Climate variables and urbanization drive changes in the distribution of *S. mansoni* competent snails in Brazil: novel results from a counterfactual analysis

Caroline Kate Glidden[1*], Alyson L. Singleton[2], Andrew Chamberlin[3], Roseli Tuan[4], Raquel Palasio[4], Roberta Caldeira[5], Antonio Miguel V Monteiro[6], Kamazima M. Lwiza[7], Ping Liu[8], Vivian Silva[6], Tejas S. Athni[9], Susanne H. Sokolow[10], Erin A. Mordecai[11] and Giulio A. De Leo[12]

(1)Stanford University, Department of Biology, Institute for Human-Centered AI, Stanford, United States, (2)Stanford University, Emmett Interdisciplinary Program in Environment and Resources, Stanford, United States, (3) Stanford University, Department of Oceans, Hopkins Marine Station, Pacific Grove, United States, (4)Pasteur Institute, São Paulo, Brazil, (5) Fiocruz Minas/Belo Horizonte, Belo Horizonte, Brazil, (6) National Institute for Space Research, São José dos Campos, Brazil, (7) Stony Brook University, Stony Brook, NY, United States, (8) Stony Brook University, Stony Brook, United States, (9) Harvard University, Cambridge, United States, (10) Stanford University, Woods Institute for the Environment, Stanford, United States, (11) Stanford University, Department of Biology, Stanford, CA, United States, (12) Stanford University, Hopkins Marine Station, Pacific Grove, CA, United States

***corresponding author: cglidden@stanford.edu**

ABSTRACT

Schistosomiasis is a debilitating neglected tropical disease caused by *Schistosoma* parasites. In the Americas, Brazil bears the largest burden of this disease with 2-6 million people currently infected. *Schistosoma* flukes are obligate parasites of freshwater *Biomphalaria* snails, thus controlling snail populations is critical to reducing schistosomiasis transmission risk. As snails are sensitive to a wide range of environmental and climatic drivers, we expect that their distribution has been, and will continue to be, significantly impacted by global environmental change. Here, we leveraged cutting edge machine learning models, remote sensing data, and a 30-year dataset on snail occurrence to map the historical and current habitat suitability of competent *Biomphalaria* intermediate hosts. We asked (i) What are the key features of the environment that influence the distribution of suitable snail habitat? (ii) How has snail distribution shifted with climate and urbanization throughout the last three decades? Our models show that variability in climate across time, particularly for variables related to precipitation, has likely influenced broad shifts in snail host range, whereas increase in urban areas has influenced fine scale, increases in *Biomphalaria* habitat suitability. Elucidating change in snail distribution through time – while accounting for non-linearities that are difficult to detect from case studies – can help to inform if and when schistosomiasis control strategies should be adapted.

INTRODUCTION

Schistosomiasis is a globally distributed, debilitating, and sometimes fatal, disease of poverty caused by *Schistosoma* blood flukes – parasitic flatworms (class: trematoda) of freshwater snails and vertebrates [1,2]. Schistosomiasis currently affects over 250 million people in tropical and subtropical regions of the Americas, Africa, and Asia [1,2]. With more than 800 million people living in regions at risk for schistosomiasis transmission, schistosomiasis is one of the most important neglected tropical diseases, second only to malaria for disease burden. In the Americas, Brazil shoulders the largest burden of this disease with 2-6 million people currently infected [3]. The World Health Organization (WHO) includes schistosomiasis in the new neglected tropical disease roadmap for elimination and control by 2030[4]. However, changes in the distribution of the disease have made areas in need of public health interventions a moving target[5], thus investment in adaptive control efforts is increasingly needed to make progress towards this goal.

A prerequisite for schistosomiasis transmission is the presence of the obligate intermediate host of the parasite, i.e., freshwater snails of the genus *Biomphalaria* - hosts for *Schistosoma mansoni*, the parasite causing intestinal schistosomiasis - and *Bulinus* spp., hosts for *S. haematobium*, the paraste responsibile of uro-genital schistosomiasis. Therefore, the distribution and abundance of the snail host underlie the observed spatiotemporal variation in schistosomiasis cases where focal transmission occurs. Understanding the spatial distribution of the intermediate host snails and their environmental, ecological and socioeconomic determinants is thus a public health priority. In Brazil, an endemic country for intestinal schistosomiasis, *S. mansoni* is transmitted via three species of *Biomphalaria* (Preston, 1910) snails: *Biomphalaria glabrata* (Say, 1818)*, B. straminea* (Dunker, 1848)*,* and *B. tenagophila* (D'Orbigny, 1835) [6]*.* As ectotherms, temperature strongly influences fundamental snail biological processes, including growth rates and reproductive rates; as freshwater animals, precipitation can influence habitat availability and phenology, all of which ultimately impact snail fecundity and survival [7,8]. Further, local case studies have demonstrated that *Biomphalaria* snails can thrive in human dominated environments within urban and peri-urban areas, such as drainage ditches and irrigation systems for small-scale agriculture, which are often found in areas with marginalized populations and informal settlements with limited access to clean water, sanitation, and waste-water treatment[9]. Thus, multiple environmental features impact snail habitat suitability and, consequently, human exposure to schistosomiasis. With climate change and land-use change on the rise, we expect that the extent of snail habitat suitability has shifted through time and will continue to shift with continued global change – continuing to move the target for public health interventions.

In Brazil, as well as globally, schistosomiasis control has been orchestrated through preventative and reactive medical treatment via administration of the drug praziquantel [2], WASH interventions (access to safe water, sanitation, and hygiene)[10], and

environmental interventions (e.g., water engineering and snail control) [10–12]. These control strategies are typically enacted at municipality or regional extents, necessitating fine-scale understanding of where and when transmission may occur. Elucidating how environmental variables have influenced past changes in snail occurrence can help to anticipate which areas may benefit from increased surveillance for disease emergence and facilitate more precise allocation of resources.

Past efforts have mapped the current climate envelope of snail habitats at the national and regional scale as well as projected the climate envelope of the *Biomphalaria* under future climate conditions[13,14]. The last publication of a national model of Brazil in 2012 included climate variables and vegetation greenness to map the foundational by static distribution of *Biomphalaria*[13]. Since then, remote sensing climate and environmental data has improved in both quality and scale (i.e., it is available at finer temporal and spatial resolutions). In particular, fine spatial and temporal resolution land-use/land-cover, human population density, and hydrological variables are now available at the continental to global scale (e.g., 30m-1km; annual)[15,16]. An increased number of spatially (100m) and temporally precise (annual) georeferenced biodiversity data for Brazil, including presence of *Schistosoma* parasite-competent snails, are now available from a number of research and public collections. Finally, advances in machine learning, particularly the accessibility of tree-based machine learning models, allow us to not only identify important variables, which is typical of machine learning models, but allows us to assess the functional form of the relationship between the environmental variable and habitat suitability (e.g, habitat suitability increases with precipitation seasonality)[17,18]. Leveraging this new wealth of data and interpretable yet cutting edge machine learning algorithm, it is now possible to develop species distribution models of *Biomphalaria* snails across multiple time steps at a geographical extent - the whole Brazil - and fine-scale spatial resolutions - 1 km^2 - that were not possible ten years ago.

The goal of this paper was to map the historical and current distribution of the three competent snail intermediate hosts of schistosomes in Brazil using a wide range of remote sensing climatological and land use data, a cutting edge, powerful and highly flexible machine learning model, i.e., Extreme Gradient Boosted Regression Trees, and a unique data set of 11 thousands georeferenced records of *Schistsome* competent snail occurrence spanning over three decades. Specifically, we identified environmental variables that most contributed to the probability of snail occurrence and used counterfactual analyses to ask: (i) what are the key features of the environment, and their functional response, that influence snail habitat suitability? And (ii) how has the snail distribution shifted with climate and urbanization throughout the last three decades?

