PLOS GLOBAL PUBLIC HEALTH

RESEARCH ARTICLE

Species distribution modeling for disease ecology: A multi-scale case study for schistosomiasis host snails in Brazil

Alyson L. Singleton[1]*, Caroline K. Glidden[2,3], Andrew J. Chamberlin[4], Roseli Tuan[5], Raquel G. S. Palasio[5], Adriano Pinter[5], Roberta L. Caldeira[6], Cristiane L. F. Mendonça[6], Omar S. Carvalho[6], Miguel V. Monteiro[7], Tejas S. Athni[2,8], Susanne H. Sokolow[4,9], Erin A. Mordecai[2,10], Giulio A. De Leo[4,10]

1 Emmett Interdisciplinary Program in Environment and Resources, Stanford University, Stanford, California, United States of America, **2** Department of Biology, Stanford University, Stanford, California, United States of America, **3** Institute for Human-centered Artificial Intelligence, Stanford University, Stanford, California, United States of America, **4** Department of Oceans, Hopkins Marine Station, Stanford University, Pacific Grove, California, United States of America, **5** Pasteur Institute, São Paulo, Brazil, **6** Fiocruz Minas/Belo Horizonte-Minas Gerais, Belo Horizonte, Brazil, **7** Geoinformation & Earth Observation Division, National Institute for Space Research (INPE), São Paulo, Brazil, **8** Harvard Medical School, Boston, Massachusetts, United States of America, **9** Marine Science Institute, University of California Santa Barbara, Santa Barbara, California, United States of America, **10** Woods Institute for the Environment, Stanford University, Stanford, California, United States of America

* asinglet@stanford.edu

Citation: Singleton AL, Glidden CK, Chamberlin AJ, Tuan R, Palasio RGS, Pinter A, et al. (2024) Species distribution modeling for disease ecology: A multi-scale case study for schistosomiasis host snails in Brazil. PLOS Glob Public Health 4(8): e0002224. https://doi.org/10.1371/journal.pgph.0002224

Editor: Kamran Shaukat Dar, The University of Newcastle, AUSTRALIA

Received: July 6, 2023

Accepted: July 17, 2024

Published: August 2, 2024

Data Availability Statement: All data and methods used in this analysis are publicly available and can be found at https://github.com/alyson-singleton/sdm-disease-ecology-multi-scale.

Funding: This work was supported by the Belmont collaborative Forum on Climate, Environment, and Health (NSF-ICER 2024383 and NSF-ICER 2024385) and by FAPESP (2019/23593-3) in a collaborative research action with Belmont Forum. GDL and EM are also supported by the Stanford

Abstract

Species distribution models (SDMs) are increasingly popular tools for profiling disease risk in ecology, particularly for infectious diseases of public health importance that include an obligate non-human host in their transmission cycle. SDMs can create high-resolution maps of host distribution across geographical scales, reflecting baseline risk of disease. However, as SDM computational methods have rapidly expanded, there are many outstanding methodological questions. Here we address key questions about SDM application, using schistosomiasis risk in Brazil as a case study. Schistosomiasis is transmitted to humans through contact with the free-living infectious stage of *Schistosoma* spp. parasites released from freshwater snails, the parasite's obligate intermediate hosts. In this study, we compared snail SDM performance across machine learning (ML) approaches (MaxEnt, Random Forest, and Boosted Regression Trees), geographic extents (national, regional, and state), types of presence data (expert-collected and publicly-available), and snail species (*Biomphalaria glabrata*, *B. straminea*, and *B. tenagophila*). We used high-resolution (1km) climate, hydrology, land-use/land-cover (LULC), and soil property data to describe the snails' ecological niche and evaluated models on multiple criteria. Although all ML approaches produced comparable spatially cross-validated performance metrics, their suitability maps showed major qualitative differences that required validation based on local expert knowledge. Additionally, our findings revealed varying importance of LULC and bioclimatic variables for different snail species at different spatial scales. Finally, we found that models using publicly-available data predicted snail distribution with comparable AUC values to models using expert-collected data. This work serves as an instructional guide to SDM

Center for Innovation in Global Health. GDL is also partially supported by NSF-DEB DEB-2011179 (EEID). EM is also supported by grants from the National Science Foundation (DEB-2011147, with Fogarty International Center), the National Institutes of Health (R35GM133439, R01AI168097, R01AI102918), and the Woods Institute for the Environment. TA is supported by the National Institute of General Medical Sciences (T32GM144273). RP is supported by FAPESP (2021/10212-1). The authors received no specific funding for this work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

methods that can be applied to a range of vector-borne and zoonotic diseases. In addition, it advances our understanding of the relevant environment and bioclimatic determinants of schistosomiasis risk in Brazil.

Introduction

Species distribution models (SDMs) have become increasingly popular tools in the field of disease ecology to profile transmission risk for vector-borne, zoonotic diseases, and environmentally-mediated diseases, i.e., diseases whose transmission involves a non-human host or vector species, such as mosquitoes (malaria, dengue, Zika), flies (leishmaniasis, sleeping sickness), ticks (Lyme disease), triatomine bugs (Chagas disease), and snails (schistosomiasis, fascioliasis) [1–6]. By using presence data of non-human hosts and remotely-sensed data of potential environmental covariates, SDMs are correlative models that can predict species habitat suitability across areas not sampled by field collection programs [7–9]. These models are typically used to create high-resolution maps of an inferred species distribution across a geographic area of interest, which can reflect areas where disease transmission may be possible. In combination with other processes that influence transmission, such as additional reservoir host distributions or other disease exposure variables, these predictions can directly inform the understanding of the pathogenic landscape of environmentally-mediated diseases [10].

SDMs are a powerful tool applied in a number of fields, including disease ecology [11, 12], epidemiology [13], and conservation [14, 15], among many others. Species distribution modeling works by using species presence/absence data to identify covariates that are predictive of a species presence. Because true absence data are not typically available, SDMs often use "background" or "pseudo-absence" data to simulate locations where an organism could have been sampled but was not [16, 17]. SDMs use various machine learning methods to identify a suite of covariates that can accurately predict the presence or absence of the organism in geographic space, using flexible functional relationships between predictors and responses that can include nonlinearities and interactions [8, 9]. Model inputs can vary in spatial and temporal resolution and extent. Many algorithms are available for model training and testing, and they differ in how they handle covariate-outcome relationships [18]. SDMs are cross-validated by leaving out part of the data in model training in order to assess model performance on out-of-sample data, often performed in a spatially-structured way [19]. The outputs of interest include geographic maps of species presence suitability, lists of variables selected as important predictors, and the functional forms of relationships between predictors and presence. A glossary of terms and concepts central to the SDM literature are summarized for reference in Table 1.

Increased access to large-scale, remotely-sensed environmental data [20, 21] and species presence databases [22], such as the Global Biodiversity Information Facility [23], has spurred rapid expansion of these methods. Further, recent decades have brought rapid development of statistical models and machine learning algorithms that can be applied to species distribution models, such as regularized regression [24], decision tree [25], Bayesian [26], neural network [27], and ensemble methods [28], among many others. Although many machine learning methods have grown in popularity due to their flexibility, ability to model covariate interactions, and increasing accessibility in common programming languages like R, no single method has fully eclipsed its counterparts [18, 29]. Due to both their popularity in the literature and their consistently high performance, we chose three modeling methods to investigate in this study: Maximum Entropy (MaxEnt), Random Forest (RF), and Boosted Regression

Table 1. Glossary of terms and concepts central to SDM methodology.

Term/concept	Summary	Citation
Presence records	Observed presence of a species, usually with latitude, longitude, and date.	[41]
True absence records	Observed absence of a species, also with latitude, longitude, and date. Often difficult to obtain with certainty.	[42]
Pseudo-absence / background records	Locations drawn from the landscape of interest where an organism could have been sampled but was not. Could by chance include presence areas.	[16]
Thinning	Reducing a set of records (presence, absence, background) so only one record is retained for each grid cell.	[43]
Environmental covariates	Variables hypothesized to impact or predict species presence, such as temperature, precipitation, or land-use type. Often remotely-sensed raster images.	[41]
Multicollinearity	When predictor variables are correlated to one another. Can potentially result in misidentification of relevant predictors, their importance, or their relationship with the outcome.	[44]
Resolution size	Size of the grid cells at which probabilities are predicted. All predictor variables need to be input into the model with the same resolution.	[35]
Geographic extent	The geographic area of interest for which probabilities are estimated.	[41]
Model type	Choice of statistical or machine learning algorithm.	[18]
Cross-validation / spatial cross-validation	Validation technique that repeatedly leaves out parts (i.e., "folds") of the data in model training in order to assess performance on out-of-sample data. Spatial cross-validation separates the folds by spatial clusters.	[19]
Discrimination	The ability for models to distinguish between presence records and absence/background records. Often estimated by ROC-AUC values (defined below).	[45]
Threshold	The cutoff value at which continuous model output probabilities (ranging from 0 to 1) are split and labelled as presence points (above the threshold) or absence/background points (below the threshold). The threshold value is not required to be 0.5, but can be optimized through AUC or pAUC calculations (discussed below).	[45]
Sensitivity	The proportion of presences correctly identified as presences. Models with high sensitivity tend to build prediction maps that look more full.	[45]
Specificity	The proportion of absences/backgrounds correctly identified as absences/backgrounds. Models with high specificity tend to build prediction maps that look more sparse.	[45]
ROC-AUC	A measure of discrimination that compares the false positive rate (i.e., 1—specificity) versus sensitivity across all possible thresholds. A value of 1 indicates perfect discrimination and 0.5 or less indicates performance is no better than random. Often referred to as "AUC."	[45]
Partial ROC-AUC	An ROC curve bounded above, below, or between sensitivity, specificity, or other threshold values (i.e., the maximum AUC value is < 1). Recommended when comparing models with varying ranges in suitability probability predictions. Often referred to as "pAUC."	[46]
AICc	Information criteria measure for regression models that can help balance model complexity and goodness-of-fit during model selection.	[47]
TSS (true skill statistic)	Sensitivity + specificity—1 (i.e., values of zero or less indicate the performance is no better than random). A measure of discrimination designed to be less sensitive to species prevalence values than ROC-AUC.	[38]
Calibration	The degree to which the observed proportion of presences in a grid cell equates to the model estimated probability. Often evaluated with a calibration graph.	[48]
Variable importance	Measurement to estimate how much each covariate contributes to model performance. SHapley Additive exPlanations (SHAP) are a particularly useful method because they are model agnostic.	[49]
Partial dependence plots	Line plots that depict the marginal effect each predictor has on suitability probabilities.	[50]

https://doi.org/10.1371/journal.pgph.0002224.t001

Tree (BRT) models [18]. MaxEnt—a regularized, regression-based model—has long been a well-established method for presence-only applications [30], while flexible, decision tree model types such as RF and BRT have gained more recent popularity [31]. Along with the expansion of model types, there has been additional SDM methodological development, including optimization of sampling techniques for "background" or "pseudo-absence" points [17, 32], increased rigor for input variable selection [33, 34], investigation on resolution size [35], defense of spatial cross-validation techniques [19], integration of ecological theory [36, 37], development of gold-standard model evaluation measures [38], and updated guidelines for method-specific reproducibility standards [39, 40].

Despite these advances, many of these new methods have not been recently documented, especially not in a cohesive, accessible manner for scientists new to SDMs or those interested

in adopting new methods [31]. To our knowledge, there has not been an analysis comparing machine learning algorithms, data sources, and geographic extents in combination and assessing the consequences for presence probabilities and covariate relationships. We hypothesize that algorithm performance will vary across geographic scales given differences in model structure, such as ability to handle covariate interactions and potential to overfit [18]. Additionally, there are very few analyses that directly compare the effects of using GBIF presence records versus records from expert-executed field collection programs. Given the known spatial bias in GBIF data, we ask how well GBIF data can approximate predictions created from expert-collected data sources [51, 52]. Finally, although there has been discussion on the effect of resolution size [35], there has been limited discussion on how SDM performance varies across areas of differing geographic extent when resolution size is held constant.

In an effort to answer these methodological questions in a biologically and epidemiologically relevant study system, we will use the intermediate hosts of *Schistosoma mansoni* Sambon, 1907—*Biomphalaria* Preston, 1910 snails—as a case study. Simultaneously, we will make substantial contributions to knowledge on predicting schistosomiasis risk in Brazil. Schistosomiasis is a debilitating parasitic disease caused, in Brazil, by *S. mansoni*, a parasite that relies on both freshwater *Biomphalaria* snails and human beings to complete its life cycle [53]. In Brazil, approximately 6 million people are infected and 25 million live in areas where they are at risk of infection [54]. The disease predominantly impacts poor communities dependent on open water sources for occupational activities or other components of daily life [55, 56]. More recently, schistosomiasis transmission has also been recorded in urban and peri-urban areas, impacting people who are either without access to basic sanitation services or whose sewage systems overflow in times of heavy rainfall [57, 58].

Because *Biomphalaria* freshwater snails are obligate intermediate hosts of *S. mansoni* parasites, SDMs of the non-human hosts of schistosomiasis allow us to predict areas of suitable snail habitat where transmission may be possible. There are three competent *Biomphalaria* snail hosts in Brazil: *Biomphalaria glabrata* Say, 1818, *Biomphalaria straminea* Dunker, 1848, and *Biomphalaria tenagophila* D'Orbigny, 1835. Because snails are ectotherms (i.e., their body temperature is dependent on their environment), their reproduction, survival, and dispersal are strongly affected by their surrounding temperature [59]. The snails live in slow-moving freshwater, including permanent and temporary sources, which are both influenced strongly by precipitation and drainage patterns [60]. Land-use and land-cover (LULC) characteristics affect snail presence through multiple pathways, including affecting temperatures through changes in tree canopy and vegetation cover and influencing water patterns through deforestation and agriculture [61]. Finally, chemical factors and soil properties—such as pH and soil water content—are known to impact the survival of *Biomphalaria* snails, due to their impact on freshwater quality [62].

