CHINOOK SALMON IN NORTHEASTERN OREGON. *Ecological Applications*  
546 9:301–319. DOI: 10.1890/1051-0761(1999)009[0301:MTRASH]2.0.CO;2.
- 547 Vila I., Fuentes L., Contreras M. 1999. Peces límnicos de Chile. *Boletín del Museo Nacional de Historia*  
548 *Natural. Chile* 48:61–75.
- 549 Vila I., Pardo R., Dyer B., Habit E. 2006. Peces límnicos: diversidad, origen y estado de conservación. In:  
550 Vila I, Veloso A, Schlatter R, Ramírez C eds. *Macrófitas y vertebrados de los sistemas límnicos de*  
551 *Chile*. Editorial Universitaria, 186.
- 552 Woolnough DA., Downing JA., Newton TJ. 2009. Fish movement and habitat use depends on water body  
553 size and shape. *Ecology of Freshwater Fish* 18:83–91. DOI: 10.1111/j.1600-0633.2008.00326.x.
- 554 Zhang Z. 2016. Variable selection with stepwise and best subset approaches. *Annals of translational*  
555 *medicine* 4:136. DOI: 10.21037/atm.2016.03.35.

## FIGURE LEGENDS

**Figura 1.** Study area.

**Figure 2.** Relationship between model fit and size fishes ( $\text{Log}_{10}$  of maximum length). Fit index is area under the curve (AUC) of the receiver operating characteristic (ROC) and the true skill statistic (TSS). *Eco* (model with only ecological predictors), *Antr* (only anthropogenic predictors), *Eco+Antr* (both sets of predictors). Solid line represents the significant relationships.

**Figure 3.** Participation of predictor variables in each model by species. From left to right, the first five anthropogenic variables that only participate in *Antr* models and *Eco+Antr* models, the second ten ecological variables that only participate in *Eco* models, and *Eco+Antr* models. Categories with suffix “use” belong to the land-use predictor, and categories with suffix F.S. belong to source-of-flow predictor.

**Figure 4.** Maps of potential distribution by species and model. Each map represents the distribution of one species. Each color represents which model or sets of models determined a species present in each river section.

**Figure 5.** Venn diagrams representing the river sections defined as having species presence by the different models that coincided in the same river segments for each species. For example, if the *Antr*, *Eco*, and *Eco+Antr* circles completely overlap, the three models selected exactly the same river segments as having species presence. The percentages of river kilometers determined as having species presence by each the model is noted for each species.