In contrast to projections of snail distribution under future climate change, that are difficult to validate, the long time series of snail data allowed us to (i) use hindcasting - i.e., the process of testing statistical model by having them compared to the actual historical observations to determine how well the models match the historical record - to

validate SDMs precision on the basis of actual field observations and (ii) counterfactual analysis to determine the relative importance of specific climatological and land use drivers in explaining the observed geographical range shift of schistosome competent snails. This work will help to formulate more reliable hypotheses about how species habitat suitability may change under current environmental trajectories, allowing us to evaluate if future projections are in line with what we expect based on the observed change through time.

RESULTS

We used machine-learning based (XGBoost) Species Distribution Models (SDMs), 30 years of expert-collected occurrence data, and remote sensing data to map the habitat suitability for each of the three *Schistosoma*-competent *Biomphalaria* species. SDMs use species occurrence coordinates and corresponding environmental variables to interpolate habitat suitability across large geographic extents, machine learning SDMs are particularly flexible models that can handle non-linear relationships, collinear features (covariates), and complex interactions among features[19,20]. We trained the models on data collected from 2000-2020. Using data collected from 1990-1999, we used a hindcasting approach to test the model.

After thinning the training dataset (data collected between 2000-2020), we included 165 snail occurrence points for the *B. glabrata* model, 283 occurrence points for the *B. straminea* model*,* and 173 occurrence points for *B. tenagophila*. Background points were 2x the number of occurrence points. After thinning the test dataset used for hindcasting (data collected between 1990-1999), *B. glabrata* had 40 occurrence points (proportional to 24% of the training set), *B. straminea* had 33 occurrence points (proportional to 12% of the training set), and *B. tenagophila* had 28 occurrence points (proportional to 16% of the training set). The number of background points used for hindcasting, which were based on retaining one point per 1km^2 grid cell, was 1688, 1691, and 1684 for *B. glabrata*, *B. straminea*, and *B. tenagophila,* respectively. Notably, these points were not used to train the model but only to test model predictions, therefore, class imbalance does not impact results. We present results based on the definition of urbanization based on an aggregation of South America definitions, but include results based on the UN's definition. Model results and interpretation are comparable across models (Supplementpp 7, 17-22).

Model validation & hindcasting performance

We used spatial cross-validation to evaluate model performance. Performance was high for each snail species: mean AUC and sensitivity was ≥ 0.80 and model sensitivity was ≥ 0.80 for all species. Mean model specificity was ≥ 0.70 for all species (Supplementp 7). Model performance was also high for the hindcasting analysis, where we used the trained model to predict out-of-sample data collected between 1990-1999. For all

species, AUC was > 0.80 (*B. glabrata* = 0.89; *B. straminea* = 0.87; *B. tenagophila* = 0.81) and sensitivity was > 0.80 (*B. glabrata* = 0.93; *B. straminea* = 0.82; *B. tenagophila* = 0.89) (Supplementp 7).

Feature contribution and functional response

Across the three species we found that, on average (i.e., across bootstrapping iterations), the features associated with climate, urbanization, and agricultural land-cover contributed the most to the predicted probability of snail occurrence, as opposed to features associated with hydrology or soil properties (Figure 1-3).

For *B. glabrata* and *B. straminea*, we found that the top five best predictors of snail occurrence were climate variables, with three of these variables associated with precipitation (precipitation seasonality, precipitation in the driest months and precipitation in the wettest quarter) and the fourth variable associated with temperature (isothermality) (Figure 1a, 2a). Our model indicates that, for both species, snail habitat suitability non-linearly increases with precipitation seasonality but decreases when precipitation in the wettest quarter was high (Figure 1b, 2b). For *B. glabrata,* habitat suitability also decreases with precipitation in the driest quarter (Figure 1b), while for *B. straminea,* habitat suitability decreases when annual precipitation was high (Figure 2b). The climate profile for *B. tenagophila* significantly differed from *B. glabrata* and *B. straminea*. Climate variables made up only two of the top five predictors (Figure 3a). *B. tenagophila* habitat suitability non-linearly decreases with precipitation seasonality and the temperature during the driest quarter (Figure 3b).

For all three snail species, we found that location along the urban to rural gradient was among the top five predictors, with location along the high density urban (>1500 per people per km^2 and > 150,000 in a contiguous area) to rural gradient also among the top five predictors for *B. tenagophila* (Figures 1a, 2a, 3a). Snail habitat suitability increased for all snails within pixels closer to urban areas (0 = in an urban area) (Figures 1b, 2b, 3b). According to the univariate PDP, the relationship between *B. tenagophila* and urban gradient followed a pattern of exponential decay, whereas the change in probability was more gradual for *B. glabrata,* and nearly linear for *B. straminea*. Agricultural crop cover and temporary crop cover were among the top six predictors for *B. straminea* and *B. tenagopila,* respectively, with habitat suitability increasing with crop cover (Figure 2-3). In the models using the UN definition of urbanization, agricultural mosaic was among the top five variables for *B. tenagophila* and *B. straminea* , with agricultural mosaic the second highest for *B. straminea* (Supplementpp 19-22).

Changing snail distribution

According to model predictions, we found that habitat suitability of *B. glabrata* decreased from 1992 - 2017 throughout western Brazil, particularly in the Eastern and Central regions between latitudes S 14°-24° and in the central regions between latitudes S 2°-8°,

and increased towards coastal areas of the Southeast and Northeast (Figure 3a-c). In contrast, we found that habitat suitability of *B. straminea* primarily expanded southward, whereas for *B. tenagophila* suitability contracted within the state of São Paulo (Figure 3d-i). In total, our model predicted a fairly large change in habitat suitability for each *Biomphalaria* host species: 58% (CI: 52%-66%) of the area of Brazil experienced at least a 1% change in *B. glabrata* suitability, 55% (CI: 45%-64%) for *B. straminea,* and 28% (CI: 19% - 34%) for *B. tenagophila* (Table 1; Supplementp 23). For *B. glabrata,* the area that decreased in suitability was equivalent to the area that increased whereas the increase in area of habitat suitability for *B. straminea* was greater than the area decrease, and the decrease in the area of habitat suitability for *B. tenagophila* was marginally higher than the increase (Table 1; Supplementp 23).

Based on our counterfactual analysis, for each species, large-scale changes in predicted habitat suitability were correlated with regional variations in climate between the two time periods (Table 1; Figure 4). The proportional increase and decrease in area of change in suitability associated with climate was nearly equivalent for all three species (Table 1; Supplementp 23), indicating broad scale shifts as opposed to net increases and decreases in suitability. The biggest growth in urbanization between these two time points was the growth of small-medium sized urban areas (300-1500 people per km^2 with > 2500 in a contiguous area) (Supplementp 26). As such, in contrast to climate response, changes in the probability of snail occurrence associated with urbanization were highly localized at the outskirts of existing cities or new cities that emerged between the two time points (Figure 5; Supplementpp 24-25), with a largely positive increase in suitability associated with urbanization (Table 1; Supplementpp 23-24). For example, in the Vale do Rio Doce, a meso-region bordering Minas Gerais and Espírito Santo States where schistosomiasis is endemic, there were numerous small-medium cities that emerged (e.g., Baxio Gaundú, Inhapim), while the area of the dense cities (Ipatinga and Governador Valadares) hardly changed (Figure 5a, d). As such, our model predicts that the most notable change in *B. glabrata* habitat suitability associated with urbanization is in the northeast and south of this mesoregion. In the meso-regions of the Recife Metropolitan Area, and area where *B. straminea* has been found infected with *S. mansoni* [21], and São Paulo Metropolitan area, an area where *B. tenagophila* has been associated with foci of transmission[22], the most notable changes in *B. straminea* and *B. tenagophila* habitat suitability were at city outskirts where new urban areas had developed around the periphery of the cities (Figure 5b, c, e, f).