SDMs capture the snails' biological relationships to these environmental factors and build predictive risk maps that can help to target disease intervention efforts such as mass drug administration [63]. There have been multiple studies using SDMs to predict suitable snail habitat across multiple geographical scales in Brazil, from national [64–66] to sub-national analyses, including those specific to areas within Pernambuco [67], São Paulo [68], and Minas Gerais [69, 70]. However, all of these analyses test only MaxEnt models, with the exception of Guimarães et al., 2009 who used an indicator kriging procedure [69]. Moreover, with the exception of Palasio et al., 2021, the quality and quantity of accessible, remotely-sensed environmental data has grown substantially since time of publication [68]. Finally, our group has collected a large dataset of presence records throughout Brazil that reflect best expert knowledge of the constraints on snail habitat, presenting an alternative to publicly available GBIF presence data. Therefore, *Biomphalaria* snails in Brazil provide a ripe opportunity to compare

and contrast current methods on SDMs, providing a rare comparative case study to guide SDM approaches for disease ecology and contributing updated risk models that can guide Brazil's schistosomiasis elimination efforts [71].

We compare multiple combinations of SDM methods—three machine learning algorithms, two data sources, and three geographic extents—and assess the consequences for suitability probabilities and covariate relationships of three snail species. We address the questions: How do statistical/machine learning models compare depending on research question or application of interest? How do model accuracy, variable importance, and geographic predictions vary across spatial scales? How does model performance compare using expert-collected data versus publicly-available data?

Methods

All data and methods used in this analysis are publicly available and can be found at https://github.com/alyson-singleton/sdm-disease-ecology-multi-scale.

Species data and background sampling

We acquired *B. glabrata*, *B. straminea*, and *B. tenagophila* presence data from two main sources (1) an ongoing, Brazil-wide field program supported by multiple government-funded groups across Brazil, including the Coleção de Malacologia Médica, Fundação Oswaldo Cruz (CMM-Fiocruz) and the Coordination for Disease Control of the State Health Secretariat of São Paulo (CCD-SP) [72–79] and (2) the Global Biodiversity Information Facility (GBIF), a database of publicly available presence records commonly used to build SDMs [22].

The Brazil-wide field collection program, hereafter referred to as the expert-collected dataset, consisted of 11,299 total snail records that spanned 1992–2019 and included 25 species. As part of national efforts to control schistosomiasis, the Brazilian Ministry of Health has approved routine collection and monitoring of *Biomphalaria* snail species. Geographical coordinates of each collection site were acquired with a Garmin eTrex GPS device and species identification was done using morphological and molecular tools. Prior to model input, all records were spatially filtered such that only one presence record was retained for each 1km grid cell (i.e. "thinned to 1km") to minimize pseudo-replication and oversampling bias [43]. After each species was separately thinned to 1km, the dataset was reduced to 972 records of our snail hosts of interest: 305 *B. glabrata*, 396 *B. straminea*, and 271 *B. tenagophila* presence points (**Fig 1**, **Table 2**).

To compare model performance between expert-collected and publicly available GBIF data and to create a background dataset (described below), we constructed a GBIF dataset by searching Brazil for all species included in the expert-collected dataset and records of all freshwater animals found in South America, as defined by the International Union for Conservation of Nature [81]. This resulted in a total of 74,960 records that spanned 1985–2020, included over 2,000 species, and reduced to 193 records of our snail hosts of interest—29 *B. glabrata*, 28 *B. straminea*, and 136 *B. tenagophila*—post thinning. Our inclusion criteria for GBIF records were (i) year was between 1985–2020, (ii) latitude and longitude each included at least three decimal places and (iii) basis of record excluded "fossil specimen" and "machine observation" to ensure that the record was field-collected at the latitude and longitude reported and was identified by a human. For our snail hosts of interest we also required a complete species taxonomic identification. We limited our comparison of expert-collected versus GBIF data to *B. tenagophila* in São Paulo due to lack of sufficient data availability in other areas.

Given a lack of true absence data, we constructed a background dataset of freshwater animals across Brazil as our comparison group, thereby representing the freshwater landscape in

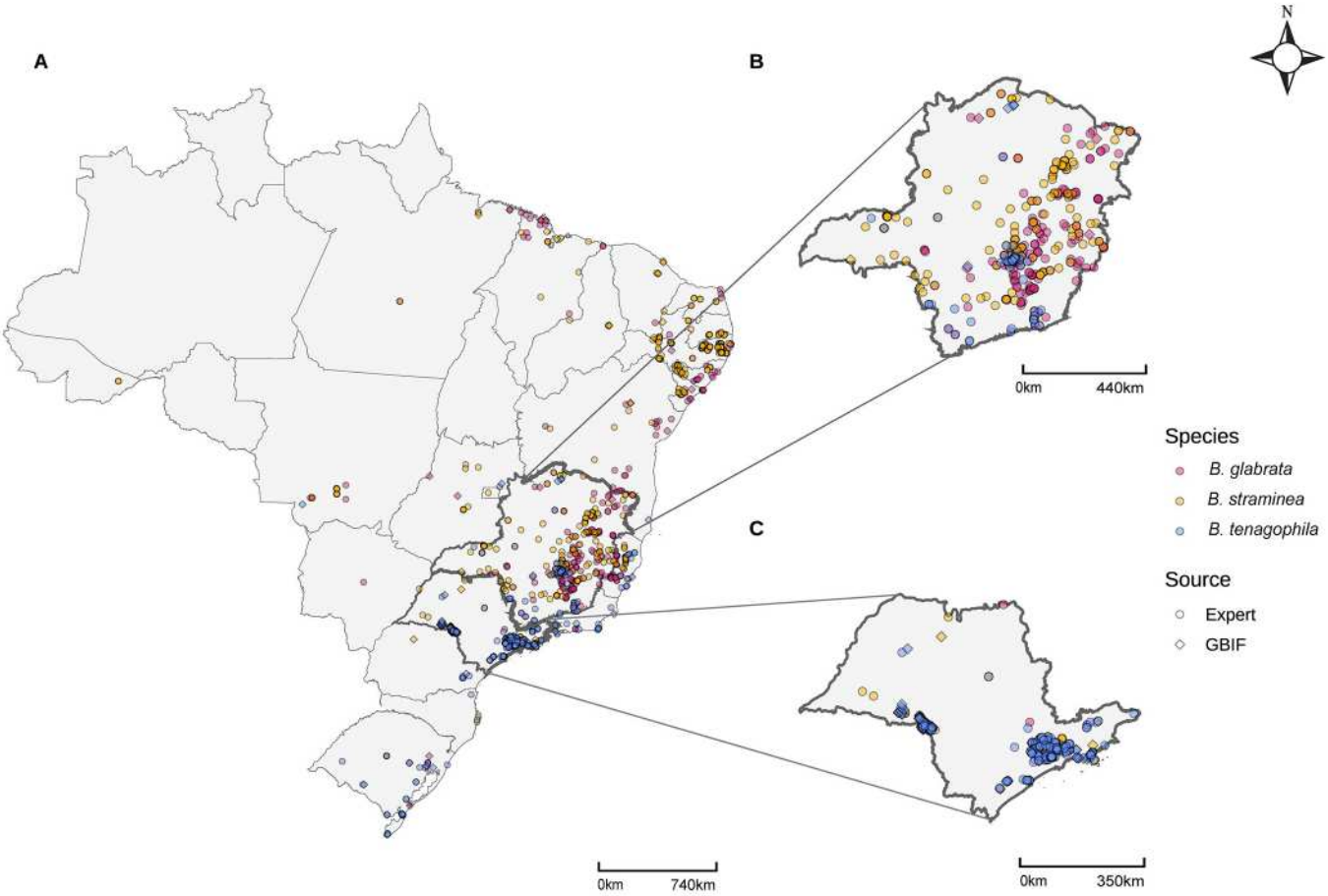

Fig 1. *Biomphalaria* presence points by species (color) and source (shape), thinned to 1 km. A) National, B) Minas Gerais, C) São Paulo. Maps were built in R (version 4.2.2) using shapefiles from the *geobr* package [80].

https://doi.org/10.1371/journal.pgph.0002224.g001

which snails could plausibly be sampled. Species distribution models are often constructed using presence data only, without data on true absences of the species. To do so, models typically calculate the probability of species presence relative to a set of randomly-sampled

Table 2. *Biomphalaria* presence point quantity by species, scale, and source, thinned to 1 km.

Species	Scale	Source	Presence point quantity	Proportion of national data
B. glabrata	National	Expert+GBIF	327	1
	Sudeste	Expert+GBIF	244	0.74
	Minas Gerais	Expert+GBIF	200	0.61
B. straminea	National	Expert+GBIF	414	1
	Sudeste	Expert+GBIF	217	0.52
	Minas Gerais	Expert+GBIF	162	0.39
B. tenagophila	National	Expert+GBIF	358	1
	Sudeste	Expert+GBIF	318	0.89
	São Paulo	Expert+GBIF	234	0.65
	São Paulo	Expert*	169	NA
	São Paulo	GBIF*	115	NA

*Source comparisons only made for *B. tenagophila* in São Paulo (the only scale where data quantities were comparable and sufficient).

https://doi.org/10.1371/journal.pgph.0002224.t002

background points across an area in which a species hypothetically could have been sampled but was not. Instead of random sampling, we use presences of ecologically similar species (i.e., freshwater animals) as "pseudo-absence" or "background" points to control for sampling effort and to capture the relationships with environmental covariates that distinguish the presence of the species of interest from that of others [16]. We would expect sampling efforts of freshwater animals to be similar for that of our species of interest. By constructing a background dataset of freshwater animals, we are better able to represent typical sampling practices of freshwater species in Brazil, rather than selecting background points randomly throughout the study area. Using background records means the model will predict whether or not a record is a presence (labeled as 1) or a background point (labeled as 0), rather than 0's representing true absences. The extent of the background dataset should be also chosen to represent the environmental variation of the study area [16]. Our background dataset was a combination of (1) the remaining expert-collected data after excluding our three species of interest (4.8%) and (2) the publicly-available GBIF data described above (95.2%), which included a total of 2,091 freshwater animal species and 77,785 presence records. Each background dataset was built by sampling two times the number of presence data points for each model (i.e., a model with 100 presence points was given 200 background points): this ratio was selected to balance the sample between groups [82], while providing sufficient data to represent all environments and promote model convergence [17, 83]. Background points were sampled without replacement across a probability distribution that maintained the frequency of background points per 1km grid cell. Therefore, we retained a maximum of one record per grid cell, generating a "background mask" that helped address sampling bias concerns [16, 84].

Environmental data and multicollinearity analysis

We used high-resolution (1km) climate, hydrology, soil property, and land-use/land-cover (LULC) data to describe the environmental conditions associated with each species presence record and background sample. We limited the number of covariates to variables previously found to impact snail presence for ease of interpretation and comparison between model design choices [44]. Climate data were obtained from CHELSA (version 2.1), a high resolution ($1km^2$) global downscaled climate data set [85]. Four climatology variables, averaged over thirty years (1981–2010), were included in the analysis: temperature seasonality (bio4), mean temperature of coldest quarter (bio11), mean precipitation of wettest quarter (bio16), and mean precipitation of driest quarter (bio17). Hydrology data (height above nearest drainage—HND—and soil water percentage) were obtained from the Merit Hydro data [86] and OpenLandMap Soil Water Content [87], respectively, and soil property data (pH and clay) was obtained from OpenLandMap Soil pH in H_2O [88] and OpenLandMap Clay Content [89], respectively. Because hydrology and soil variables were measured at finer spatial resolution than the climate data, we scaled them up to the maximum value (HND) or mean value (water content, pH, clay content) for each $1km^2$ grid cell. Finally, our two LULC covariates—distance to high population density and proportion of temporary crop cover during the year of sampling—were constructed from WorldPop [90] and MapBiomas [91], respectively. High population density was defined as a 1km grid cell with a density of at least 1500 inhabitants per km^2, per the World Bank definition [92]. Proportion of temporary crop cover was defined—in natural areas—as farming areas where it was not possible to distinguish between pasture and agriculture and—in urban areas—as areas of urban vegetation, including cultivated vegetation, natural forest, and non-forest vegetation [91]. We selected these two LULC variables based on our team's on-the-ground knowledge of snail presence [93]. In total, we provided our models with 12 environmental covariates (**Fig A in S1 Text**), none of which had pairwise Pearson

correlation coefficients above 0.7 with any of the other covariates [44]. Although our studied models can handle multicollinearity when calculating probabilities, collinear variables obscure the variable importance and partial dependence plot interpretation [44].

Geographic extent

To investigate model performance across varying geographic extent, we created models spanning national, regional (Sudeste, composed of four states: Espírito Santo, Minas Gerais, Rio de Janeiro and São Paulo), and state (Minas Gerais and São Paulo) extents in Brazil. The region and states of interest were chosen based on the quantity of data available to input into the models. Past studies have shown that model performance substantially declines with fewer than 30–50 presence records [94, 95]. We selected only states with greater than 100 presence records for a species of interest: Minas Gerais for *B. glabrata* and *B. straminea* and São Paulo for *B. tenagophila* (Table 2).

Statistical model type

To compare between machine learning modeling methods, we built three model types: Maximum Entropy (MaxEnt), Random Forest (RF), and Boosted Regression Tree (BRT). All models were built using the R program (version 4.2.2).

MaxEnt uses a maximum-entropy approach to estimate a species' relative probability distribution in response to environmental covariates [24]. MaxEnt models create smooth fitted curves, which can facilitate straightforward ecological interpretation [16]. The degree to which this "smoothness" is enforced can be controlled through choice of regularization settings and by which feature types are provided, where options include linear, quadratic, hinge, threshold, and product features [16]. Product features are equivalent to interaction terms in regression, and they allow for limited interactions between covariates [16]. We allow MaxEnt all five of these options and use the *trainMaxNet* function from the *enmSdmX* package, which includes an L1 regularization feature [96].