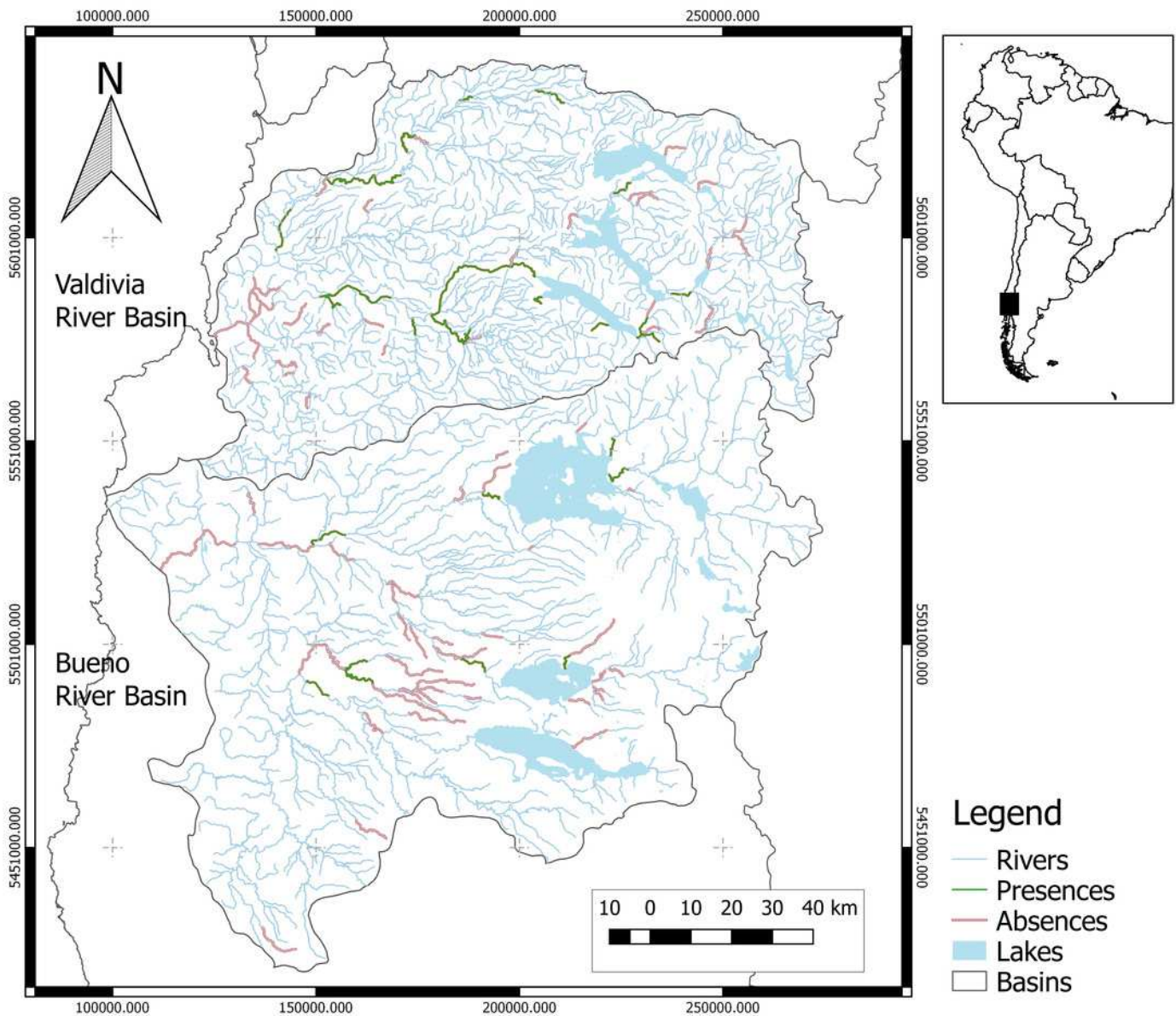
## TABLE LEGENDS

Table 1. Modeled species, modeled presences, and TSS and AUC values for each model with different set predictors.

577 Table 2. Predictor variables used in SDMs, indicating variable type (ecological or anthropogenic), spatial  
578 categories, statistical description, and mean participation by models with different set predictor  
579 variables.