DISCUSSION

Schistosomiasis snail intermediate host distributions depend strongly on climate and land use variation and, as a result, have shifted substantially in the last 30 years in Brazil (Figure 3). Using remote sensing data, machine learning, and a long-term dataset of expert-collected snail occurrence data, we mapped the distribution of snail habitat at a fine spatial resolution (1km^2) and large geographic extent (national) with high accuracy (Figure 3, Supplementp 7). We found that climate and urbanization features best predict

Biomphalaria habitat suitability across all three species, after controlling for sampling bias (Figure 1-3). Large-scale shifts in snail habitat suitability occurred due to regional changes in climatic variables, whereas urbanization influenced fine-scale, highly localized increases in habitat suitability within and around small-medium sized urban areas, most likely due to changing population sizes and lack of infrastructure in rapidly growing informal settlement, as people move from rural and urban areas both within and across Brazilian states[21,23] (Figure 4, Figure 5, Supplementpp 23-24, 26).

In regard to climatologies snail habitat suitability was influenced mostly by precipitation patterns for *B. straminea* and *B. glabrata* and by temperature for *B. tenagophila* (Figures 1-3). Counter-intuitively, as snails inhabit freshwater habitats, the *B. glabrata* habitat suitability was highest with low precipitation of the driest month (Figure 1b). Low precipitation may create small pockets of surface water ideal for snail habitat. Additionally, *B. glabrata* habitat suitability was highest in areas with high precipitation seasonality, i.e., a strong difference in precipitation between the dry season and the wet season (Figure 2b). Perhaps freshwater habitats in areas with low precipitation in the driest quarter but high precipitation seasonality accumulate enough water during the wet season to support snail survival and reproduction throughout the dry season. *B. straminea* was most likely to occur in areas with annual precipitation < 6000 kg m^2 $year^{-1}$(Figure 2b). Its habitat suitability is high across a large range of precipitation values but likely gets flushed out with intense precipitation. Further, *B. straminea* habitat is most suitable at high values of isothermality, suggesting that these snails are most likely to inhabit areas where seasonal temperature variation is low compared to diurnal variation (Figure 2b). Our model indicates that *B. tenagophila* has a distinct climatic niche as it is found in areas with low temperatures of the driest quarter, consistent with *B. tenagophila* almost exclusively found in the south of Brazil (Figures 3b). However, these results contrast previously published national models, which found variables associated with temperature to be the strongest contributors to model performance. This may be related to the mismatch of *Biomphalaria* occurrence points and the climatologies, which were retrieved from WorldClim that spans 1960-2000[24], whereas our climatologies overlapped with the decade the snail was collected. Yet, because this analysis is correlational, we cannot rule out the influence of unobserved variables that are correlated with these climatic conditions and may drive the discrepancy between the models.

The location along the urban to rural gradient was among the top three predictors of *Biomphalaria* habitat suitability for all three species (Figure 1-3). The distance to a small-medium urban area (300-1,500 people per 1km^2 with > 2,500 people in contiguous pixels) was important for all three snail species, and distance to high density urban area (>1500 people per 1km^2 with > 150,000 people in contiguous pixels) was among the top five predictors of *B. tenagophila* habitat suitability. Although the rate of change in snail habitat suitability and urban gradient differed between the three snail species, in general, habitat suitability for the three snails is higher closer to urban areas than rural areas (Figure 1-3). Case studies indicate that schistosomiasis, a disease that has been historically associated with rural areas, is urbanizing in Brazil [9], with empirical evidence

hypothesizing that urban and peri-urban areas provide the necessary abiotic factors for snail survival, while reducing the number of snail competitors and predators, ultimately increasing snail habitat suitability[9,25]. Small scale agricultural land-use found in peri-urban and urban areas – agricultural mosaics and temporary crops – were top predictors of snail habitat suitability, providing some insight on the types of habitats that urban areas provide[26]. However, as the location along the urban to rural gradient contributed more to model predictions than agricultural land-use and land-cover, there are likely other aspects of urban and peri-urban infrastructure that provide snail habitat, such as the lack of paved roads and wastewater treatment facilities in rapidly growing settlement, which favor the dispersion in the environment of *S. mansoni* eggs with fecal waste[21,27]. The location along the urban to rural gradient is a rough proxy for estimating the impact of urban dynamics and should be updated in future studies to better describe urban and peri-urban areas and estimate edge effects, especially as they may lag behind change in population size, and determine specific socio-ecological mechanisms predicting fine-scale snail habitat suitability.

Our models suggest that climate and urbanization may have significantly contributed to a shift in *Biomphalaria* habitat suitability over the last three decades (Figure 4-5, Supplement24-25). Distributions of *B. glabrata,* the most competent intermediate host, have shifted to coastal areas, particularly of the southeast (Figure 3), with our counterfactual analysis indicating that this change is correlated with both climate, particularly increased precipitation seasonality and drier dry seasons, and urbanization (Figures 4, Supplementp 24). Of the three host snails, the distribution of the habitat suitability for *B. straminea* has changed the most, with a large increase in habitat suitability throughout the southeast, reflective of decreasing annual precipitation and increasing isothermality within Central West and southern Brazil across the two time points, and local changes associated with urbanization (Figure 2, Figure 4-5, Supplementpp 15-16). While *B. glabrata* is the most competent host, the expansive and moving habitat suitability of *B. straminea* suggests that this species may play a larger role in schistosomiasis persistence under a changing environment. In contrast, the geographic extent of *B. tenagophila* habitat suitability has decreased throughout the last three decades, most likely due to temporal variation in temperatures and reduced precipitation seasonality in southeastern Brazil (Figure 2, Figure 4, Supplementpp 15-16).

While this analysis indicates large geographic shifts in snail habitat suitability, on-the-ground observations may not detect this shift as land-use, particularly urbanization, may create localized pockets that additively amplify or dilute snail habitat suitability in adjacent areas (Figure 5; supplementpp 13). For example, around some new cities we observe increases, roughly 10 km wide, in suitable habitat associated with both climate and urbanization. For instance, the cities of Inhapim and Baixo Guandú grew from rural areas to small cities with > 20,000 people between the two time points. Both cities have reported some of the highest cases of schistosomiasis between 2007-2017 in their respective states (Minas Gerais and Espírito Santo)[28,29]. Notably, our analysis shows

that there has been a decrease in suitability associated with the “disappearance” of small-medium urban areas between the time points. This warrants further on-the-ground investigation and validation as this result may be reflective of change in infrastructure associated with population size or might be an artifact of our definition of urbanization, where the population size and/or density is only marginally drops below our urban population size and density thresholds, thus showing up on maps as a dissolved urban area when in reality there has been little change in infrastructure that impacts schistosomiasis risk. Overall, this analysis provides a fine scale yet bird’s-eye view of changing snail habitat suitability, allowing us to tease apart these interactions and better understand change in snail distribution through time.

Despite the three intermediate host snails having distinct socio-ecological niches within Brazil, our work underscores the importance of species-specific combinations of temperature, precipitation, and urban conditions that are highly predictive of snail occurrence, and therefore important indicators of schistosomiasis risk. Proactively anticipating shifts in snail habitat suitability and schistosomiasis risk is a major priority for targeting snail control and public health services and for mitigating disease emergence. Long-term snail surveillance and control programs face shifting demands as the seasonality, intensity, and geography of snail populations change with climate, land-use, and socio-ecological conditions. Our study shows that snail habitat suitability has changed throughout the last three decades, and provides profiles of exposure risk that can help shift the spatial priorities of control efforts with changing socio-ecological conditions, land-use, and climate.