On the other hand, RF, BRT, and other tree-based methods provide enhanced flexibility that allow for automatic fitting of precise interactions between the environmental covariates [97]. RF models take bootstrap samples from the training data and fit a decision tree to each sample [83]. These individual trees can have high variance (i.e., depend heavily on the training data), but have strong generalizability when averaged together to make a prediction over all fitted trees [97]. RF models use random subsets of the available predictor variables (parameter *mtry*) on each decision tree split, which results in decorrelated trees and subsequently improves model performance [83, 98]. Due to its relative ease of implementation and conceptual simplicity, RF has become a common SDM approach [18]. However, RF models have the potential to overfit, especially when provided data with high class imbalance (e.g., many more background points than presences) [83]. We use the *trainRF* function from the *enmSdmX* package [96], which is a wrapper of the *randomForest* function from the *randomForest* package [99].

BRT is similar in structure to RF, but the decision trees are recursively updated as the algorithm learns. During each step of the learning process, BRT fits new trees to the residuals for the previously fitted trees, which allows the algorithm to improve on the observations that are not yet predicted correctly [25]. We use the *trainBRT* function from the *enmSdmX* package [96], which is a wrapper of the *gbm.step* function from the *dismo* package [100]. Similar to RF, BRT also has the potential to overfit to training data but can better handle class imbalance and missing data due to its additional hyperparameters [25]. While these hyperparameters make BRT the most flexible model of the three included in our analysis, they require an additional

tuning step that can be computationally expensive [25]. As of now, no one model type has fully eclipsed the others as the SDM standard, but tree-based methods have been shown to improve performance in multiple settings [18].

Model evaluation

Our goals in model evaluation were first to assess the accuracy of each model in classifying presence versus background (how well does each model classify snail distribution?), second to compare model accuracy among methods (which machine learning approach represents the data best?), third to assess the importance of different environmental covariates and the shapes of their relationships with presence (what environmental characteristics are associated with the observed snail distribution and with what functional form?), and fourth to compare this variable importance and functional form among model methods (are the relationships between predictors and snail presence consistent among models?). Before quantifying accuracy, we first assessed model biological realism qualitatively by using expert opinion to visually compare maps where each pixel shows the mean value across 10 bootstrapping iterations in which models were provided 80% of presence records available for each species at the scale of interest. Our group of experts consisted of scientists from CMM-Fiocruz and CCD-SP who have studied and organized field collection of *Biomphalaria* snails in Brazil for over three decades. Second, we assessed accuracy using four out-of-sample model performance metrics, as described below, calculated through ten-fold spatial cross-validation (a process where folds are divided in space instead of through random sampling, to avoid inflating SDM performance measures due to spatial autocorrelation of environmental covariates [19]). We determined the ten spatial folds using a k-means clustering algorithm where the size of folds was allowed to vary [101]. We choose this mode of data partitioning to prioritize the degree of spatial separation while also minimizing unnecessary computational time, as compared to checkerboard, *n—1* jackknife, or block methodologies [101, 102]. Each fold was required to have at least one presence and one background point.

To determine each model's discrimination ability, we calculated sensitivity, specificity, the area under the receiver operator characteristic curve (ROC-AUC, often referred to as AUC), the partial ROC-AUC (pAUC), and true skill statistic (TSS). Sensitivity is the proportion of presences correctly identified as presences, and specificity is the number of background points correctly identified as background records. ROC-AUC measures the false positive rate (i.e., 1—specificity) versus sensitivity across all possible thresholds [45]. An AUC value of 1 indicates perfect discrimination and 0.5 or less indicates the performance is no better than random). We allowed AUC threshold values to vary across each fold for each model [103]. TSS is defined as *sensitivity + specificity—1* (i.e., values of zero or less indicate the performance is no better than random) and is designed to be less sensitive to species prevalence values [38]. Given our interest in comparing each of our models' ability to distinguish relative suitability of sites, output suitability probabilities were scaled such that all distributions ranged from 0 to 1 [48, 95]. We also calculated partial ROC-AUC (pAUC) to better compare performance between model types, as pAUC calculates the AUC values bounded between each model's range of predicted probabilities [46]. We implemented pAUC as described in Peterson et al., 2008 [46] using the *NicheToolbox* package [104], which also substitutes "proportion of area predicted as present" for the *1 –specificity* x-axis. Background points are not actual absences; they can, in fact, represent areas of suitable species habitat. This substitution eliminates their impact on pAUC values. Instead, models are only evaluated on their ability to correctly identify presence points, while still being penalized for overprediction of presence areas [46]. With the intention of achieving an "out-of-sample" pAUC measure, we compared each test fold's presence

predictions with the training data's total range of predictions. We also include a measure of pAUC significance to establish if models are performing better than random since pAUC null hypotheses can be <0.5 due to bounding between each model's range of predicted probabilities [46]. AICc is another common, useful measure for balancing model complexity and goodness-of-fit during model selection [47]. However, we are unable to calculate AICc for two of our models (RF and BRT) due to their model structure (i.e., no obvious likelihood function or number of parameters) and therefore do not calculate the measure in this study. Calibration is a measure of how well the observed proportion of presence records in a grid cell equals the model estimated probability (i.e., 60% of grid cells predicted with a probability of 0.6 contain a presence record [48]). The main calibration evaluation technique is a calibration graph, which plots model probability estimates against the observed proportion of presences, and is predominantly used in studies with true absence data [48]. Although not applicable for this analysis, there are other situations where the calibration of the model is an additional aspect that should be tested, such as when evaluating estimates of true prevalence [48].

Finally, partial dependence plots and variable importance measures were calculated across the ten folds for each model to investigate each covariate's contribution to model accuracy and functional relationship with presence probability. Partial dependence plots (PDP) were drawn using the *pdp* R package and show the marginal effect of each predictor on model probabilities [105]. Partial dependence plots allow for comparison of inferred relationships between covariates and suitability probability with *a priori* knowledge of factors that drive snail ecological niche suitability. Variable importance measures were calculated using the *vi_shap* function from the *vip* package [106], which calculates SHapley Additive exPlanations (SHAP) variable importance values (a method of calculating how much covariates contribute to model predictions) [49, 107]. Notably, SHAP values are model agnostic and can estimate comparable values of variable contribution for both regression-based and tree-based methods [49].

Results

Model types produce remarkably different national prediction maps for all species despite using the same presence and background records and environmental data (Fig 2). Although probability prediction varies widely (Fig 2), spatially cross-validated AUC (Fig 3) and TSS (**Fig B in** S1 Text) values of national models do not substantially differ across model types. RF models tended to have somewhat higher sensitivity—they were more likely to accurately predict presence points—than MaxEnt and BRT across species (Table 3 **and Fig B in** S1 Text). RF models also had consistently higher pAUC values—they were more likely to correctly predict presence points relative to the proportion of area predicted as present—across all species and scales (Table 3 **and Fig B in** S1 Text). BRT models tended to have higher specificity—they were more likely to accurately predict background points—than MaxEnt and RF across species (Table 3 **and Fig B in** S1 Text). When comparing de-identified national prediction maps, expert opinion selected BRT maps for *B. glabrata*, *B. straminea*, and *B. tenagophila* as best matching *a priori* knowledge of current suitable snail habitat.

Compared to these national models, model accuracy remained consistent at smaller geographic scales for *B. glabrata* and *B. straminea* and increased at smaller geographic scales for *B. tenagophila* (Fig 3 **and Fig B in** S1 Text), as measured by spatially cross-validated out-of-sample AUC, sensitivity, specificity, and TSS. Spatially cross-validated out-of-sample pAUC values decreased somewhat at smaller geographic scales for all species (**Fig B in** S1 Text). pAUC significance values indicate that almost all spatially cross-validated models were better than random at correctly classifying presence points, with the exception of *B. glabrata* MaxEnt and BRT models (Table 3). However, when testing models fit to national-scale data at predicting

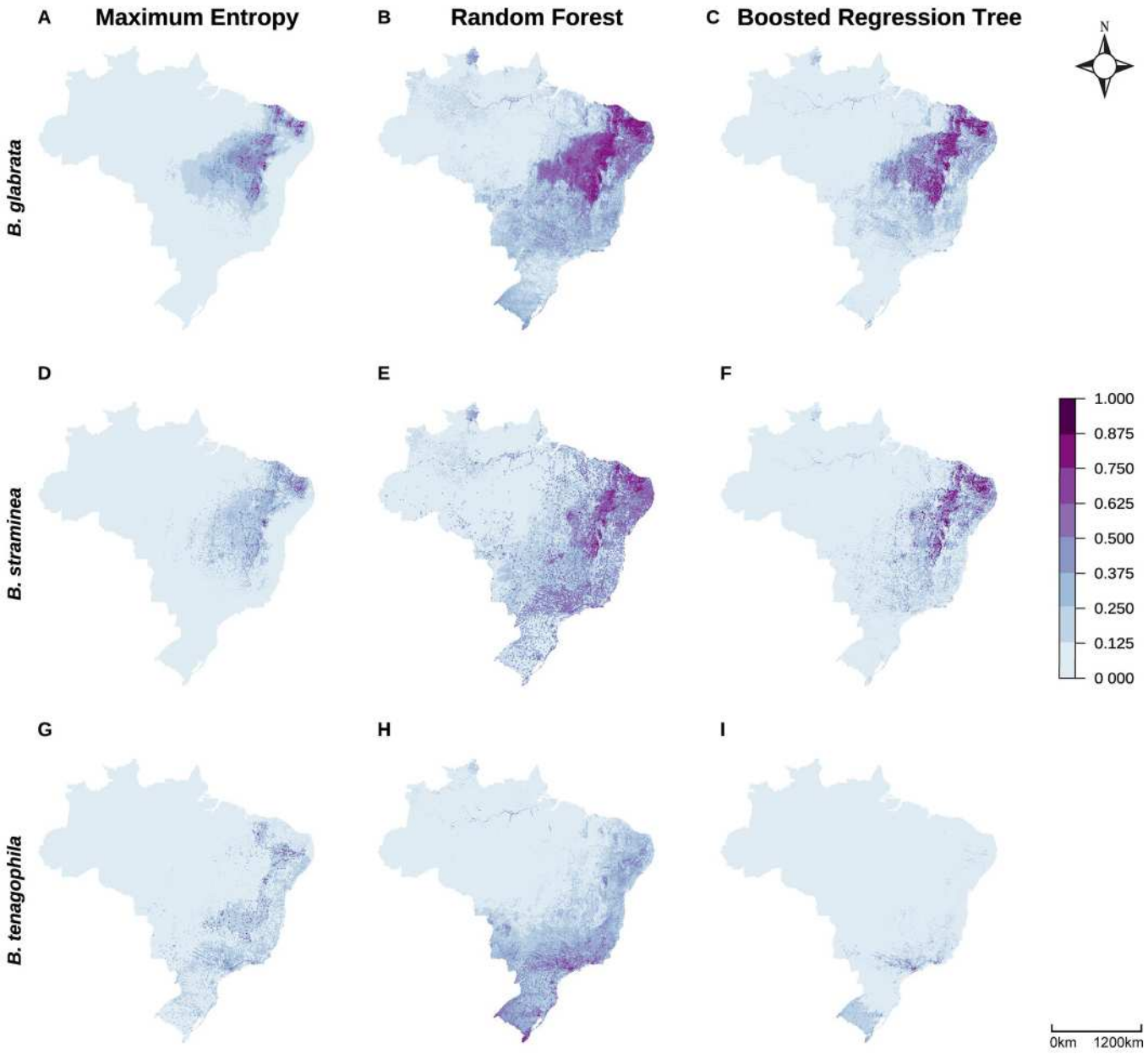

Fig 2. Large variation in snail suitability probabilities at a national scale. National prediction maps of *B. glabrata* (A–C), *B. straminea* (D–F), and *B. tenagophila* (G–I) suitability probabilities by model type (MaxEnt: A, D, G; Random Forest: B, E, H; Boosted Regression Tree: C, F, I). Each pixel shows the mean value across 10 bootstrapping iterations in which models were provided 80% of the available species presence records. Maps were built in R (version 4.2.2) using shapefiles from the *geobr* package [80].

https://doi.org/10.1371/journal.pgph.0002224.g002

state-level presences, all models for all species produced lower in-sample AUC and pAUC values than state models (**Table B in** S1 Text). State models also generally produced higher in-sample sensitivity and specificity values but the nationally-fit models occasionally produced higher sensitivity values (i.e., sometimes the nationally-fit models were able to correctly identify presence points that the state-specific models missed). Differences in predictive accuracy between models trained on state versus national data when tested on state data occurs due to differences in predicted suitability maps, which are visually apparent (**Fig 4 and Figs C and D in** S1 Text). We also directly measure the differences between state suitability maps produced