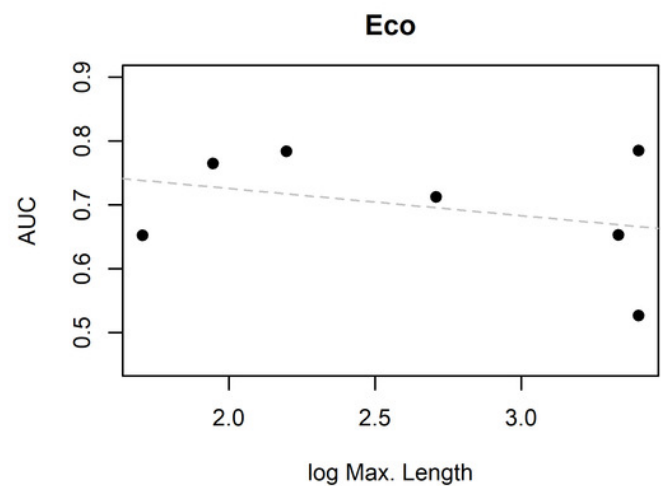
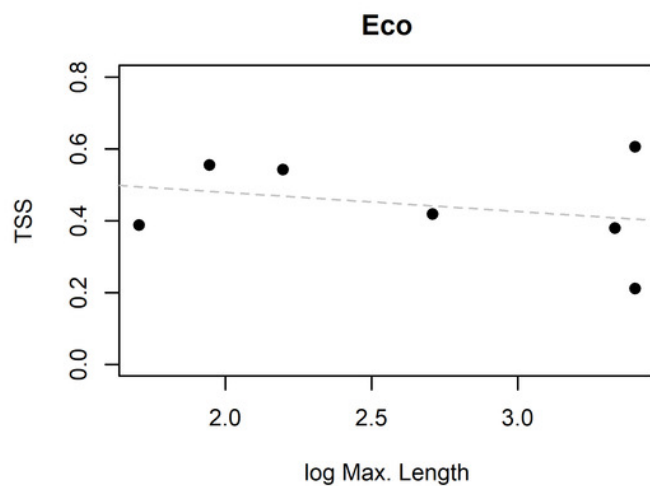
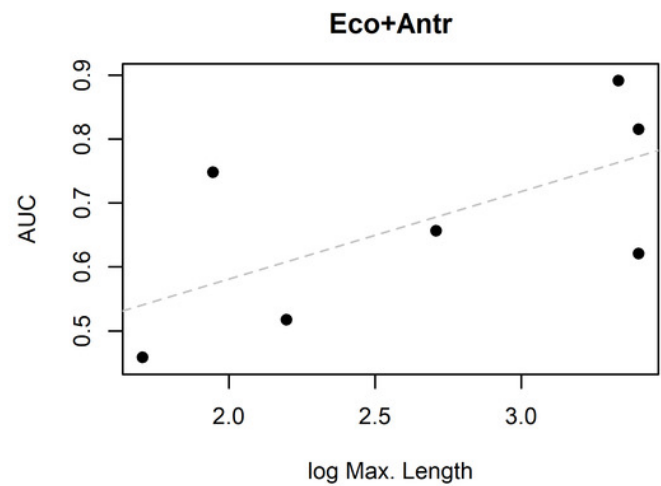
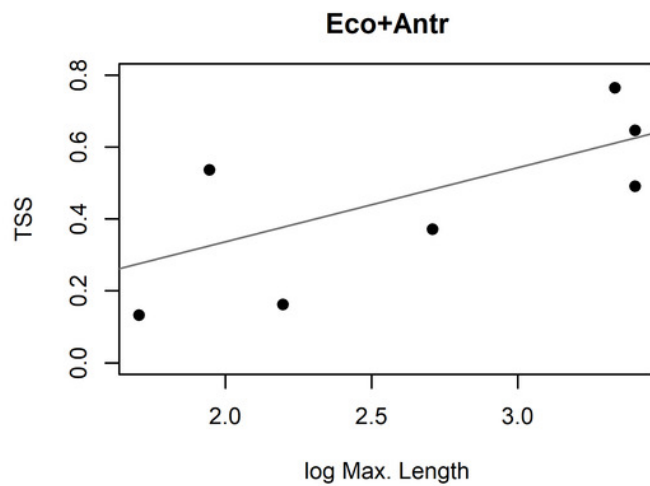
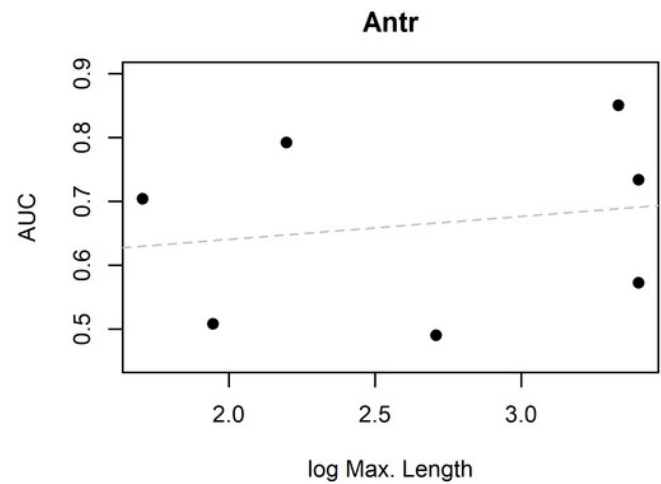
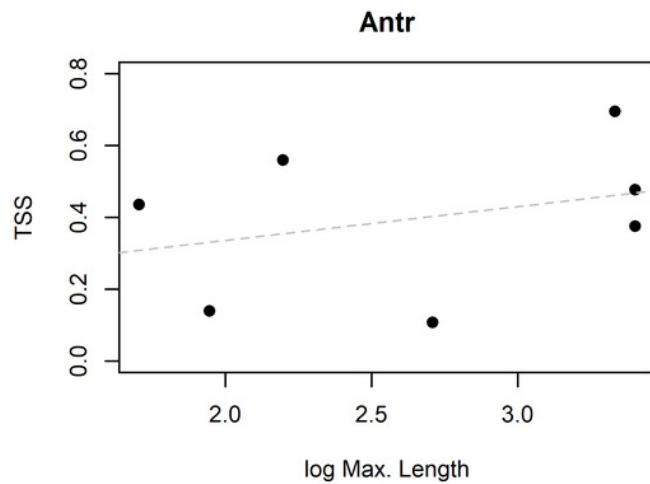
# Figure 1