METHODS

Occurrence points and data thinning

We used a combination of snail occurrence points collected by the Brazilian government programs, FioCruz Minas Gerais and São Paulo Department of Public Health. In total, the programs collected 11,299 snail records that spanned 1990–2020 (Supplementpp 1;8). For our analysis, for each focal species, we thinned data so that only one occurrence point was retained per 1km^2 grid cell using the R package *dismo*[30]. Thinning data is essential to control for pseudo-replication[31].

Background points

There is often a lack of true species absence data for model input and thus background points or pseudo-absences are used as a proxy for where the focal species does not occur[20]. To generate background points, we created a background mask using species records from the Global Biodiversity Information Facility (GBIF)[32] as well as expert collected snails that were not the focal species. Specifically, from GBIF, we used georeferenced records from all freshwater animals in Brazil. We downloaded 55,057 coordinates, spanning 1992-2020, from 2135 species. Records from all freshwater animals allowed us to create a landscape of freshwater animal sampling locations and, thus, where snail species could plausibly have been sampled. These methods help to

account for sampling bias[20]. We first overlaid a 1km^2 resolution grid over Brazil. Each cell was then assigned a probability value based on the number of background points (records) within each grid cell. Using these probabilities, we selected the background points by probabilistically sampling twice the number of occurrence points from unique 1km^2 grid cells. This procedure was repeated for each of the three competent host snail species.

Environmental data processing

We included features related to climate, land use/land cover (LULC), hydrology, topography, and soil properties in our distribution models (Supplementpp 1-2; 4-6). We used the CHELSA v2.1 monthly products to calculate 10-year climatologies spanning four decades (1980-1989, 1990-1999, 2000-2009, 2010-2017)[33]. The CHELSA dataset occurred at the largest spatial resolution: 1 km^2. To minimize issues with spatial autocorrelation, we aggregated all other spatial covariates to a spatial resolution of 1 km^2.

To estimate land-use and land-cover (LULC) variables, we used the MAPBIOMAS land-cover rasters collection 7.0 v2 to calculate the percent cover of temporary crops (excluding soy, sugarcane, rice, and cotton) and mosaic of use – two land-use types associated with free standing, freshwater water of urban and peri-urban farms[34] – within the 1km^2 grid cell during the year the data point was collected[16]. We also incorporated hydrology and water availability using the Global Surface Water dataset[35], MERIT Hydro data (100m resolution)[36], and WWF HydroSHEDS Free Flowing River dataset[37], topography using the NASA NASADEM Digital Elevation dataset (30m resolution)[38], and the OpenLandMap dataset (250 m resolution) to calculate soil properties[39–44]. We removed highly correlated features ($r > 0.75$) (Supplementpp 1-2; 4-6).

Urban-to-Rural gradient index

Globally, schistosomiasis is considered a rural disease but has been observed close to or in urban areas in Brazil. To measure the impact of urbanization on *Biomphalaria* occurrence, and because a urban:rural dichotomy does not capture the complexity of urbanization in Brazil[45], we derived a urban to rural gradient, where 0 indicated that the snail was observed in an urban area and 1 indicated the snail was in a rural area, and continuous values from 0-1 indicate the relative location of the snail along this gradient. We defined urban and rural areas using population density and population size as a proxy. First, we defined an urban area based on an aggregation of urban definitions throughout South America: an urban area had population densities from 300-1500 people within 1 km^2 with contiguous pixels totalling at least 2500 people[46]. Next, we included high density urban areas in both analyses: > 1500 people per grid cell with contiguous grid cells totalling more than 150,000 people[47]. As population-based definitions of urban are inconsistent, as a separate analysis, we repeated the analysis with alternative definitions of urbanization as described in supplementp 2, and the results yielded similar results (Supplementpp 7, 17-22). We used WorldPop to estimate population size and cumulative cost mapping – where every pixel is assigned the total

cost of the lowest cost path (distance) to an urban pixel – to estimate the distance to the urban pixel for points collected between 2000 - 2020[15]. WorldPop data is not available before 2000; as such, we derived population estimates from 1990 - 1999 by linearly interpolating between the 2000 WorldPop raster and the Global Human Development Layer, Population Grid 1975 raster[34]. To calculate the final variable - location along the urban to rural gradient - we divided distance to an urban area by distance to a rural area and standardized it from 0-1 using a max-min transformation. All environmental data was processed and aggregated to 1km^2 grid cells around each data point using Google Earth Engine[48] and the GEE Python API in Google Co-Laboratory, with the exception of CHELSA, which was downloaded from their AWS repository[33,49].

Extreme Gradient Boosted Regression Tree SDMs

We used extreme gradient boosted regression tree models, using package *Xgboost* [50], for each SDM as boosted regression models were determined to yield the best performance and interpretability for modeling snail habitat in Brazil[19]. Extreme gradient boosted regression is a machine learning algorithm that creates an ensemble of weak decision trees to form a stronger prediction model by iteratively learning from weak classifiers and combining them into a strong classifier (i.e., boosting)[17]. Gradient boosted regression is flexible in that it allows for non-linearity, both among features (i.e., interactions) and between features and predictions, higher collinearity among features than statistical models, and non-random patterns of missing data, characteristics frequently observed in ecological data. Spatial cross-validation (5-fold), using the R package *blockCV*[51], was used to estimate the model's predictive power in geographic regions not used for training so as to minimize inflating model performance values due to spatial autocorrelation of environmental predictors[51]. Model parameters were tuned and trained using Bayesian optimization via the R package *bayesianOptimization*[52]. The suite of tuned hyperparameters is described in supplementpp 2. Model sensitivity (i.e., the proportion of occurrences the model correctly identifies as occurrences), model specificity (i.e., the proportion of background points that the model correctly identifies as background), and model area-under-the-curve (AUC; i.e., a measure that calculates how well the model correctly distinguishes occurrence points from background points, where AUC ≤ 0.5 indicates the model performs no better than a coin flip) were calculated using the R package *pROC*[53] and *caret*[54]. Model training and validation was conducted using data from 2000-2020.

To evaluate the average contribution of each covariate to model predictions, we first identified important features using SHAP (Shapley Additive Explanations) Values calculated via the R package *SHAPforxgboost*[55] (supplementp 3). Next, to generate environmental profiles for each snail, we constructed partial dependence plots (PDPs) using the R package *pdp*[56]. PDPs illustrate the average relationship between habitat suitability and the feature of interest. The average relationship is calculated by fitting models with all combinations of the non-focal variables for each value of the focal feature. To account for uncertainty in the model, feature contribution (SHAP Values) and PDPs were calculated 25 times, where, for each iteration, the model was trained on a randomly selected 80% of the data (bootstrapping). Data and code are available at https://github.com/ckglidden/biomphalaria-sdm-brazil.

Hindcasting & Counterfactuals

Our ultimate objective was to use historical information to estimate how *Biomphalaria* species distributions have changed with land-use and climate. To evaluate whether our model could track changes in snail distribution through time, we trained the model on all data from 2000-2020 and tested the model on data collected from 1992-1999. We calculated model sensitivity, specificity, and AUC to evaluate how well the model could predict suitable snail habitats on historic out-of-sample data.