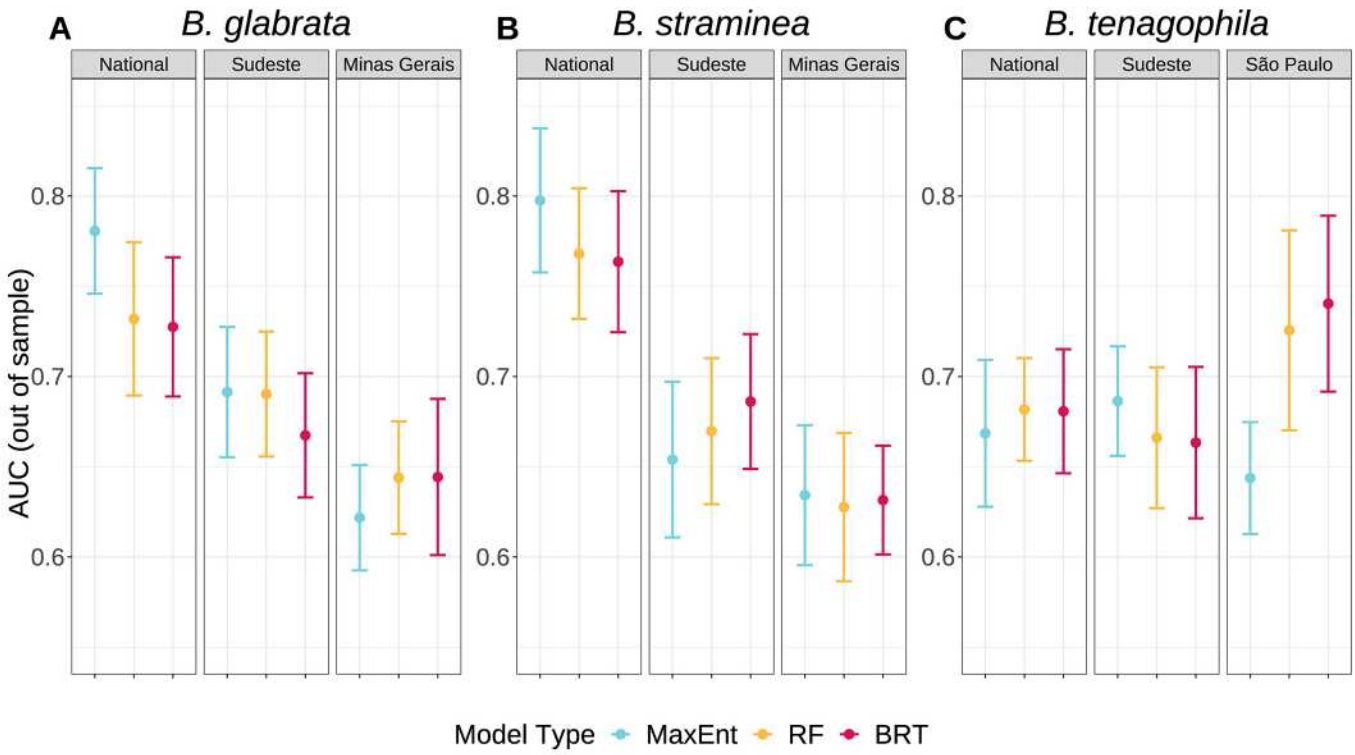

Fig 3. Scale and species drive SDM performance metrics more than model type. Plots of ten-fold spatially cross-validated, out-of-sample AUC values across species (A, B, C), scales (panels), and model types (colors). Plots display mean (point) and +/- standard error (error bars).

https://doi.org/10.1371/journal.pgph.0002224.g003

by state and national models across species and model types by calculating pixel by pixel Pearson correlation coefficients and their significance (Table 4). As evident in the predicted suitability maps, MaxEnt suitability maps remain most similar across scales for all species, while RF and BRT maps change more readily at smaller scales and produce larger improvements in model performance (**Table B in** S1 Text). Suitability maps of *B. tenagophila* were the most correlated for all model types (**Fig D in** S1 Text).

Mean, 95% confidence intervals, and significance values of in-sample pixel by pixel Pearson correlation coefficients between the state suitability maps produced by state and national models across species and machine learning model type. Values are calculated across 10 bootstrapping iterations in which models were provided 80% of the species presence records available at the given scale. The values correspond to comparisons between the two columns of suitability maps in Fig 4 **and Fig C in** S1 Text, and **Fig D in** S1 Text, displaying the comparisons for *B. glabrata*, *B. straminea*, and *B. tenagophila*, respectively. P-values were calculated using a t-test, comparing the 10 bootstrapped values with correlations from 1000 randomly permuted maps.

Despite similar overall accuracy across machine learning model types within geographic extents (i.e., MaxEnt national compared to RF national (Fig 3)), the models infer strikingly different relationships between covariates and suitability probability, which imply distinct biological relationships. We use three specific examples to illustrate how responses differ across model types, spatial extents, and focal species, by comparing plots in each column of Fig 5. First, model types produce different curve shapes: MaxEnt often fits smoother or linear forms in comparison to the flexible, nonlinear shapes produced by RF and BRT, as illustrated by distance to high population density (Fig 5A). Second, functional forms vary across scales: both *B. glabrata* and *B. tenagophila* responses to soil clay percentage are directionally opposite at

Table 3. Model performance of national *Biomphalaria* snail models across machine learning model types.

Species	Model Type	AUC	Sensitivity	Specificity	pAUC	pAUC p-value*	TSS
B. glabrata	MaxEnt	0.78	0.69	0.67	0.52	0.72	0.36
		(0.75–0.82)	(0.61–0.78)	(0.58–0.76)	(0.44–0.60)	(0.00–1.00)	(0.20–0.52)
	RF	0.73	0.62	0.60	0.60	0.00	0.22
		(0.69–0.77)	(0.49–0.76)	(0.54–0.66)	(0.56–0.64)	(0.00–0.96)	(0.05–0.40)
	BRT	0.73	0.64	0.60	0.52	0.46	0.24
		(0.69–0.76)	(0.53–0.75)	(0.53–0.67)	(0.45–0.59)	(0.00–1.00)	(0.08–0.40)
B. straminea	MaxEnt	0.80	0.86	0.72	0.62	0.00	0.58
		(0.76–0.84)	(0.81–0.91)	(0.67–0.77)	(0.55–0.69)	(0.00–1.00)	(0.51–0.65)
	RF	0.77	0.76	0.77	0.64	0.00	0.53
		(0.73–0.80)	(0.70–0.81)	(0.72–0.83)	(0.60–0.67)	(0.00–0.35)	(0.46–0.59)
	BRT	0.76	0.73	0.79	0.61	0.00	0.52
		(0.73–0.80)	(0.68–0.79)	(0.74–0.84)	(0.54–0.67)	(0.00–1.00)	(0.45–0.59)
B. tenagophila	MaxEnt	0.67	0.53	0.47	0.54	0.00	-0.01
		(0.63–0.71)	(0.39–0.66)	(0.36–0.57)	(0.45–0.64)	(0.00–1.00)	(-0.22–0.20)
	RF	0.68	0.53	0.52	0.63	0.00	0.05
		(0.66–0.71)	(0.39–0.66)	(0.42–0.63)	(0.58–0.68)	(0.00–0.00)	(-0.15–0.25)
	BRT	0.68	0.46	0.59	0.59	0.00	0.05
		(0.65–0.72)	(0.35–0.57)	(0.47–0.72)	(0.52–0.66)	(0.00–0.18)	(-0.15–0.25)

Mean and +/- standard errors of spatially cross-validated out-of-sample AUC, sensitivity, specificity, pAUC, and true skill statistic (TSS) values for national models of all species across machine learning model type. *Median and 10%–90% percentiles of pAUC significance reported due to better represent large outlier values. The corresponding table for in-sample estimates for all species can be found in **Table A in** S1 Text.

https://doi.org/10.1371/journal.pgph.0002224.t003

national versus state scales for all model types (Fig 5B). It is important to note that the range of environmental covariates may differ remarkably across geographic extents. Third, species differ in the functional forms: *B. glabrata* and *B. tenagophila* suitability both respond nonlinearly to temperature in the coldest quarter, but with different functional responses that vary between scales (Fig 5C). By contrast, other functional forms remain relatively consistent across species and scale, such as the response to distance to high population density (Fig 5A). These differences in inferred biological relationships highlight the potential pitfalls of using SDMs to extrapolate environmental suitability beyond the scope of the data, and of assuming generality from a single modeling approach.

Given the importance of understanding how *Biomphalaria* snails are responding to land use and land cover (LULC) change, we investigated how the relative importance of LULC variables changes with scale. We hypothesized that LULC variables would become increasingly important compared to climatic gradients at relatively smaller scales. Evidence for this prediction was mixed. Supporting this prediction, the relative importance of LULC variables increased consistently from national to regional to state scales for *B. tenagophila* models (Fig 6C). However, LULC variable importance for *B. glabrata* models (Fig 6A) remained more constant across scales and decreased at the state scale when using a MaxEnt model. Similar to *B. glabrata*, LULC variable importance for *B. straminea* models (Fig 6B) dipped in regional models and was equivalently high in national and state models.

To investigate impacts of using presence data from an expert-executed field collection program versus from a publicly-available species presence database, we constructed models using two distinct datasets: expert-collected and GBIF. As anticipated, each dataset produced distinct predictions of presence probability. Limiting these analyses to *B. tenagophila* in São Paulo, model accuracy was similar across both datasets (expert-collected mean AUC = 0.84,+/-

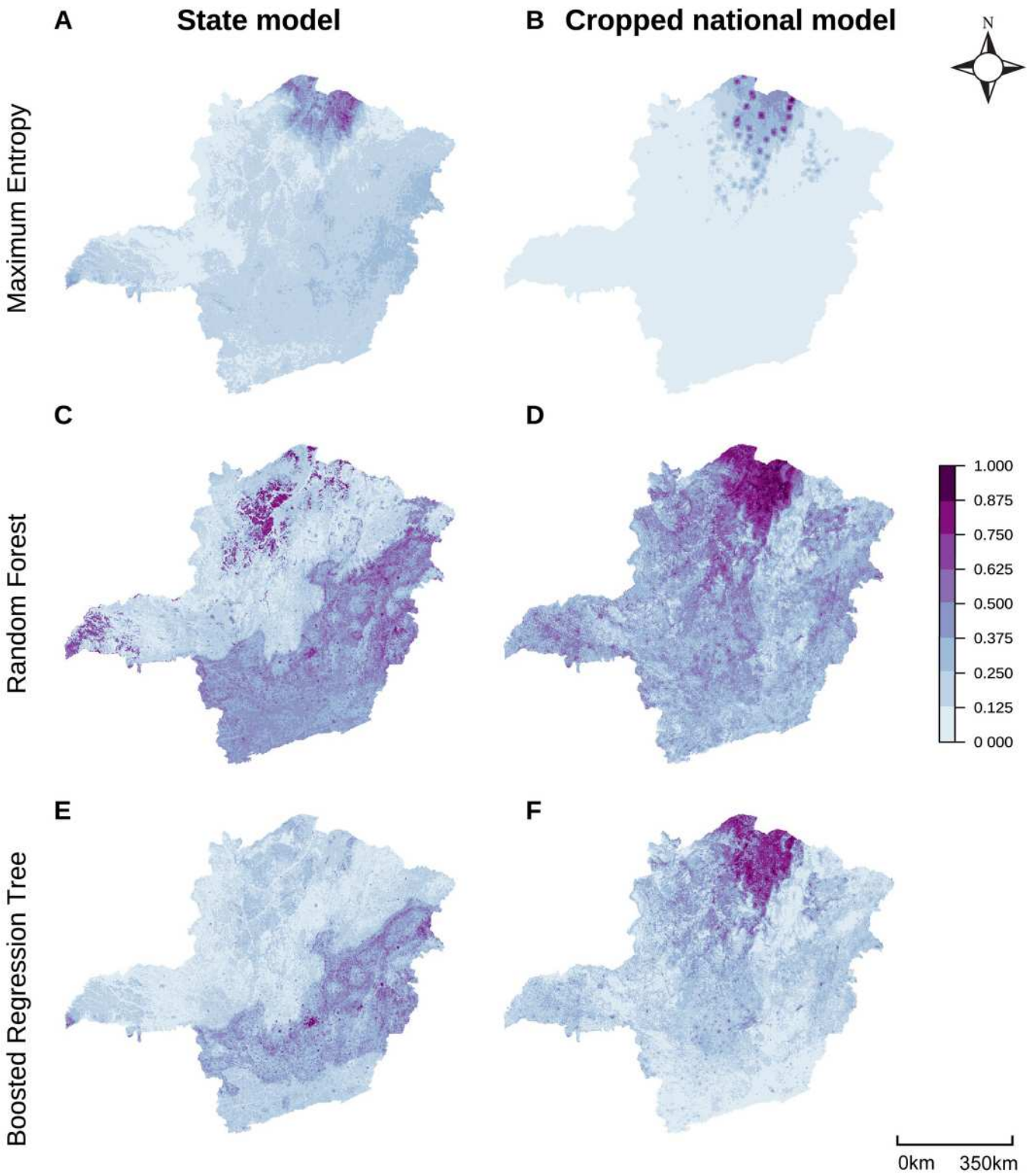

Fig 4. State and national models produce substantially different state-level prediction maps. Minas Gerais prediction maps of *B. glabrata* suitability probabilities by model type (rows) and model geographic extent (columns) Each pixel shows the mean value across 10 bootstrapping iterations in which models were provided 80% of the species presence records available at a given scale. Parallel prediction maps of *B. straminea* in Minas Gerais and *B. tenagophila* in São Paulo can be found in **Figs C and D in** S1 Text. Compared to national models (Fig 2), at smaller geographic scales it becomes more obvious that suitability probabilities can be highly localized, producing points of high suitability surrounded by areas with low suitability. Maps were built in R (version 4.2.2) using shapefiles from the *geobr* package [80].

https://doi.org/10.1371/journal.pgph.0002224.g004

Table 4. Correlation values between state and national models when predicting state-level suitability probabilities across species and model types.

Species	Model Type	Mean correlation (95% CI)	P-value
B. glabrata	MaxEnt	0.48 (0.31–0.65)	0.00
	RF	-0.04 (-0.10–0.02)	0.26
	BRT	-0.02 (-0.09–0.05)	0.63
B. straminea	MaxEnt	0.59 (0.53–0.65)	0.00
	RF	0.27 (0.21–0.33)	0.00
	BRT	0.31 (0.24–0.38)	0.00
B. tenagophila	MaxEnt	0.80 (0.68–0.92)	0.00
	RF	0.69 (0.66–0.73)	0.00
	BRT	0.76 (0.73–0.78)	0.00

https://doi.org/10.1371/journal.pgph.0002224.t004

standard errors: (0.82, 0.86, publicly-available GBIF mean AUC = 0.79, [0.77, 0.82]), yet the prediction maps show substantial variation regardless of model type (Fig 7). Despite somewhat lower AUC values when the two datasets were combined (0.70, [0.68, 0.73]), experts judged the suitability maps as preferable when data from both sources is included, across all model types (**Fig 7C, 7F, and 7I**). Notably, the two data sets have different data quantities, with the expert-collected dataset ($n = 169$) and combined dataset ($n = 234$) containing more presence points than the GBIF dataset ($n = 115$). The same maps and AUC comparisons for models with data quantity held constant ($n = 115$) can be found in the (**Fig E in S1 Text**), with slightly more variation between model suitability maps but generally small changes to above results.