Study area



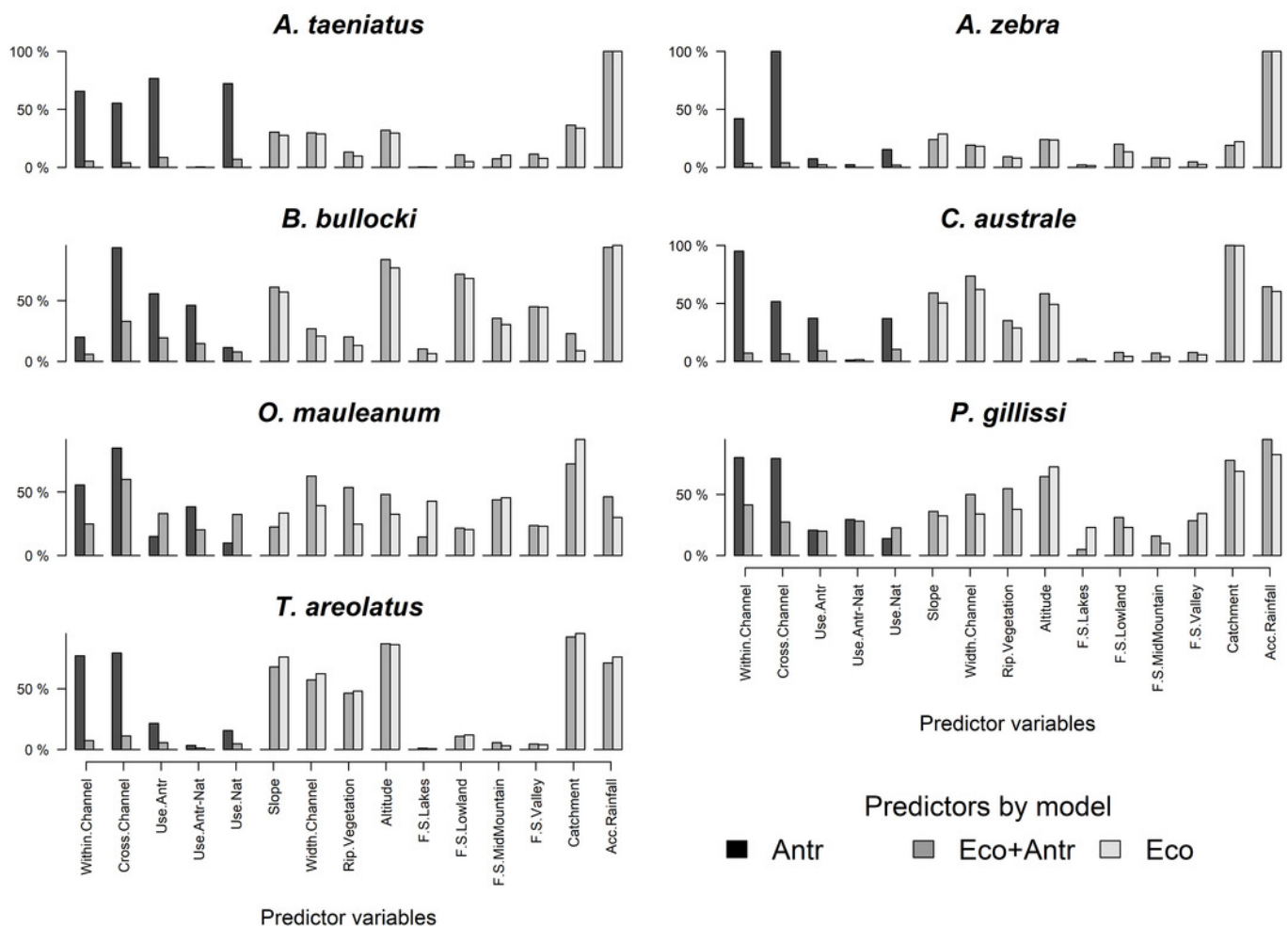
# Figure 2

Relationship between model fit and size fishes ( $\text{Log}_{10}$  of maximum length). Fit index is area under the curve (AUC) of the receiver operating characteristic (ROC) and the true skill statistic (TSS). *Eco* (model with only ecological predictors)