We then conducted a counterfactual analysis to examine how changes in climate and urbanization contributed to observed changes in *Biomphalaria* distribution over time. First, using the model trained on 2000-2020 data, we created prediction maps for the historical distribution of each snail (using data from 1992) and the current distribution of the snail (using data from 2017). We then created a climate counterfactual map by estimating model predictions using current data for all variables except climate, for which we used historical data (i.e., predicted what the current distribution would be if all variables but climate changed between the two time points). We repeated this process for urban variables to create an urban counterfactual. Finally, we compared the observed change in distribution between 1992-2017 with the observed change in distribution using the counterfactual scenario to estimate the change in distribution that was correlated with each set of features (i.e., climate or urban). As such, for example, in location *x*, if the change of probability of occurrence in the observed map is higher than change in probability in the counterfactual map (when climatology remained at historical values), then we infer that change in climatology is, at least in part, correlated with the increased probability observed between the two observed time periods. We created prediction maps for each of the 25 iterations described in the bootstrapping protocol above.

ACKNOWLEDGEMENTS

This work was supported by the Belmont Collaborative Forum on Climate, Environment and Health (US-NSF ICER-2024383, FAPESP), by a grant of the Stanford Center for Innovation in Global Health, and the Stanford Program for Disease Ecology, Health and the Environment. GADL was partially supported also by an NSF EEDI grant (DEB – 2011179). EAM and CKG were supported by the National Science Foundation and the Fogarty International Center (grant no. DEB-2011147). EAM was additionally supported by the National Institute of Allergy and Infectious Diseases (grant nos R01AI168097 and R01AI102918), the National Institutes of Health (grant no. R35GM133439), and by seed grants from the Stanford Woods Institute for the Environment, King Center on Global Development, Center for Innovation in Global Health and Terman Award. CKG was additionally supported by a Stanford Institute for Human-centered Artificial Intelligence Postdoctoral Fellowship.

AUTHOR CONTRIBUTIONS

CKG, ALS, AC, RT, RP, MM, EM, SHS, GADL conceived and designed the analysis; RT, RP, RC collected the data; KML, PL, VS, TA contributed data or analysis tools; CKG, ALS, AC performed the analysis; CKG, ALS, RT, RP, MM, EM, GADL wrote the paper.

DATA & CODE AVAILABILITY

https://github.com/ckglidden/biomphalaria-sdm-brazil/

535 **REFERENCES**

536 1. Colley, D. G., Bustinduy, A. L., Secor, W. E. & King, C. H. Human schistosomiasis. *Lancet*
537 *Lond. Engl.* **383**, 2253–2264 (2014).
538 2. World Health Organization. Schistosomiasis: key facts. https://www.who.int/news-room/fact-
539 sheets/detail/schistosomiasis (2019).
540 3. Pan American Health Organization. Schistosomiasis. *Pan American Health Organization /*
541 *World Health Organization*
542 https://www3.paho.org/hq/index.php?option=com_content&view=article&id=9474:schistoso
543 miasis-factsheet&Itemid=0&lang=en#gsc.tab=0 (2014).
544 4. World Health Organization. *Ending the neglect to attain the Sustainable Development*
545 *Goals: A road map for neglected tropical diseases 2021–2030*. (World Health Organization,
546 2020).
547 5. De Leo, G. A. *et al.* Schistosomiasis and climate change. *The BMJ* **371**, m4324 (2020).
548 6. Paraense, W. L. The schistosome vectors in the Americas. *Mem. Inst. Oswaldo Cruz* **96**
549 **Suppl**, 7–16 (2001).
550 7. Nguyen, K. H. *et al.* Interventions can shift the thermal optimum for parasitic disease
551 transmission. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2017537118 (2021).
552 8. Barbosa, F. S. & Barbosa, C. S. The bioecology of snail vectors for schitosomiasis in Brazil.
553 *Cad. Saúde Pública* **10**, 200–209 (1994).
554 9. Klohe, K. *et al.* A systematic literature review of schistosomiasis in urban and peri-urban
555 settings. *PLoS Negl. Trop. Dis.* **15**, e0008995 (2021).
556 10. WASH and snail control interventions. in *WHO guideline on control and elimination of*
557 *human schistosomiasis [Internet]* (World Health Organization, 2022).
558 11. Sokolow, S. H. *et al.* To Reduce the Global Burden of Human Schistosomiasis, Use 'Old
559 Fashioned' Snail Control. *Trends Parasitol.* **34**, 23–40 (2018).
560 12. Sokolow, S. H. *et al.* Global Assessment of Schistosomiasis Control Over the Past Century

Shows Targeting the Snail Intermediate Host Works Best. *PLoS Negl. Trop. Dis.* **10**, e0004794 (2016).

13. Scholte, R. G. C., Carvalho, O. S., Malone, J. B., Utzinger, J. & Vounatsou, P. Spatial distribution of Biomphalaria spp., the intermediate host snails of Schistosoma mansoni, in Brazil. *Geospatial Health* **6**, S95–S101 (2012).
14. Palasio, R. G. S., de Azevedo, T. S., 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.* **214**, 105764 (2021).
15. Sorichetta, A. *et al.* High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. *Sci. Data* **2**, 150045 (2015).
16. Azevedo, T., Sr., Souza, C. M., Jr., Shimbo, J. & Alencar, A. MapBiomas initiative: Mapping annual land cover and land use changes in Brazil from 1985 to 2017. **2018**, B22A-04 (2018).
17. Molnar, C. *Interpretable Machine Learning*.
18. Elith, J. *et al.* A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* **17**, 43–57 (2011).
19. Singleton, A. L. *et al.* Species distribution modeling for disease ecology: a multi-scale case study for schistosomiasis host snails in Brazil. *medRxiv* 2023.07.10.23292488 (2023) doi:10.1101/2023.07.10.23292488.
20. Elith, J. *et al.* A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* **17**, 43–57 (2011).
21. Favre, T. C. *et al.* A Longitudinal Study on the Natural Infection of Biomphalaria straminea and B. glabrata by Schistosoma mansoni in an Endemic Area of Schistosomiasis in Pernambuco, Brazil.
22. Palasio, R. G. S., de Jesus Rossignoli, T., Di Sessa, R. C. S., Ohlweiler, F. P. & Chiaravalloti-Neto, F. Spatial analysis of areas at risk for schistosomiasis in the Alto Tietê

Basin, São Paulo, Brazil. *Acta Trop.* **224**, 106132 (2021).

23. de Lima Amaral, E. F. Brazil: internal migration. in *The Encyclopedia of Global Human Migration* (Blackwell Publishing Ltd., 2013).
24. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978 (2005).
25. De Sousa dos Santos, S. *et al.* Schistosomiasis mansoni and hydrographical conditions in São Carlos, São Paulo, Brazil. *Trans. R. Soc. Trop. Med. Hyg.* **117**, 391–400 (2023).
26. Njoh, D. B. *et al. Urban and peri-urban agriculture in Bamenda: A policy narrative*. (2018).
27. de Souza Gomes, E. C. *et al.* Risk analysis for occurrences of schistosomiasis in the coastal area of Porto de Galinhas, Pernambuco, Brazil. *BMC Infect. Dis.* **14**, 101 (2014).
28. Ferrari, D. F. *et al.* ESQUISTOSSOMOSE – PERFIL EPIDEMIOLÓGICO DOS CASOS ESPÍRITO SANTO – 2007 A 2017. in vol. 1 1 (2018).
29. Alves, M. R. Socio-environmental factors and schistosomiasis hospitalizations in Minas Gerais, Brazil (2007-2017)/ Factores socioambientales y hospitalizaciones por esquistosomiasis en Minas Gerais, Brasil (2007-2017)/ Fatores socioambientais e internações por esquistossomose em Minas Gerais, Brasil (2007-2017). *J. Health NPEPS* **7**, (2022).
30. Hijmans, R. J., Phillips, S., Leathwick, J., Elith, J. & Hijmans, M. R. J. Package ‘dismo’. *Circles* **9**, 1–68 (2017).
31. Boria, R. A., Olson, L. E., Goodman, S. M. & Anderson, R. P. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecol. Model.* **275**, 73–77 (2014).
32. GBIF. https://www.gbif.org/.
33. Karger, D. N., Schmatz, D. R., Dettling, G. & Zimmermann, N. E. High-resolution monthly precipitation and temperature time series from 2006 to 2100. *Sci. Data* **7**, 248 (2020).
34. European Commission, Joint Research Centre (JRC) & Columbia University, Center for

International Earth Science Information Network - CIESIN. GHS-POP R2015A - GHS population grid, derived from GPW4, multitemporal (1975, 1990, 2000, 2015) - OBSOLETE RELEASE. (2015).