Discussion

SDMs are increasingly used in disease ecology to understand environmental drivers of reservoir host or vector species distributions and to project how they might change with anthropogenic modification. We showed, by systematically comparing SDM approaches that employed different modeling techniques, spatial extents, data types, and species, that both the spatial predictions and the inferred relationships with environmental features can vary substantially across methods, even when performance measures (i.e., AUC, sensitivity, specificity, pAUC, and TSS) are very similar.

A first important result is that even when given the same presence, background, and covariate data, the three model types produce remarkably different suitability maps despite similar accuracy. Although differences in spatially-cross validated mean AUC values were minimal when compared within geographic extents (i.e., MaxEnt national compared to RF national), we found that RF models had higher pAUC and tended to have somewhat higher sensitivity, producing more 'dense' maps of predicted suitable habitat, than MaxEnt or BRT across species and scales (**Table 3 and Fig B in S1 Text**). Consistently higher pAUC values highlight that RF models are the best at predicting presence points correctly, even when controlling for overprediction of presence areas. On the other hand, BRT models tended to have higher specificity, producing more 'sparse' predictions, as compared to MaxEnt and RF across species and scales (**Table 3 and Fig B in S1 Text**).

Our analysis demonstrates the importance of individually investigating sensitivity, specificity, and pAUC (separate from AUC), especially if models are intended to inform disease control policy [108, 109]. If total elimination is of high priority, high sensitivity and/or pAUC—the ability for models to accurately identify all presence locations—might be emphasized to safely capture all presence areas, with less concern for mistakenly implementing control interventions in places that actually contained only background records, which in this case would

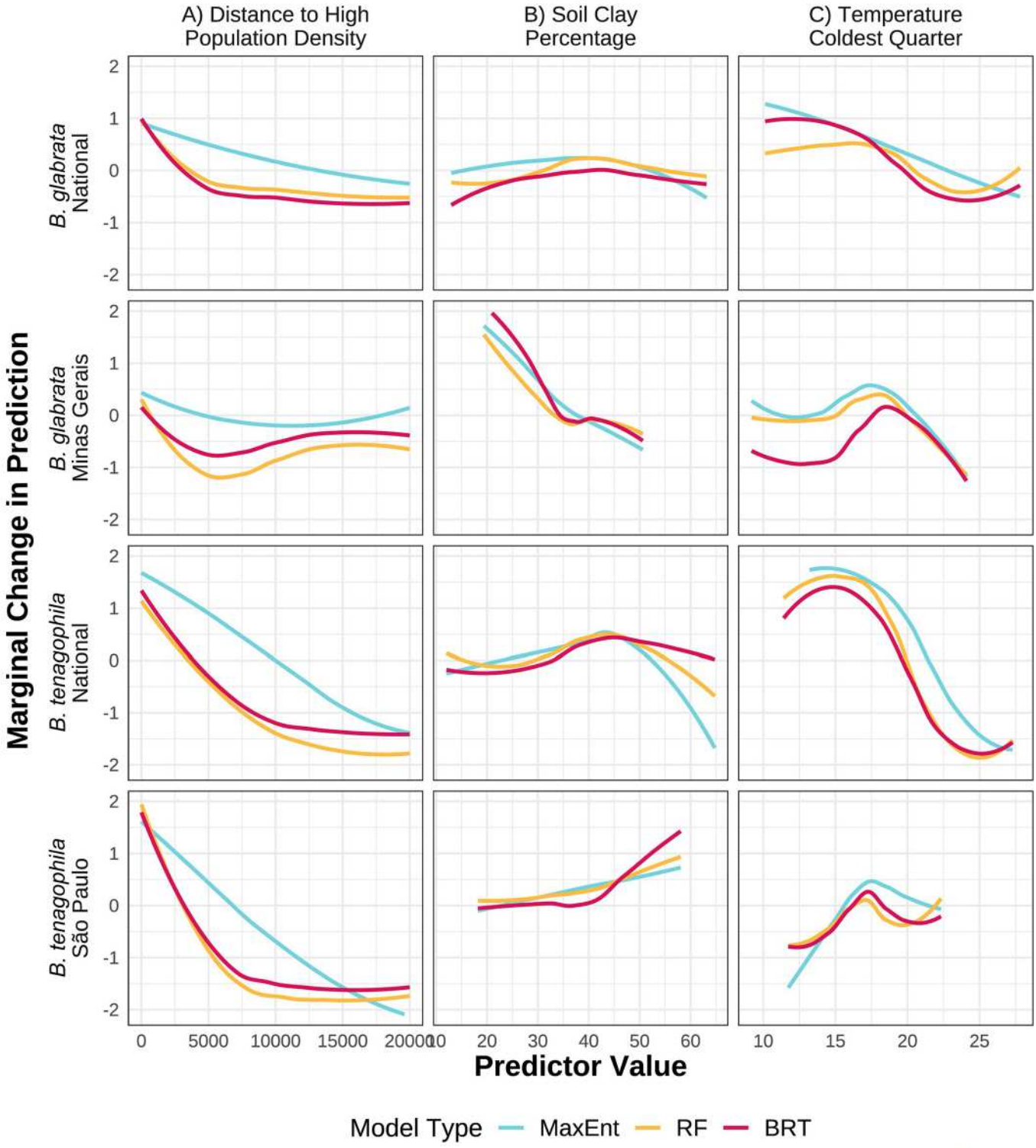

Fig 5. Examples of marginal effects of covariates on suitability probabilities that vary across model type (A), geographic scale (B), and species (C). Partial dependence plots for three covariates (columns) across model types (color), species (top two rows vs. bottom two rows), and scale (first row vs. second and third row vs. fourth).

https://doi.org/10.1371/journal.pgph.0002224.g005

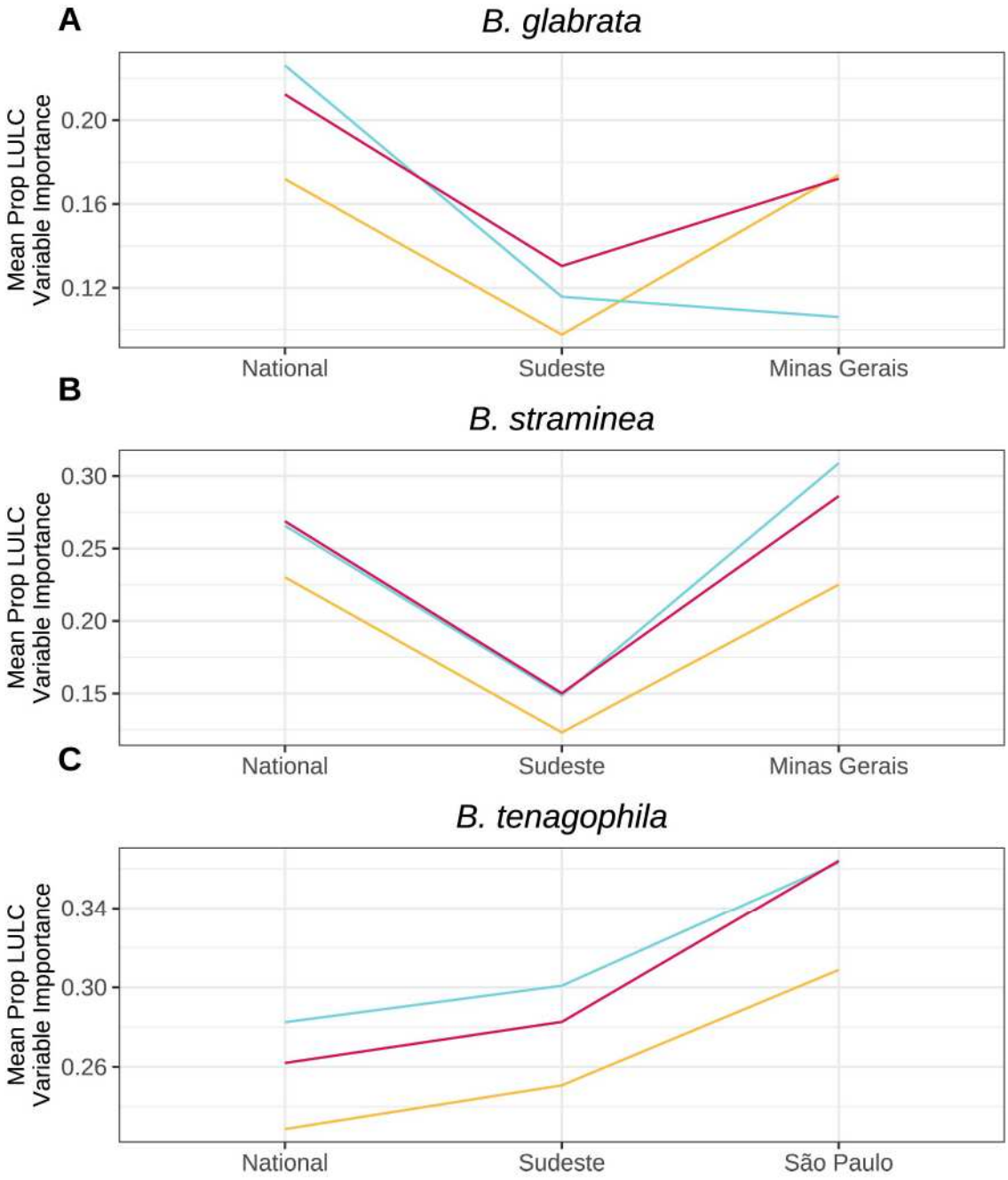

Fig 6. Variable importance of land use/land cover (LULC) variables can increase at smaller scales. Proportion of total variable importance averaged across all training folds attributable to distance to high population density and proportion of temporary crop cover. Displayed for all species (A,B, C) and model types (color).

https://doi.org/10.1371/journal.pgph.0002224.g006

generally suggest using RF models for most species and scales (Table 3 **and Fig B in S1 Text**). Alternatively, with more limited resources, policymakers might prioritize models with high specificity (i.e., the ability to accurately identify locations where the species is not expected), such as the BRT models at all scales for *B. tenagophila* and *B. straminea* (Table 3 **and Fig B in** S1 Text). These models would minimize potential efficiency losses that could result from control programs deploying available resources in places that do not actually contain the species

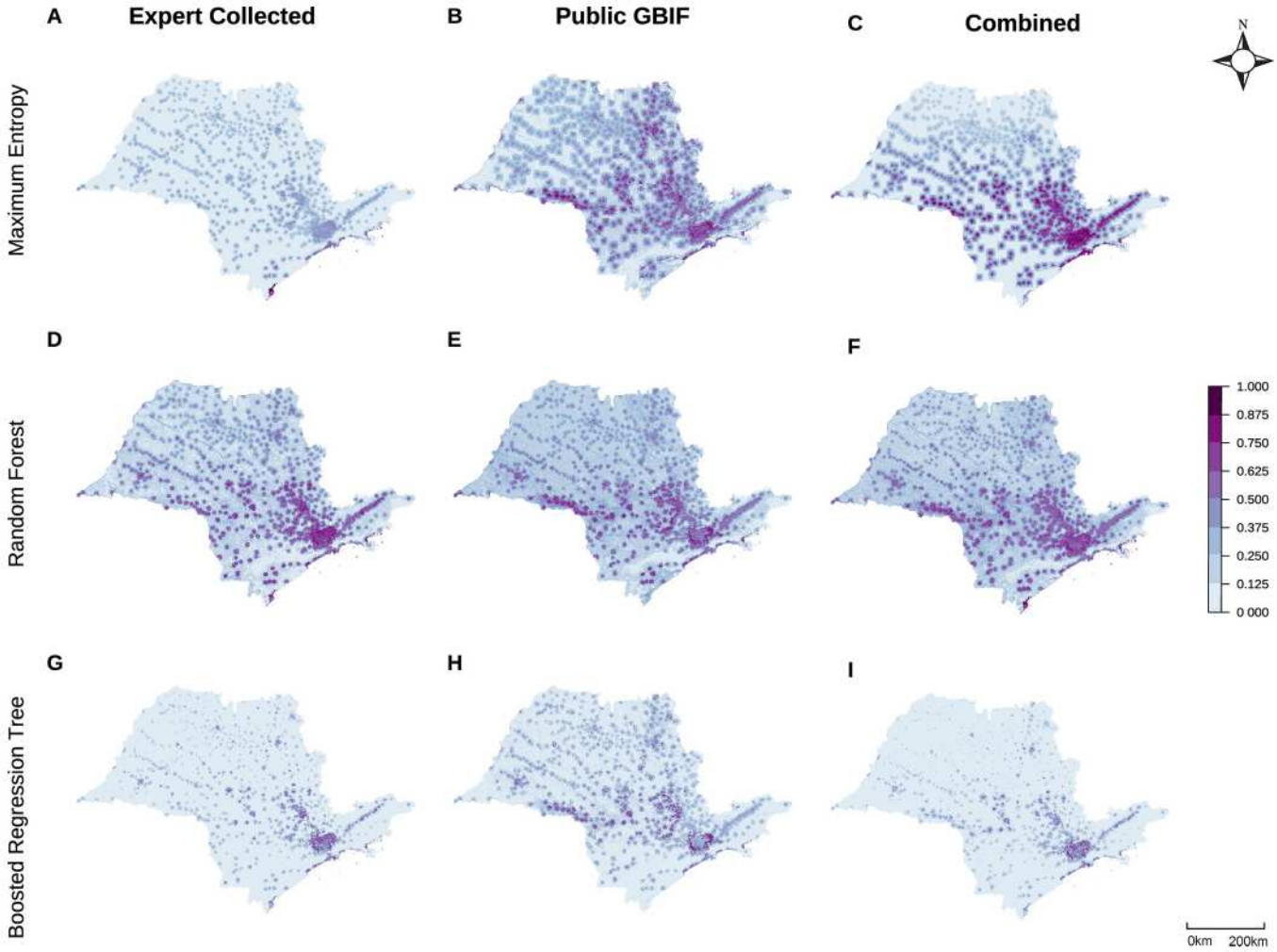

Fig 7. Expert collected and public GBIF data produce visually different suitability maps for *B. tenagophila* in São Paulo across model type. Predicted suitability maps with varying input data (columns) supplied to all model types (rows). Compared to national models (Fig 2), at smaller geographic scales it becomes more obvious that suitability probabilities can be highly localized, producing points of high suitability surrounded by areas with low suitability. Maps were built in R (version 4.2.2) using shapefiles from the *geobr* package [80].

https://doi.org/10.1371/journal.pgph.0002224.g007

of interest. Notably, decisions regarding prevention and intervention efforts will change depending on the species of interest, and our discussion centers around snail control. Conservation efforts, for example, would likely consider different policy decisions based on modeled species distribution maps, as they are concerned with maintaining the presence of species as opposed to the absence. Finally, our experts consistently selected de-identified BRT models as producing maps that best aligned with their *a priori* knowledge of suitable snail habitat across multiple geographic contexts (national and São Paulo scales): these models tended to have higher specificity and lower sensitivity, making their suitability predictions relatively more sparse. Overall, our findings align with previous comparisons of statistical model types in the SDM literature: MaxEnt, RF, and BRT can all produce high model performance measures, although which is the best can vary across species types [18, 41]. Therefore, we encourage modelers to use the suite of SDM resources (including the R packages *dismo*, *enmSdmX*, etc.) to draft multiple models for their application and explicitly test which model type is best suited for their question in close collaboration with experts in the field, particularly those who have extensive expertise in on-the-ground surveillance, as detailed further below.