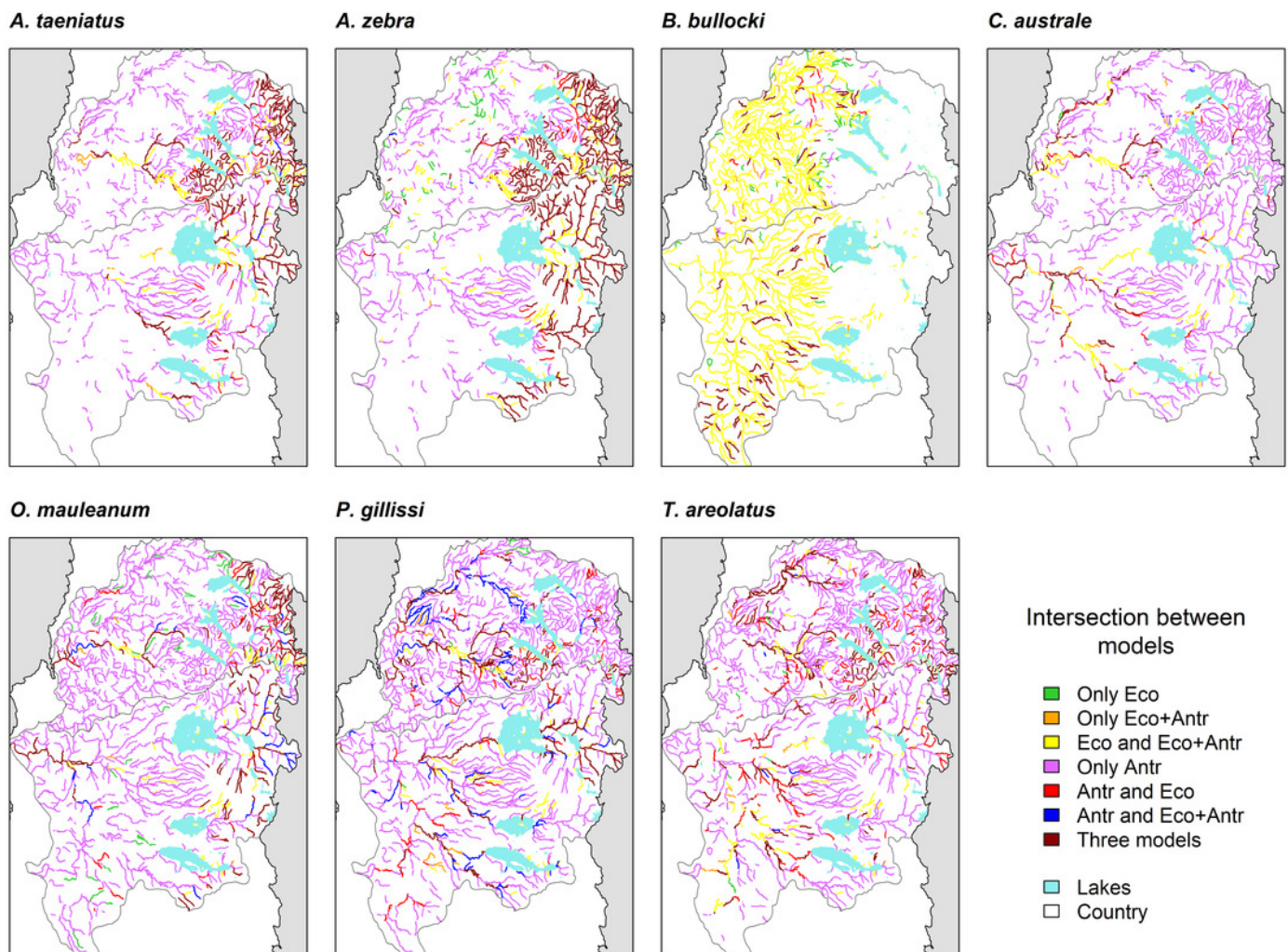
# Figure 3

Participation of predictor variables in each model by species. From left to right, the first five anthropogenic variables that only participate in *Antr* models and *Eco+Antr* models, the second ten ecological variables that only participate in [i