35. Pekel, J.-F., Cottam, A., Gorelick, N. & Belward, A. S. High-resolution mapping of global surface water and its long-term changes. *Nature* **540**, 418–422 (2016).
36. MERIT Hydro: Global Hydrography Datasets | Earth Engine Data Catalog. *Google for Developers* https://developers.google.com/earth-engine/datasets/catalog/MERIT_Hydro_v1_0_1.
37. Mapping the world's free-flowing rivers: data set and technical documentation. (2019) doi:10.6084/m9.figshare.7688801.v1.
38. NASADEM: NASA NASADEM Digital Elevation 30m | Earth Engine Data Catalog. *Google for Developers* https://developers.google.com/earth-engine/datasets/catalog/NASA_NASADEM_HGT_001.
39. Hengl, T. & Wheeler, I. Soil organic carbon content in x 5 g / kg at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution. (2018) doi:10.5281/zenodo.2525553.
40. Hengl, T. Sand content in % (kg / kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution. (2018) doi:10.5281/zenodo.2525662.
41. 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. https://zenodo.org/records/2784001.
42. Hengl, T. Clay content in % (kg / kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution. (2018) doi:10.5281/zenodo.2525663.
43. Hengl, T. Soil bulk density (fine earth) 10 x kg / m-cubic at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution. (2018) doi:10.5281/zenodo.2525665.
44. Hengl, T. Soil pH in H2O at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution. (2018) doi:10.5281/zenodo.2525664.

45. Cardoso, A. C. D., dal'Asta, A. P. & Monteiro, A. M. V. O que é o urbano na Amazônia contemporânea? Implicações para a vigilância em saúde no bioma. *Cad. Saúde Pública* **39**, (2023).
46. United Nations Statistics Division - Demographic and Social Statistics. https://unstats.un.org/unsd/demographic/products/dyb/dyb2005.htm.
47. The World Bank. Sustainable cities and communities. https://datatopics.worldbank.org/sdgatlas/archive/2017/SDG-11-sustainable-cities-and-communities.html (2017).
48. Gorelick, N. *et al.* Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**, 18–27 (2017).
49. s3_browsing. https://envicloud.wsl.ch/#/?prefix=chelsa%2Fchelsa_V2%2FGLOBAL%2F.
50. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (2016). doi:10.1145/2939672.2939785.
51. Valavi, R., Elith, J., Lahoz-Monfort, J. J. & 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.* **10**, 225–232 (2019).
52. Yan, Y. rBayesianOptimization. (2023).
53. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
54. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **28**, 1–26 (2008).
55. Liu, Y. SHAPforxgboost. (2023).
56. Greenwell, B. M. pdp: An R Package for Constructing Partial Dependence Plots. *R J.* **9**, 421–436 (2017).

Table 1. The % of the area of Brazil that is predicted to have experienced a change in habitat suitability between 1992-2017. The counterfactuals indicate the % of area change associated with those features. The table presents the mean values, with the 95% confidence intervals in parentheses across 25 bootstrap iterations.

predicted distribution (% change in area)			
	total change	**positive change**	**negative change**
B. glabrata	0.58 (0.52-0.66)	0.29 (0.36-0.34)	0.29 (0.24-0.36)
B. straminea	0.55 (0.4-0.64)	0.31 (0.27-0.34)	0.24 (0.16-0.30)
B. tenagophila	0.28 (0.19-0.34)	0.12 (0.08 - 0.16)	0.25 (0.11-0.19)
climate counterfactual (% change in area)			
	total change	**positive change**	**negative change**
B. glabrata	0.54 (0.46-0.62)	0.25 (0.21-0.30)	0.29 (0.23-0.37)
B. straminea	0.45 (0.33-0.56)	0.24 (0.17-0.30)	0.21 (0.12-0.29)
B. tenagophila	0.18 (0.13-0.24)	0.07 (0.04-0.10)	0.11 (0.08-0.16)
urban counterfactual (% change in area)			
	total change	**positive change**	**negative change**
B. glabrata	0.16 (0.12-0.20)	0.11 (0.08-0.14)	0.05 (0.03-0.07)
B. straminea	0.20 (0.15-0.24)	0.14 (0.11-0.18)	0.06 (0.04 - 0.07)
B. tenagophila	0.04 (0.02-0.07)	0.03 (0.05-0.04)	0.01 (0.006-0.024)

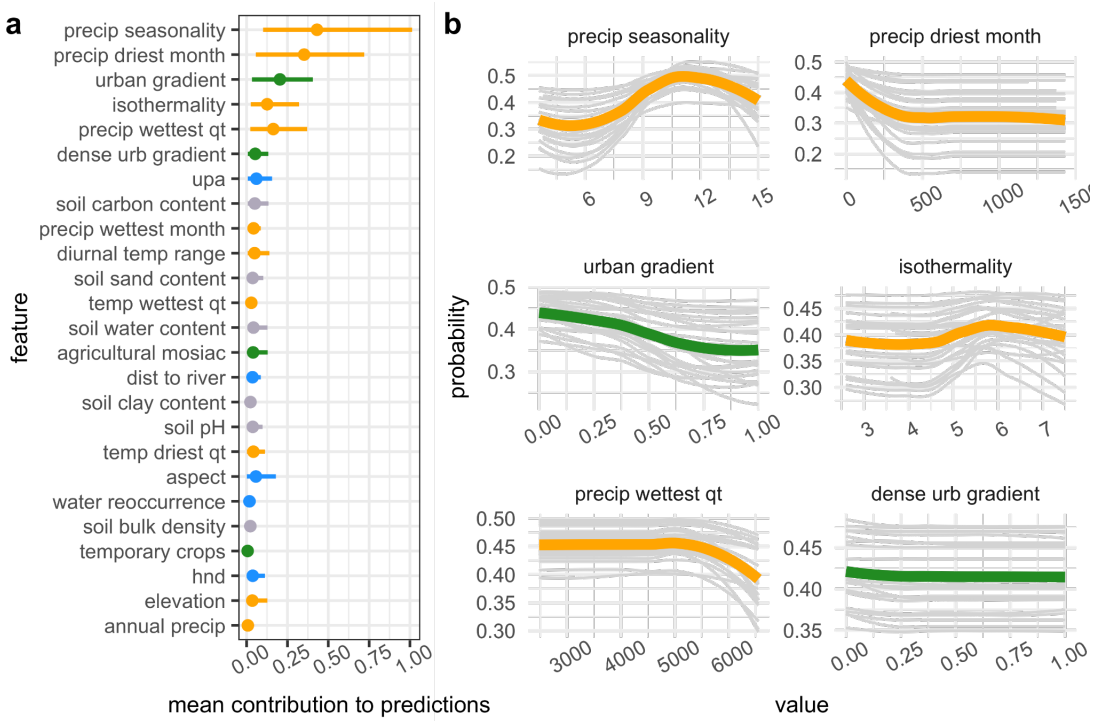

Figure 1. Model summary for the *B. glabrata* model. Feature contribution for each feature. Points are the absolute value of the mean contribution to predictions (mean |SHAP Value|) for the covariate across all data points (i.e., global feature contribution), bars represent the 95% confidence interval **(a)**. Response of snail occurrence to top six contributing variables. For each univariate plot, gray lines represent the probability of occurrence after controlling for all other covariates for each bootstrapping iteration and green (land-use variable) or orange (climate variable) represent the average across all iterations **(b)**. Grey color is used for soil type/cover and blue for water associated features.