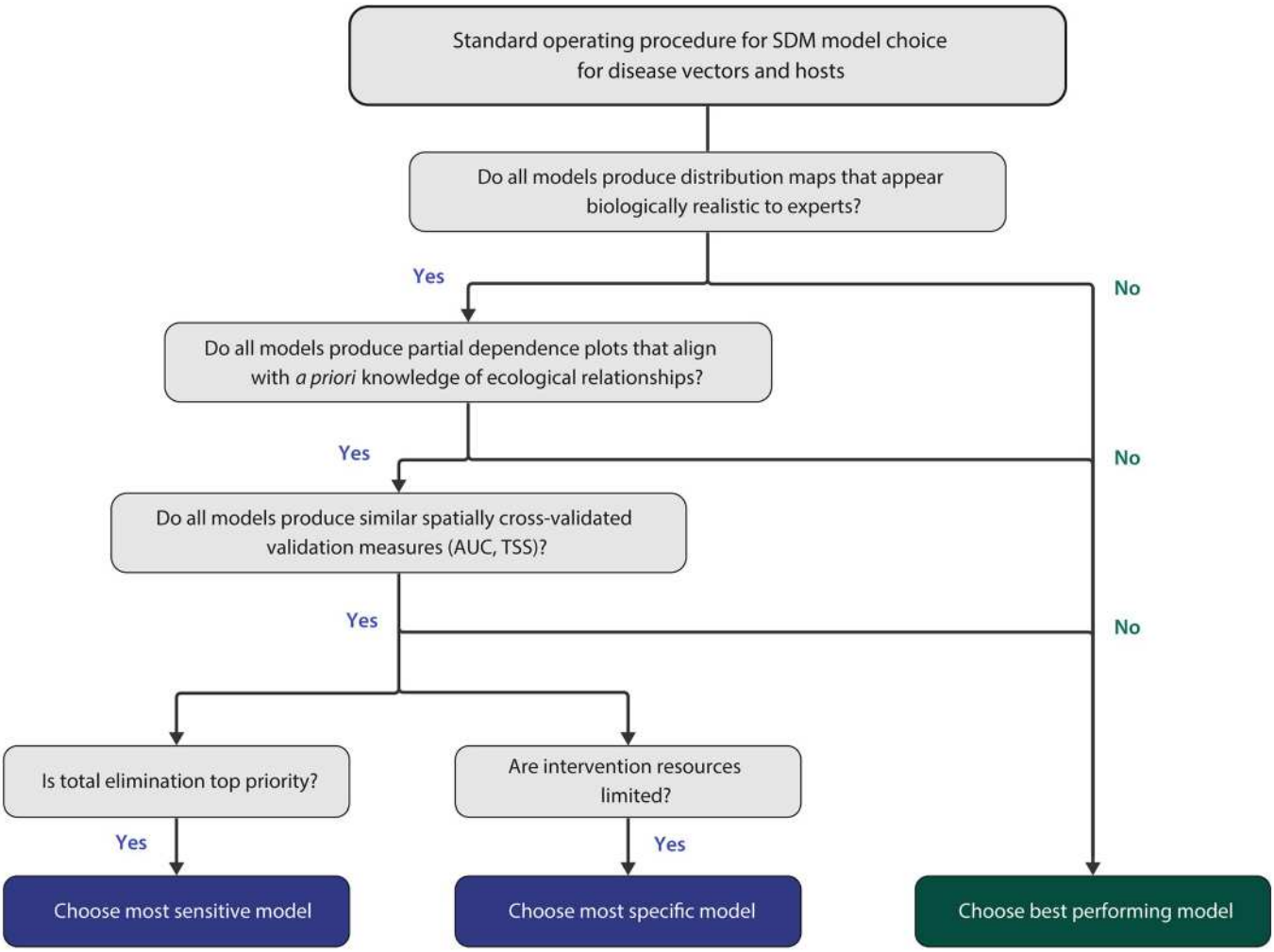

Fig 8.

https://doi.org/10.1371/journal.pgph.0002224.g008

When prediction maps are used to inform intervention and/or funding decisions, significant differences in the suitability maps could warrant radically differing deployment of control strategies [63, 108, 109]. Therefore, in addition to evaluating multiple model performance measures (AUC, sensitivity, specificity, pAUC, TSS, etc.), it is crucial to leverage local ecological knowledge to assess the biological realism of each model's predicted suitability map, as well as of the estimated ecological relationships derived from partial dependence plots [93]. Other analyses have leveraged expert assessment of model outputs when AUC was unable to clearly rank models by performance [51]. This aligns with the well-known but underemployed guideline that remotely-sensed, big data models need to be integrated with local, on-the-ground knowledge to create the best understanding of the system of interest [93]. Fig 8 displays our recommended Standard Operating Procedure summarizing SDM model choice for disease vector and host control efforts (Fig 8).

Subtle differences in performance across scales suggest that the most relevant geographic extent may depend on the application and the relative distribution of data at different geographical scales; yet we also found that model performance could be high from national down to state scales. Comparing across geographic scales, spatially cross-validated AUC values

decreased at smaller geographic scales for *B. glabrata* and *B. straminea*, but increased at smaller geographic scales for *B. tenagophila* (Fig 3; **Fig B in** S1 Text). This phenomenon can likely be attributed to the varying proportion of presence data for each species within each state (Table 2). While 89% of national *B. tenagophila* data is from within the Sudeste region and 65% is within São Paulo state, only 74% of national *B. glabrata* data is from the Sudeste region and 61% is from Minas Gerais. *B. straminea* had an even smaller proportion of total national data at the Sudeste (52%) and state level (39%). Accordingly, we hypothesize that larger amounts of localized data for *B. tenagophila* Sudeste and São Paulo models improved model accuracy, while limiting the ability for national models to capture ecological heterogeneity across the entirety of Brazil. On the other hand, *B. glabrata* and *B. straminea* records are more widely distributed across the nation, allowing for improved national predictions, whereas the smaller data set from Minas Gerais limits the relative performance of state and regional models. When specifically aiming to create best predictions for small geographic regions, we demonstrate that locally-fit SDMs moderately increase model discrimination ability (**Table B in** S1 Text) and create maps with substantially different predictions as compared to nationally-fit models (Table 4 **and** Fig 4 **and Figs C and D in** S1 Text). However, when data are more uniformly distributed at the national scale, national scale models can be cropped to smaller scales relatively effectively, indicating that building national models can also be warranted when needed for large-scale applications or when investigating smaller geographic regions that have limited local data. A final key factor affecting choice of geographic extent is whether the aim is to identify covariate relationships specific to a geographic area of interest or to see generalized covariate relationships that span heterogeneous habitats and geographies, including ranges not yet observed in a given geographic region. This is particularly important when researchers aim to use SDMs to project species distributions under scenarios of future climate change, which include temperature and precipitation patterns not yet experienced in a given region.

Covariate relationships not only varied depending on geographic extent, but also by species and machine learning model used. Even for two snail species in the same genus, their responses to environmental covariates varied in both magnitude and direction (Fig 5), contributing to the large suitability map differences (Fig 2 **and** Fig 7). Compounding these true biological differences among species is the fact that different model structures produce differently shaped partial dependence plot curves, weighting interpretability versus flexibility and differentially favoring nonlinearity and interactions [16, 30, 97]. For example, even when providing our MaxEnt models maximum flexibility in fitting the observed data, the resulting PDP curves still exhibit more limited shapes than RF or BRT. MaxEnt's smooth curves offer simple, interpretable predictor relationships—potentially preferred for modelers whose primary interest is general mechanisms that underlie habitat suitability and/or ease of explanation for policymakers who need to make decisions with limited time [16]. On the other hand, the hyper-flexible curves produced by RF and BRT (and other tree-based methods) can produce improved model performance and variable interactions, especially when models include suites of variables known to interact in nonlinear ways, such as temperature and precipitation or sets of LULC variables [25, 83, 97]. If model classification ability is held at the highest priority and modelers believe it is ecologically feasible for predictors to have flexible relationships, partial dependence plots and the other model evaluation methods discussed here can assist in retaining clear model interpretation [49, 107]. Finally, we note that SDMs are correlative analyses. Therefore, modeled covariate relationships may not be directly related to species presence but with other environmental variables not included in the model. SDMs should be followed by causal analyses if the goal is to understand true causes of species presence.

LULC variables became proportionally more important for predicting *B. tenagophila* snail presence at smaller geographic scales as compared to bioclimatic variables. However, LULC

variable importance remained relatively constant across scales for both *B. glabrata* and *B. straminea*. Given that remotely-sensed bioclimatic variables predominantly change at larger spatial scales (i.e., they are highly spatially-autocorrelated), we expected that models of smaller geographic extent would rely more heavily on LULC variables, which contribute more localized variation (**Fig A in** S1 Text). This hypothesized effect may have been mitigated for two of the three snail species due to the fact that we held spatial resolution ($1km^2$) constant over the three geographic extents. Other analyses varying resolution size have shown that biotic interactions dominate at local scales, while abiotic factors dominate regionally [35]. Holding resolution constant likely allowed even national scale models to leverage localized variation derived from LULC variables.

Given an adequate number of presence points, publicly-available GBIF data creates models with comparable snail distribution predictions and model performance measures as models given an expert-collected dataset. This is a very encouraging finding given that expert-executed field collection programs can be logistically infeasible and public species presence resources have grown in size and popularity [22]. Moreover, even when expert field collection is feasible, it is often not possible to execute surveillance programs across large areas, such as the entirety of Brazil. GBIF cannot always guarantee the same level of species identification accuracy as the morphological and molecular tools often used in expert-executed field sampling, but the accessibility of large amounts of species data has dramatically increased the potential for species distribution analyses [22]. Although only a singular case-study, our findings support the utility of GBIF data for producing accurate SDMs without targeted field collection programs. It is critical to employ methods to overcome spatial biases inherent in these publicly available data sources [51, 52], such as through geographically stratified background sampling and careful inclusion criteria, but our findings support the growing use of these resources [34]. Importantly, many of these conclusions rest on our example where there was sufficient quantity of GBIF data, which was only true of one species in São Paulo state. Our findings demonstrate the value that GBIF data can offer to disease control and elimination efforts and we support ongoing initiatives working to increase access, precision, and quality of GBIF data across all species and geographies [110].

Although our analysis contributes substantially to describing and quantifying current best practices in the SDM literature, there are several limitations. First, the smallest geographic extent we investigated was at the state level, which is still a large area. Other modeling studies, including some of specific *Biomphalaria* species, have been conducted at the municipality or intra-municipality scale [67, 68]. Although infeasible due to data quantity constraints across species for this study, it is possible that our comparisons could have been augmented for local specificity if we had included models built for specific municipalities. Secondly, we included a limited number of predictors in this case study for ease of interpretation, especially given our plan to compare models across geographic extents, machine learning models, and data sources. However, some of our findings could be sensitive to the number, resolution, and/or spatial-autocorrelation of predictors included [111]. For example, a set of predictors dominated by LULC variables—rather than our models that included only two—could come to differing conclusions on changes in variable importance or partial dependence relationships. However, our set of predictors was chosen to be biologically relevant, sufficient to capture ecological relationships, and sufficiently general to be representative for other species distribution modeling studies. A combination of bioclimatic, LULC, and other variables is very common in the body of literature informing this analysis [41]. Lastly, while this analysis does compare results across three species of snails, the species are very similar in that they are all from the same genus and are all freshwater mollusks. Other species, even among those relevant to disease ecology, could vary in their response to our analyses across machine learning models,

spatial extents, and data sources [18]. However, our analysis shows that even species in the same genus may have significantly different ecological niches, indicating that modeling decisions need to be grounded in system-specific ecological and biological knowledge.

There rightfully remains no single gold-standard of SDM methods suitable for all species, geographic locations, and applications because differing contexts and intended uses warrant differing modeling decisions. Making species distribution models that are useful and accurate for a given question of interest requires careful design and in-depth evaluation. This paper aims to serve as a resource and reference for current methods in species distribution modeling, with applications to disease ecology. Given the extent to which these models are used to inform fieldwork, policy, funding, and intervention strategies, continuous assessment and model evaluation are imperative. Species distribution models are powerful tools if used appropriately, and this work illustrates the importance of three key dimensions of variation—model type, spatial extent, and data source—highlighting that the former two can have large implications for model predictions and interpretation.

Supporting information

S1 Text.
(DOCX)

Acknowledgments

We would like to thank the Medical Malacology Collection of Fiocruz Minas for sharing *Biomphalaria* data used in this study.