# Figure 4

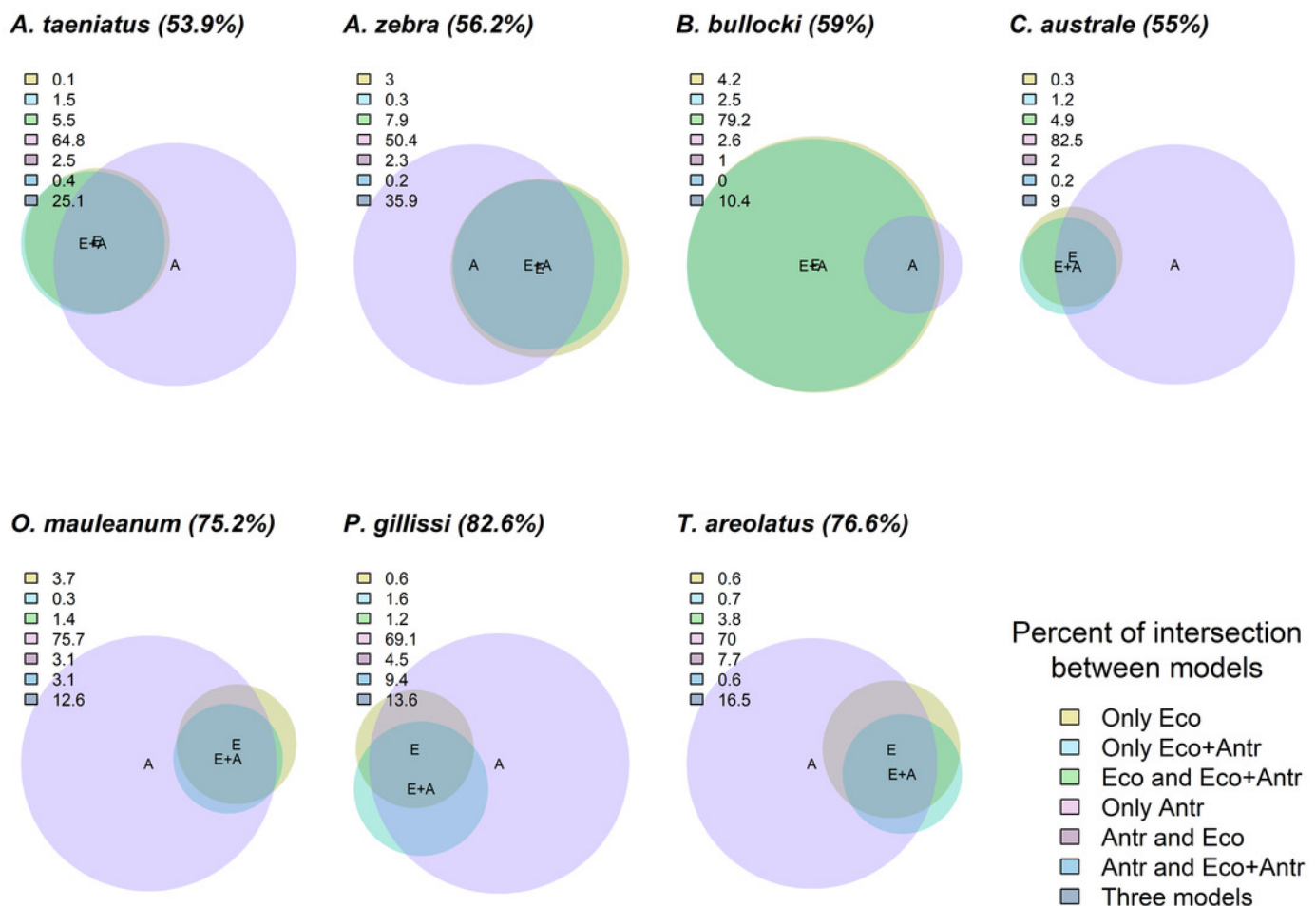
Maps of potential distribution by species and model. Each map represents the distribution of one species. Each color represents which model or sets of models determined a species present in each river section.



# Figure 5

Venn diagrams representing the river sections defined as having species presence by the different models that coincided in the same river segments for each species.