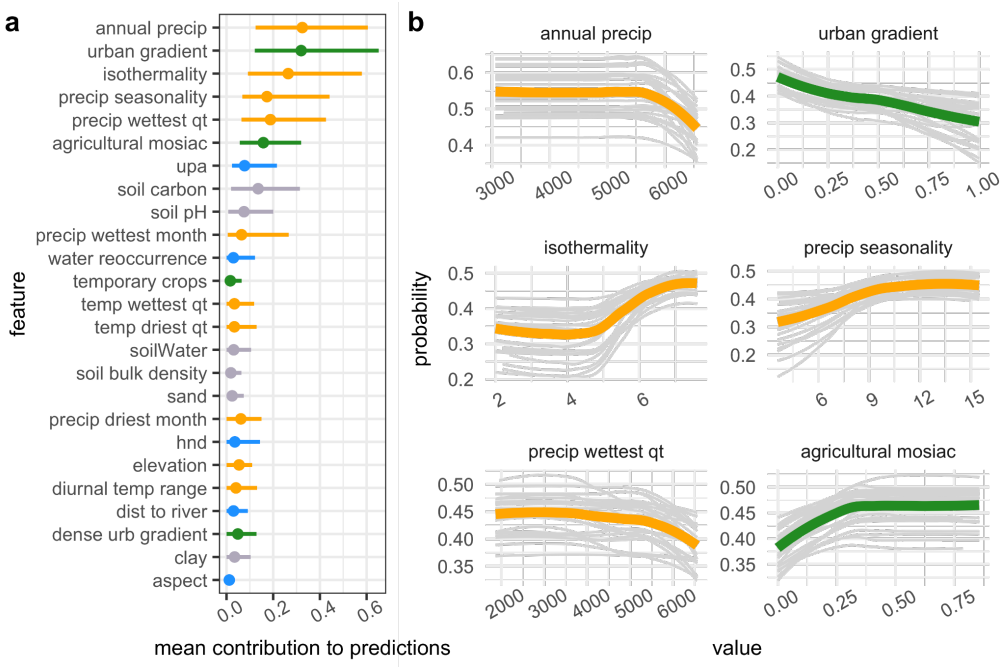

Figure 2. Model summary for the *B. straminea* model. Feature contribution for each feature. Points are the absolute value of the mean contribution to predictions (mean |SHAP Value|) for the covariate across all data points (i.e., global feature contribution), bars represent the 95% confidence interval. **(a)**. Response of snail occurrence to top six contributing variables. For each univariate plot, gray lines represent the probability of occurrence after controlling for all other covariates for each bootstrapping iteration and green (land-use variable) or red (climate variable) represent the average across all iterations **(b)**. Color coding as described in Fig. 1.

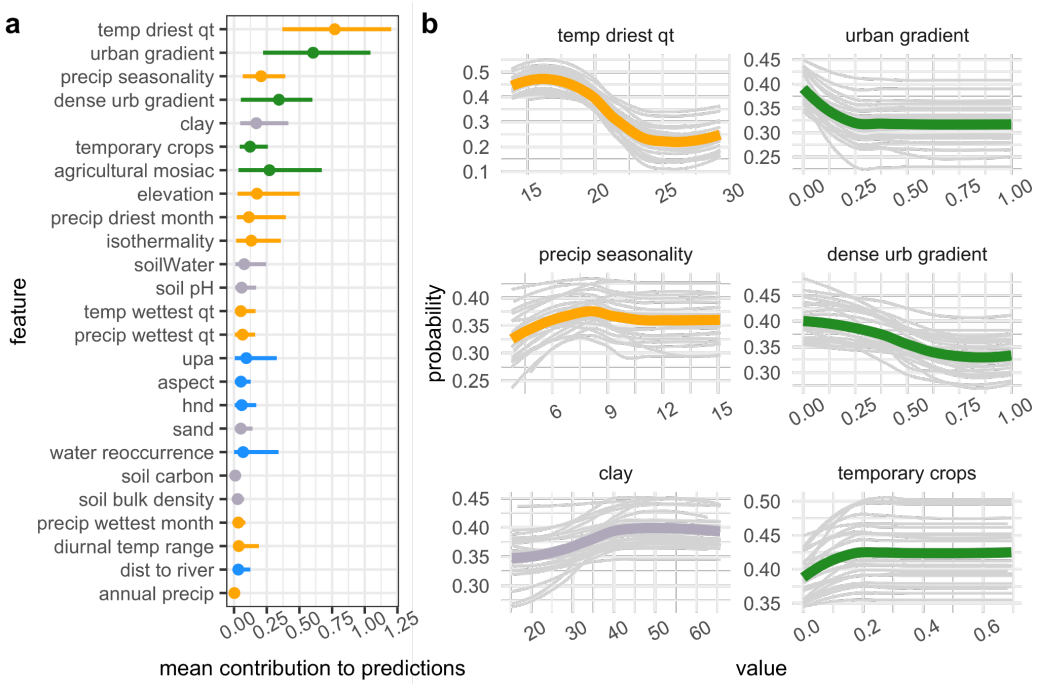

Figure 3. Model summary for the *B. tenagophila* model. Feature contribution for each feature. Points are the absolute value of the mean contribution to predictions (mean |SHAP Value|) for the covariate across all data points (i.e., global feature contribution), bars represent the 95% confidence interval. **(a)**. Response of snail occurrence to top six contributing variables. For each univariate plot, gray lines represent the probability of occurrence after controlling for all other covariates for each bootstrapping iteration and green (land-use variable) or red (climate variable) represent the average across all iterations **(b)**. Color coding as described in Fig. 1.