Author Contributions

Conceptualization: Alyson L. Singleton, Caroline K. Glidden, Andrew J. Chamberlin, Roseli Tuan, Raquel G. S. Palasio, Miguel V. Monteiro, Tejas S. Athni, Erin A. Mordecai, Giulio A. De Leo.

Data curation: Alyson L. Singleton, Caroline K. Glidden, Roseli Tuan, Raquel G. S. Palasio, Adriano Pinter, Roberta L. Caldeira, Cristiane L. F. Mendonça, Omar S. Carvalho.

Formal analysis: Alyson L. Singleton, Erin A. Mordecai.

Funding acquisition: Roseli Tuan, Adriano Pinter, Roberta L. Caldeira, Cristiane L. F. Mendonça, Omar S. Carvalho, Miguel V. Monteiro, Susanne H. Sokolow, Giulio A. De Leo.

Investigation: Alyson L. Singleton.

Methodology: Alyson L. Singleton, Caroline K. Glidden, Miguel V. Monteiro, Tejas S. Athni.

Project administration: Roseli Tuan, Adriano Pinter, Roberta L. Caldeira, Cristiane L. F. Mendonça, Omar S. Carvalho, Miguel V. Monteiro, Susanne H. Sokolow, Giulio A. De Leo.

Resources: Roseli Tuan, Raquel G. S. Palasio, Roberta L. Caldeira, Cristiane L. F. Mendonça, Omar S. Carvalho, Miguel V. Monteiro, Erin A. Mordecai, Giulio A. De Leo.

Supervision: Caroline K. Glidden, Roseli Tuan, Roberta L. Caldeira, Miguel V. Monteiro, Erin A. Mordecai, Giulio A. De Leo.

Validation: Alyson L. Singleton, Caroline K. Glidden, Andrew J. Chamberlin, Roseli Tuan, Raquel G. S. Palasio, Roberta L. Caldeira, Miguel V. Monteiro, Erin A. Mordecai, Giulio A. De Leo.

Visualization: Alyson L. Singleton.

Writing – original draft: Alyson L. Singleton.

Writing – review & editing: Alyson L. Singleton, Caroline K. Glidden, Andrew J. Chamberlin, Roseli Tuan, Raquel G. S. Palasio, Roberta L. Caldeira, Miguel V. Monteiro, Tejas S. Athni, Erin A. Mordecai, Giulio A. De Leo.

References

1. Lippi CA, Mundis SJ, Sippy R, Flenniken JM, Chaudhary A, Hecht G, et al. Trends in mosquito species distribution modeling: insights for vector surveillance and disease control. Parasit Vectors. 2023 Aug 28; 16(1):302. https://doi.org/10.1186/s13071-023-05912-z PMID: 37641089
2. Hollings T, Robinson A, Andel M van, Jewell C, Burgman M. Species distribution models: A comparison of statistical approaches for livestock and disease epidemics. PLOS ONE. 2017 Aug 24; 12(8): e0183626. https://doi.org/10.1371/journal.pone.0183626 PMID: 28837685
3. de Almeida TM, Neto IR, Consalter R, Brum FT, Rojas EAG, da Costa-Ribeiro MCV. Predictive modeling of sand fly distribution incriminated in the transmission of *Leishmania* (Viannia) *braziliensis* and the incidence of Cutaneous Leishmaniasis in the state of Paraná, Brazil. Acta Trop. 2022 May 1; 229:106335.
4. MacDonald AJ, McComb S, Sambado S. Linking Lyme disease ecology and epidemiology: reservoir host identity, not richness, determines tick infection and human disease in California. Environ Res Lett. 2022 Nov; 17(11):114041.
5. de la Vega GJ, Medone P, Ceccarelli S, Rabinovich J, Schilman PE. Geographical distribution, climatic variability and thermo-tolerance of Chagas disease vectors. Ecography. 2015; 38(8):851–60.
6. Ayob N, Burger RP, Belelie MD, Nkosi NC, Havenga H, Necker L de, et al. Modelling the historical distribution of schistosomiasis-transmitting snails in South Africa using ecological niche models. PLOS ONE. 2023 Nov 30; 18(11):e0295149. https://doi.org/10.1371/journal.pone.0295149 PMID: 38033142
7. Guisan A, Zimmermann NE. Predictive habitat distribution models in ecology. Ecol Model. 2000 Dec 5; 135(2):147–86.
8. Jeschke JM, Strayer DL. Usefulness of Bioclimatic Models for Studying Climate Change and Invasive Species. Ann N Y Acad Sci. 2008; 1134(1):1–24. https://doi.org/10.1196/annals.1439.002 PMID: 18566088
9. Elith J, Leathwick JR. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annu Rev Ecol Evol Syst. 2009; 40(1):677–97.
10. Lambin EF, Tran A, Vanwambeke SO, Linard C, Soti V. Pathogenic landscapes: Interactions between land, people, disease vectors, and their animal hosts. Int J Health Geogr. 2010 Oct 27; 9(1):54. https://doi.org/10.1186/1476-072X-9-54 PMID: 20979609
11. Childs ML, Nova N, Colvin J, Mordecai EA. Mosquito and primate ecology predict human risk of yellow fever virus spillover in Brazil. Philos Trans R Soc B Biol Sci. 2019 Aug 12; 374(1782):20180335. https://doi.org/10.1098/rstb.2018.0335 PMID: 31401964
12. Martínez-Bello D, López-Quílez A, Prieto AT. Spatiotemporal modeling of relative risk of dengue disease in Colombia. Stoch Environ Res Risk Assess. 2018 Jun 1; 32(6):1587–601.
13. Gosoniu L, Vounatsou P, Sogoba N, Smith T. Bayesian modelling of geostatistical malaria risk data. Geospatial Health. 2006 Nov 1; 1(1):127–39. https://doi.org/10.4081/gh.2006.287 PMID: 18686238
14. Parviainen M, Luoto M, Ryttäri T, Heikkinen RK. Modelling the occurrence of threatened plant species in taiga landscapes: methodological and ecological perspectives. J Biogeogr. 2008; 35(10):1888–905.
15. Gotelli NJ, Anderson MJ, Arita HT, Chao A, Colwell RK, Connolly SR, et al. Patterns and causes of species richness: a general simulation model for macroecology. Ecol Lett. 2009; 12(9):873–86. https://doi.org/10.1111/j.1461-0248.2009.01353.x PMID: 19702748
16. Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ. A statistical explanation of MaxEnt for ecologists. Divers Distrib. 2011; 17(1):43–57.
17. Barbet-Massin M, Jiguet F, Albert CH, Thuiller W. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol Evol. 2012; 3(2):327–38.
18. Valavi R, Guillera-Arroita G, Lahoz-Monfort JJ, Elith J. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecol Monogr. 2022; 92(1): e01486.

19. Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G. blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods Ecol Evol. 2019; 10(2):225–32.

20. Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens Environ. 2017 Dec 1; 202:18–27.

21. Anderson CB. Biodiversity monitoring, earth observations and the ecology of scale. Ecol Lett. 2018; 21(10):1572–85. https://doi.org/10.1111/ele.13106 PMID: 30004184

22. Lippi CA, Rund SSC, Ryan SJ. Characterizing the Vector Data Ecosystem. J Med Entomol. 2023 Mar 1; 60(2):247–54. https://doi.org/10.1093/jme/tjad009 PMID: 36752771

23. GBIF. GBIF [Internet]. GBIF. [cited 2023 Jul 5]. Available from: https://www.gbif.org/

24. Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. Ecol Model. 2006 Jan 25; 190(3):231–59.

25. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol. 2008; 77 (4):802–13. https://doi.org/10.1111/j.1365-2656.2008.01390.x PMID: 18397250

26. Golding N, Purse BV. Fast and flexible Bayesian species distribution modelling using Gaussian processes. Methods Ecol Evol. 2016; 7(5):598–608.

27. Park YS, Céréghino R, Compin A, Lek S. Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecol Model. 2003 Feb 15; 160(3):265–80.

28. Hao T, Elith J, Guillera-Arroita G, Lahoz-Monfort JJ. A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. Divers Distrib. 2019; 25(5):839–52.

29. Norberg A, Abrego N, Blanchet FG, Adler FR, Anderson BJ, Anttila J, et al. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecol Monogr. 2019; 89(3):e01370.

30. Merow C, Smith MJ, Silander JA Jr. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography. 2013; 36(10):1058–69.

31. Elith J, Graham CH. Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. Ecography. 2009; 32(1):66–77.

32. Fourcade Y, Engler JO, Rödder D, Secondi J. Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias. PLOS ONE. 2014 May 12; 9(5):e97122. https://doi.org/10.1371/journal.pone.0097122 PMID: 24818607

33. Guisande C, García-Roselló E, Heine J, González-Dacosta J, Vilas LG, García Pérez BJ, et al. SPEDInstabR: An algorithm based on a fluctuation index for selecting predictors in species distribution modeling. Ecol Inform. 2017 Jan 1; 37:18–23.

34. Smith AM, Capinha C, Kramer AM. Predicting species distributions with environmental time series data and deep learning [Internet]. bioRxiv; 2022 [cited 2023 Mar 24]. p. 2022.10.26.513922. Available from: https://www.biorxiv.org/content/10.1101/2022.10.26.513922v1

35. Cohen JM, Civitello DJ, Brace AJ, Feichtinger EM, Ortega CN, Richardson JC, et al. Spatial scale modulates the strength of ecological processes driving disease distributions. Proc Natl Acad Sci. 2016 Jun 14; 113(24):E3359–64. https://doi.org/10.1073/pnas.1521657113 PMID: 27247398

36. Bell DM, Schlaepfer DR. On the dangers of model complexity without ecological justification in species distribution modeling. Ecol Model. 2016 Jun 24; 330:50–9.

37. Johnson EE, Escobar LE, Zambrana-Torrelio C. An ecological framework for modeling the geography of disease transmission. Trends Ecol Evol. 2019 Jul 1; 34(7):655–68. https://doi.org/10.1016/j.tree.2019.03.004 PMID: 31078330

38. Allouche O, Tsoar A, Kadmon R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J Appl Ecol. 2006; 43(6):1223–32.

39. Feng X, Park DS, Walker C, Peterson AT, Merow C, Papeş M. A checklist for maximizing reproducibility of ecological niche models. Nat Ecol Evol. 2019 Oct; 3(10):1382–95. https://doi.org/10.1038/s41559-019-0972-5 PMID: 31548646

40. Araújo MB, Anderson RP, Márcia Barbosa A, Beale CM, Dormann CF, Early R, et al. Standards for distribution models in biodiversity assessments. Sci Adv. 2019 Jan 16; 5(1):eaat4858. https://doi.org/10.1126/sciadv.aat4858 PMID: 30746437

41. Elith* J, H. Graham* C, P. Anderson R, Dudík M, Ferrier S, Guisan A, et al. Novel methods improve prediction of species' distributions from occurrence data. Ecography. 2006; 29(2):129–51.

42. Jiménez-Valverde A, Lobo JM, Hortal J. Not as good as they seem: the importance of concepts in species distribution modelling. Divers Distrib. 2008; 14(6):885–90.

43. Boria RA, Olson LE, Goodman SM, Anderson RP. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecol Model. 2014 Mar 10; 275:73–7.

44. Brun P, Thuiller W, Chauvier Y, Pellissier L, Wüest RO, Wang Z, et al. Model complexity affects species distribution projections under climate change. J Biogeogr. 2020; 47(1):130–42.

45. Fielding AH, Bell JF. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv. 1997 Mar; 24(1):38–49.

46. Peterson AT, Papeş M, Soberón J. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecol Model. 2008 Apr 24; 213(1):63–72.

47. Galante PJ, Alade B, Muscarella R, Jansa SA, Goodman SM, Anderson RP. The challenge of modeling niches and distributions for data-poor species: a comprehensive approach to model complexity. Ecography. 2018; 41(5):726–36.

48. Jiménez-Valverde A, Acevedo P, Barbosa AM, Lobo JM, Real R. Discrimination capacity in species distribution models depends on the representativeness of the environmental domain. Glob Ecol Biogeogr. 2013; 22(4):508–16.

49. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2017 [cited 2023 Mar 26]. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

50. Greenwell BM, Boehmke BC, McCarthy AJ. A simple and effective model-based variable importance measure [Internet]. arXiv; 2018 [cited 2023 Mar 26]. Available from: http://arxiv.org/abs/1805.04755

51. Beck J, Böller M, Erhardt A, Schwanghart W. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecol Inform. 2014 Jan 1; 19:10–5.

52. Daru B, Rodriguez J. Specimens trump field observations in capturing biodiversity trends. Nat Ecol Evol. 2023 Jun; 7(6):802–3. https://doi.org/10.1038/s41559-023-02051-7 PMID: 37127768

53. Morgan J a. T, Dejong RJ, Snyder SD, Mkoji GM, Loker ES. *Schistosoma mansoni* and *Biomphalaria*: past history and future trends. Parasitology. 2001 Nov; 123(7):211–28.