For example, if the *Antr*, *Eco*, and *Eco+Antr* circles completely overlap, the three models selected exactly the same river segments as having species presence. The percentages of river kilometers determined as having species presence by each the model is noted for each species



# **Table 1**(on next page)

Modeled species, modeled presences, and TSS and AUC values for each model with different set predictors.

1 Table 1. Modeled species, modeled presences, and TSS and AUC values for each model with different set predictors.

Species	Distribution in Chile	Max. Length (cm)	Presences	Algorithms selected	Antr AUC	Antr TSS	Eco+Antr AUC	Eco+Antr TSS	Eco AUC	Eco TSS
<i>Aplochiton taeniatus</i> <sup>1</sup>	38° - 55° Lat. S	30	17	RF	0.73	0.48	0.82	0.65	0.79	0.61
<i>Aplochiton zebra</i> <sup>2</sup>	35.88° - 55° Lat. S	28	15	RF	0.85	0.69	0.89	0.77	0.65	0.38
<i>Brachygalaxias bullocki</i> <sup>3</sup>	35.88° - 43.81° Lat. S	5.5	27	GLM	0.70	0.44	0.46	0.13	0.65	0.39
<i>Cheirodon australe</i> <sup>4</sup>	39.32° - 43.81° Lat. S	7	21	RF	0.51	0.14	0.75	0.54	0.76	0.56
<i>Odontesthes mauleanum</i> <sup>5</sup>	32.25° - 43.81° Lat. S	30	9	NNET	0.57	0.38	0.62	0.49	0.53	0.21
<i>Percilia gillissi</i> <sup>6</sup>	32.25° - 43.81° Lat. S	9	33	NNET	0.79	0.56	0.52	0.16	0.78	0.54
<i>Trichomycterus areolatus</i> <sup>7</sup>	29.13° - 43.81° Lat. S	15	36	RF	0.49	0.11	0.66	0.37	0.71	0.42

2 Reference to fish size:

3 <sup>1</sup> Ministerio del Medio Ambiente (2011a)

4 <sup>2</sup> Ministerio del Medio Ambiente (2011b)

5 <sup>3</sup> Ministerio del Medio Ambiente (2008d)

6 <sup>4</sup> Ministerio del Medio Ambiente (2008a)

7 <sup>5</sup> Ministerio del Medio Ambiente (2008b)

8 <sup>6</sup> Froese & Pauly (2017)

9 <sup>7</sup> Ministerio del Medio Ambiente (2008c)

10

## Table 2 (on next page)

Predictor variables used in SDMs, indicating variable type (ecological or anthropogenic), spatial categories, statistical description, and mean participation by models with different set predictor variables

- 1 Table 2. Predictor variables used in SDMs, indicating variable type (ecological or anthropogenic), spatial categories, statistical description, and
- 2 mean participation by models with different set predictor variables.

Predictive Variable	Type	Spatial Scale	Unit	Description	Variable participation		
					<i>Antr</i>	<i>Eco+Antr</i>	<i>Eco</i>
Accumulated rainfall	Ecological	inter-basin	mm	Min: 954 Median: 2302 Max: 5099159		81.5	77.8
Catchment	Ecological	intra-basin	km2	Min: 0.11 Median: 38.022 Max: 15033		60.1	60.0
Source-of-flow	Ecological	intra-basin	Categories: lake, glacier, mountain, foothills, valley, plains	lake: 163 Plains: 891 Foothills: 538 Valley: 619		75.7	35.9
Altitude	Ecological	intra-basin	m.a.s.l.	Min: 2 Median: 264 Max: 1751		56.8	52.9
Slope	Ecological	Inter-segment	m/m	Min: 0 Median: 0.018 Max: 3		43.0	43.7
Channel width	Ecological	Inter-segment	m	Min: 1 Median: 8 Max: 1500		45.6	38.0
Percent riparian vegetation	Ecological	Inter-segment	%	Min: 0 Median: 182 Max: 200		33.2	24.3
Segment land-use	Anthropic	Inter-segment	Categories	Antr: 371 Antr_Nat: 189 Nat: 1252	75.7	35.9	
Cross-channel constructions	Anthropic	Inter-segment	Number of Works	Min: 0 Median: 0	62.2	13.7	

				Max: 8			
Within-channel constructions	Anthropic	Inter-segment	Number of Works	Min: 0 Median: 0 Max: 10	77.7	20.9	

3