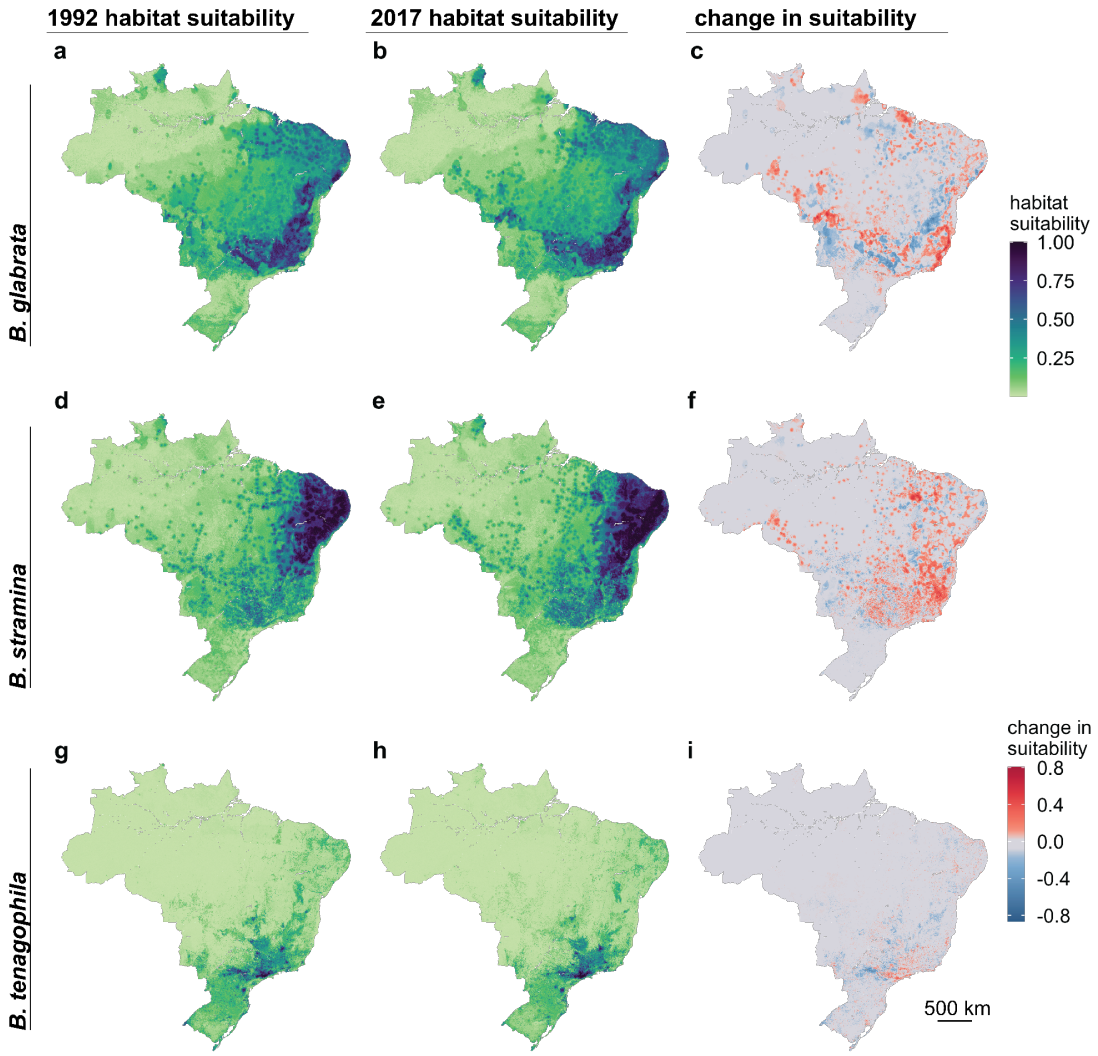

Figure 3. Predicted habitat suitability of the three competent snail hosts. The figure includes the distribution of each snail species in 1992 (first column), 2017 (second column), and the change in habitat suitability between the two time points (third column; blue indicates an increase in habitat suitability and red indicates a decrease in habitat suitability). Spatial resolution is 1km^2. Figures are mean values per pixel across 25 bootstrapping iterations. Predicted values are scaled from 0 to 1 using a min-max transformation, using the minimum and maximum value across the two time points to rescale across all predicted probabilities of snail occurrence for each species.

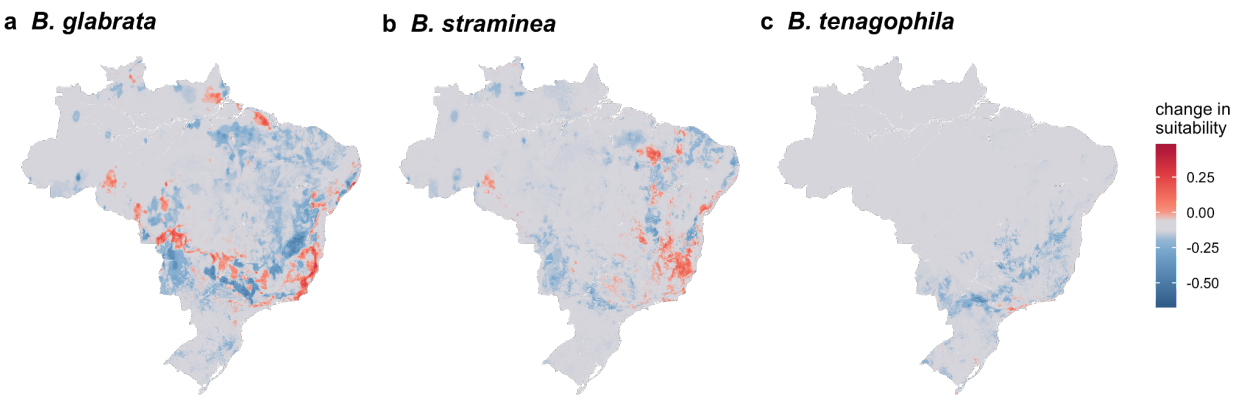

Figure 4. Climate influences large scales change in suitability. The difference in distribution between the predicted distribution of 2017 and when holding climate features (i.e., as they were observed in 1992). Blue indicates the habitat that became more suitable with change in climate and red indicates the climate became less suitable with the observed change in climate (i.e., the change in distribution is correlated with the change in climate). Figures are mean values per pixel across 25 bootstrapping iterations.

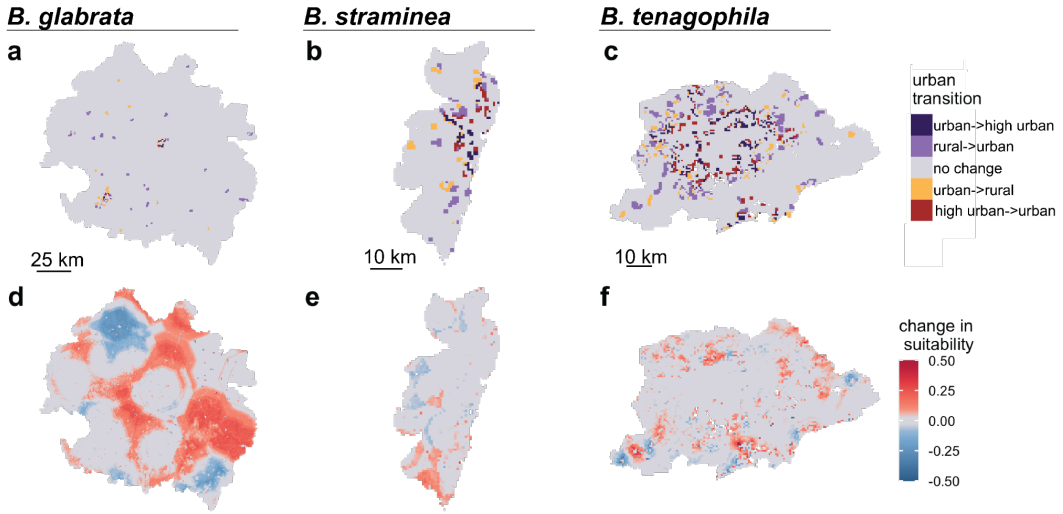

Figure 5. An example of how urbanization influences change in habitat suitability on small scales. The change in urban extent for the Vale do Rio Doce, Recife Metropolitan, and São Paulo Metropolitan mesoregions between 1992 and 2017 (**a-c)** and the difference in snail probability for 2017 when using urban-rural gradient values for 2017 versus 1992 (i.e, the change in snail probability attributed to change in urban distribution and extent) for *B. glabrata* in Vale do Rio Doce **(d),** *B. straminea* in Recife **(e)**, and *B. tenagophila* in São Paulo **(f)**. The Vale de Rio Doce region is much larger than the other two mesoregions, but the change in urban extent highlights that there was a significant amount of emergence of small-large cities (300-1500 people per 1km^2 pixel, with >2500 people in contiguous pixels) and this is a associated with large changes in habitat suitability. We chose each area for each *Biomphalaria* species as they have been associated with foci of transmission and/or infected snail in each respective mesoregion.