54. Mitchell C. PAHO/WHO | Schistosomiasis [Internet]. Pan American Health Organization / World Health Organization. 2014 [cited 2023 May 26]. Available from: https://www3.paho.org/hq/index.php?option=com_content&view=article&id=9474:schistosomiasis-factsheet&Itemid=0&lang=en#gsc.tab=0

55. Kloos H, Correa-Oliveira R, Oliveira Quites HF, Caetano Souza MC, Gazzinelli A. Socioeconomic studies of schistosomiasis in Brazil: A review. Acta Trop. 2008 Nov 1; 108(2):194–201. https://doi.org/10.1016/j.actatropica.2008.07.002 PMID: 18694715

56. Silva da Paz W, Duthie MS, Ribeiro de Jesus A, Machado de Araújo KCG, Dantas dos Santos A, Bezerra-Santos M. Population-based, spatiotemporal modeling of social risk factors and mortality from schistosomiasis in Brazil between 1999 and 2018. Acta Trop. 2021 Jun 1; 218:105897. https://doi.org/10.1016/j.actatropica.2021.105897 PMID: 33753030

57. Santos IG de A, Bezerra LP, Cirilo TM, Silva LO, Machado JPV, Lima PD, et al. New epidemiological profile of schistosomiasis from an area of low prevalence in Brazil. Rev Soc Bras Med Trop. 2020 Oct 21; 53:e20200335. https://doi.org/10.1590/0037-8682-0335-2020 PMID: 33111913

58. Klohe K, Koudou BG, Fenwick A, Fleming F, Garba A, Gouvras A, et al. A systematic literature review of schistosomiasis in urban and peri-urban settings. PLoS Negl Trop Dis. 2021 Feb 25; 15(2):e0008995. https://doi.org/10.1371/journal.pntd.0008995 PMID: 33630833

59. McCreesh N, Booth M. The effect of simulating different intermediate host snail species on the link between water temperature and schistosomiasis risk. PloS One. 2014; 9(7):e87892. https://doi.org/10.1371/journal.pone.0087892 PMID: 24988377

60. Kloos H, Souza C de, Gazzinelli A, Soares Filho BS, Temba P da C, Bethony J, et al. The distribution of *Biomphalaria* spp. in different habitats in relation to physical, biological, water contact and cognitive factors in a rural area in Minas Gerais, Brazil. Mem Inst Oswaldo Cruz. 2001 Sep; 96:57–66.

61. Gomes E, Leal-Neto OB, Albuquerque J, Silva H da, Barbosa CS. Schistosomiasis transmission and environmental change: a spatio-temporal analysis in Porto de Galinhas, Pernambuco—Brazil. Int J Health Geogr. 2012 Nov 20; 11(1):51.

62. Rowel C, Fred B, Betson M, Sousa-Figueiredo JC, Kabatereine NB, Stothard JR. Environmental epidemiology of intestinal schistosomiasis in Uganda: population dynamics of *biomphalaria* (gastropoda: planorbidae) in Lake Albert and Lake Victoria with observations on natural infections with digenetic trematodes. BioMed Res Int. 2015; 2015:717261.

63. Soares Magalhães RJ, Salamat MS, Leonardo L, Gray DJ, Carabin H, Halton K, et al. Geographical distribution of human *Schistosoma japonicum* infection in The Philippines: tools to support disease control and further elimination. Int J Parasitol. 2014 Nov 1; 44(13):977–84.

64. Scholte RGC, Carvalho OS, Malone JB, Utzinger J, Vounatsou P. Spatial distribution of *Biomphalaria* spp., the intermediate host snails of *Schistosoma mansoni*, in Brazil. Geospatial Health. 2012 Sep 1; 6 (3):S95–101.

65. Scholte RGC, Gosoniu L, Malone JB, Chammartin F, Utzinger J, Vounatsou P. Predictive risk mapping of schistosomiasis in Brazil using Bayesian geostatistical models. Acta Trop. 2014 Apr 1; 132:57–63. https://doi.org/10.1016/j.actatropica.2013.12.007 PMID: 24361640

66. Rumi A, Vogler RE, Beltramino AA. The South-American distribution and southernmost record of Biomphalaria peregrina—a potential intermediate host of schistosomiasis. PeerJ. 2017 May 30; 5: e3401. https://doi.org/10.7717/peerj.3401 PMID: 28584726

67. Barbosa VS, Guimarães RJ de PS e, Loyo RM, Barbosa CS. Modelling of the distribution of *Biomphalaria glabrata* and *Biomphalaria straminea* in the metropolitan region of Recife, Pernambuco, Brazil. Geospatial Health [Internet]. 2016 Nov 25 [cited 2023 Mar 24]; 11(3). Available from: https://geospatialhealth.net/index.php/gh/article/view/490

68. Palasio RGS, de Azevedo TS, Tuan R, Chiaravalloti-Neto F. Modelling the present and future distribution of *Biomphalaria* species along the watershed of the Middle Paranapanema region, São Paulo, Brazil. Acta Trop. 2021 Feb 1; 214:105764.

69. Guimarães RJPS, Freitas CC, Dutra LV, Felgueiras CA, Moura ACM, Amaral RS, et al. Spatial distribution of *Biomphalaria* mollusks at São Francisco River Basin, Minas Gerais, Brazil, using geostatistical procedures. Acta Trop. 2009 Mar 1; 109(3):181–6.

70. Guimarães RJ de PS, Freitas CC, Dutra LV, Scholte RGC, Martins-Bedé FT, Fonseca FR, et al. A geoprocessing approach for studying and controlling schistosomiasis in the state of Minas Gerais, Brazil. Mem Inst Oswaldo Cruz. 2010 Jul; 105:524–31. https://doi.org/10.1590/s0074-02762010000400030 PMID: 20721503

71. Nascimento GL, Pegado HM, Domingues ALC, Ximenes RA de A, Itria A, Cruz LN, et al. The cost of a disease targeted for elimination in Brazil: the case of schistosomiasis mansoni. Mem Inst Oswaldo Cruz. 2019 Jan 14; 114:e180347. https://doi.org/10.1590/0074-02760180347 PMID: 30652735

72. Tuan R, Pires F, Sanches Palasio RG, Dalla R, Almeida Guimaraes MCD. Pattern of Genetic Divergence of Mitochondrial DNA Sequences in *Biomphalaria tenagophila* Complex Species Based on Barcode and Morphological Analysis. In: Rokni MB, editor. Schistosomiasis [Internet]. InTech; 2012 [cited 2023 Apr 17]. Available from: http://www.intechopen.com/books/schistosomiasis/pattern-of-genetic-divergence-of-mitochondrial-dna-sequences-in-biomphalaria-tenagophila-complex-spe

73. Oliveira-Júnior JF de, Correia Filho WLF, Monteiro L da S, Shah M, Hafeez A, Gois G de, et al. Urban rainfall in the Capitals of Brazil: Variability, trend, and wavelet analysis. Atmospheric Res. 2022 Apr 1; 267:105984.

74. Ohlweiler FP, Eduardo JM, Takahashi FY, Holcman MM, Costa CBTL da. Gastrópodes dulciaquícolas e helmintos associados, em coleções hídricas de Santo André, São Paulo, Brasil. Rev Biociências [Internet]. 2012 Nov 1 [cited 2023 Apr 17]; 18(1). Available from: http://revistas.unitau.br/ojs/index.php/biociencias/article/view/1497

75. Palasio RGS, Casotti MO, Rodrigues TC, Menezes RMT, Zanotti-Magalhaes EM, Tuan R. The current distribution pattern of *Biomphalaria tenagophila* and *Biomphalaria straminea* in the northern and southern regions of the coastal fluvial plain in the state of São Paulo. Biota Neotropica. 2015 Jul 31; 15:e20140153.

76. Palasio RGS, Guimarães MC de A, Ohlweiler FP, Tuan R. Molecular and morphological identification of *Biomphalaria* species from the state of São Paulo, Brazil. ZooKeys. 2017 Apr 12;(668):11–32.

77. Palasio RGS, Zanotti-Magalhães EM, Tuan R. Genetic diversity of the freshwater snail *Biomphalaria tenagophila* (d'Orbigny, 1835) (Gastropoda: Hygrophila: Planorbidae) across two coastal areas of southeast Brazil. Folia Malacol. 2018 Dec 4; 26(4):221–9.

78. Palasio RGS, Xavier IG, Chiaravalotti-Neto F, Tuan R. Diversity of *Biomphalaria* spp. freshwater snails and associated mollusks in areas with schistosomiasis risk, using molecular and spatial analysis tools. Biota Neotropica. 2019 Aug 15; 19:e20190746.

79. Palasio RGS, de Jesus Rossignoli T, Di Sessa RCS, Ohlweiler FP, Chiaravalloti-Neto F. Spatial analysis of areas at risk for schistosomiasis in the Alto Tietê Basin, São Paulo, Brazil. Acta Trop. 2021 Dec; 224:106132.

80. Pereira RHM, Gonçalves CN, et al [Internet]. 2019 [cited 2024 May 10]. geobr: Loads Shapefiles of Official Spatial Data Sets of Brazil. 2019. Available from: https://github.com/ipeaGIT/geobr

81. IUCN. IUCN [Internet]. [cited 2023 Jul 5]. Available from: https://www.iucn.org/

82. He H, Garcia EA. Learning from Imbalanced Data. IEEE Trans Knowl Data Eng. 2009 Sep; 21 (9):1263–84.

83. Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G. Modelling species presence-only data with random forests. Ecography. 2021; 44(12):1731–42.

84. Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, et al. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol Appl. 2009; 19(1):181–97. https://doi.org/10.1890/07-2153.1 PMID: 19323182

85. CHELSA. CHELSA [Internet]. Chelsa Climate. 2020 [cited 2023 May 26]. Available from: https://chelsa-climate.org/downloads/

86. Yamazaki D, Ikeshima D, Sosa J, Bates PD, Allen GH, Pavelsky TM. MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset. Water Resour Res. 2019; 55 (6):5053–73.

87. Hengl T, Gupta S. Soil water content (volumetric %) for 33kPa and 1500kPa suctions predicted at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution [Internet]. Zenodo; 2019 [cited 2023 Jul 5]. Available from: https://zenodo.org/record/2784001

88. Hengl T. Soil pH in H2O at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution [Internet]. Zenodo; 2018 [cited 2023 Jul 5]. Available from: https://zenodo.org/record/2525664

89. Hengl T. Clay content in % (kg / kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution [Internet]. Zenodo; 2018 [cited 2023 Jul 5]. Available from: https://zenodo.org/record/2525663

90. Open Spatial Demographic Data and Research. WorldPop [Internet]. WorldPop. [cited 2023 Jul 5]. Available from: https://www.worldpop.org/

91. Souza CM, Z. Shimbo J, Rosa MR, Parente LL, A. Alencar A, Rudorff BFT, et al. Reconstructing three decades of land use and land cover changes in Brazilian biomes with landsat archive and Earth Engine. Remote Sens. 2020 Jan; 12(17):2735.

92. Dijkstra L, Florczyk AJ, Freire S, Kemper T, Melchiorri M, Pesaresi M, et al. Applying the Degree of Urbanisation to the globe: A new harmonised definition reveals a different picture of global urbanisation. J Urban Econ. 2021 Sep 1; 125:103312.

93. Chaves LF, Gottdenker NL, Runk JV, Bergmann LR. Reifications in disease ecology 2: Towards a decolonized pedagogy enabling science by, and for, the people. Capital Nat Social. 2023 Jan 3; 0 (0):1–18.

94. Guisan A, Thuiller W, Zimmermann NE. Habitat suitability and distribution models: with applications in R [Internet]. Cambridge: Cambridge University Press; 2017 [cited 2023 Mar 25]. (Ecology, Biodiversity and Conservation). Available from: https://www.cambridge.org/core/books/habitat-suitability-and-distribution-models/A17F74A3418DBF9ADA191A04C35187F9

95. Steen VA, Tingley MW, Paton PWC, Elphick CS. Spatial thinning and class balancing: Key choices lead to variation in the performance of species distribution models with citizen science data. Methods Ecol Evol. 2021; 12(2):216–26.

96. Smith AB, Murphy SJ, Henderson D, Erickson KD. Including imprecisely georeferenced specimens improves accuracy of species distribution models and estimates of niche breadth. Glob Ecol Biogeogr. 2023; 32(3):342–55.

97. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods. 2009 Dec; 14(4):323–48.

98. Hastie T, Friedman J, Tibshirani R. The Elements of Statistical Learning [Internet]. New York, NY: Springer; 2001 [cited 2023 Mar 27]. (Springer Series in Statistics). Available from: http://link.springer.com/10.1007/978-0-387-21606-5

99. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002; 2.

100. Hijmans R, Phillips S, Leathwick J, Elith J. Package "dismo." Circles. 2017;1–68.

101. Muscarella R, Galante PJ, Soley-Guardia M, Boria RA, Kass JM, Uriarte M, et al. ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. Methods Ecol Evol. 2014; 5(11):1198–205.

102. Brenning A. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In: 2012 IEEE International Geoscience and Remote Sensing Symposium [Internet]. 2012 [cited 2024 Feb 9]. p. 5372–5. Available from: https://ieeexplore.ieee.org/abstract/document/6352393?casa_token=9cLUp0sh44sAAAAA:Aj5KmRmaAN2Z92tMkLuyIRMxj7VEy6DziMlv6JSrjy97JJyyGh-QGhdZxn_hfZc2HVkFmwM5ng

103. Jiménez-Valverde A, Lobo JM, Hortal J. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. Community Ecol. 2009 Dec 1; 10(2):196–205.

104. Osorio-Olvera L. luismurao/ntbox: From getting biodiversity data to evaluating species distribution models in a friendly GUI environment version 0.7.1 from GitHub [Internet]. [cited 2024 Feb 26]. Available from: https://rdrr.io/github/luismurao/ntbox/

105. Greenwell BM. pdp: An R package for constructing partial dependence plots. R J. 2017;

106. Greenwell B, Boehmke B, Gray B. Package "vip." Var Importance Plots. 12(1):343–66.

107. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst. 2014 Dec 1; 41(3):647–65.

108. Rhodes CG, Loaiza JR, Romero LM, Gutiérrez Alvarado JM, Delgado G, Rojas Salas O, et al. *Anopheles albimanus* (Diptera: Culicidae) Ensemble Distribution Modeling: Applications for Malaria Elimination. Insects. 2022 Mar; 13(3):221.

109. Ren Z, Wang D, Ma A, Hwang J, Bennett A, Sturrock HJW, et al. Predicting malaria vector distribution under climate change scenarios in China: Challenges for malaria elimination. Sci Rep. 2016 Feb 12; 6 (1):20604.

110. Anderson RP, Araújo M, Guisan A, Lobo JM, Martínez-Meyer E. Final report of the task group of GBIF data fitness for use in distribution modelling. 2016;

111. Merow C, Smith MJ, Edwards TC, Jr, Guisan A, McMahon SM, Normand S, et al. What do we gain from simplicity versus complexity in species distribution models? Ecography. 2014; 37(12):1267–81.