

MR. CHAZ HYSENI (Orcid ID : 0000-0003-2567-8013)  
DR. PAUL O MIREJI (Orcid ID : 0000-0002-7965-2428)  
DR. NORAH SAARMAN (Orcid ID : 0000-0001-8974-0301)

Article type : Original Article

**Title:** A machine learning approach to integrating genetic and ecological data in tsetse flies (*Glossina pallidipes*) for spatially explicit vector control planning

**Running title:** Tsetse fly habitat use and connectivity

**Authors:** Anusha Bishop<sup>1,2</sup>, Giuseppe Amatulli<sup>3</sup>, Chaz Hyseni<sup>4</sup>, Evelyn Pless<sup>1,5</sup>, Rosemary Bateta<sup>6</sup>, Winnie A. Okeyo<sup>6,7</sup>, Paul O. Mireji<sup>6,8</sup>, Sylvance Okoth<sup>6</sup>, Imna Malele<sup>9</sup>, Grace Murilla<sup>6</sup>, Serap Aksoy<sup>10</sup>, Adalgisa Caccone<sup>1</sup>, Norah P. Saarman<sup>1,11\*</sup>

<sup>1</sup>Department of Ecology and Evolutionary Biology, Yale University, CT, USA. <sup>2</sup>Department of Environmental Science, Policy, & Management, University of California, Berkeley, CA, USA.

<sup>3</sup>School of Forestry and Environmental Studies, Yale University, CT, USA. <sup>4</sup>Department of Ecology and Genetics, Uppsala University, Sweden. <sup>5</sup>Department of Anthropology, University of California, Davis, CA, USA <sup>6</sup>Biotechnology Research Institute, Kenya Agricultural and Livestock Research Organization, Kikuyu, Nairobi, Kenya. <sup>7</sup>Department of Biomedical Sciences and Technology, School of Public Health and Community Development, Maseno University, Maseno, Kisumu, Kenya. <sup>8</sup>Centre for Geographic Medicine Research Coast, Kenya Medical Research Institute, Kilifi, Kenya. <sup>9</sup>Vector and Vector Borne Diseases Research Institute, Tanzania Veterinary Laboratory Agency, Tanga, Tanzania. <sup>10</sup>Department of Epidemiology of Microbial Diseases, Yale School of Public Health, CT, USA. <sup>11</sup>Department of Biology, Utah State University, UT, USA.

\*Corresponding author norah.saarman@usu.edu, Utah State University, Logan, UT, USA

## ABSTRACT

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1111/eva.13237](https://doi.org/10.1111/eva.13237)

This article is protected by copyright. All rights reserved

**Introduction** - Vector control is an effective strategy for reducing vector-borne disease transmission, but requires knowledge of vector habitat use and dispersal patterns. Our goal was to improve this knowledge for the tsetse species *Glossina pallidipes*, a vector of human and animal African trypanosomiasis, which are diseases that pose serious health and socioeconomic burdens across sub-Saharan Africa.

**Methods and Results** - We used random forest regression to: (i) Build and integrate models of *G. pallidipes* habitat suitability and genetic connectivity across Kenya and northern Tanzania, and (ii) provide novel vector control recommendations. Inputs for the models included field-survey records from 349 trap locations, genetic data from 11 microsatellite loci from 659 flies and 29 sampling sites, and remotely sensed environmental data. The suitability and connectivity models explained approximately 80% and 67% of the variance in the occurrence and genetic data, and exhibited high accuracy based on cross-validation. The bivariate map showed that suitability and connectivity vary independently across the landscape and inform vector control recommendations. Post-hoc analyses show spatial variation in the correlations between the most important environmental predictors from our models and each response variable (e.g. suitability and connectivity) as well as heterogeneity in expected future climatic change of these predictors.

**Discussion** - The bivariate map suggests that vector control is most likely to be successful in the Lake Victoria Basin, and supports the previous recommendation that *G. pallidipes* from most of eastern Kenya should be managed as a single unit. We further recommend that future monitoring efforts should focus on tracking potential changes in vector presence and dispersal around the Serengeti and the Lake Victoria basin based on projected local climatic shifts. The strong performance of the spatial models suggests potential for our integrative methodology to be used to understand future impacts of climate change in this and other vector systems.

**Keywords** - disease vector, gene flow, habitat suitability, landscape genetics, random forest, spatial modelling

## 1 INTRODUCTION

2 Worldwide, vector-borne diseases account for more than 17% of all infectious diseases in  
3 humans, and represent a significant socioeconomic burden through decreases in livestock milk  
4 production, birth rates, weight gain and survival (Chanie et al., 2013; Narladkar, 2018; Rohr et al.,  
5 2019). The potential of a vector to transmit a pathogen is heterogeneous across the landscape because  
6 of variation in the disease, vector, and risk of contact between host and vector. Variation in  
7 distribution is caused by complex evolutionary and ecological interactions between the organism and  
8 the local environment over multiple generations. Ultimately, variation in vector survival and dispersal  
9 are two components that most strongly influence long-term disease transmission. Both survival and  
10 dispersal can be modeled spatially as estimates of habitat suitability and genetic connectivity (Bouyer  
11 et al., 2015; Dicko et al., 2014; Hirzel et al., 2008), which can improve our ability to plan and  
12 implement disease control interventions.

13 Tsetse flies (genus *Glossina*) are obligate vectors of animal and human African  
14 trypanosomiasis (AAT and HAT, respectively). These diseases pose serious socioeconomic and  
15 health burdens to sub-Saharan Africa. In Kenya and Tanzania, HAT and AAT is transmitted most  
16 often by tsetse of the species *Glossina pallidipes*. Although there have only been a few cases of HAT  
17 reported recently in the study area (Franco et al., 2014; World Health Organization, 2020), both  
18 Kenya and Tanzania remain classified by the World Health Organization (WHO) as regions of HAT  
19 public health concern because of lack of control and surveillance activities (Franco et al., 2020). In  
20 contrast to HAT, AAT is widespread throughout the *G. pallidipes* range in Kenya and Tanzania.  
21 Previous empirical studies and mathematical modeling have indicated that *G. pallidipes* populations  
22 could be reduced to levels that minimize AAT transmission through vector control strategies such as  
23 bush clearance, ground spraying using insecticides, odor baited traps and insecticide impregnated  
24 targets (Bourn et al., 2001; Davis et al., 2011; Gilbert et al., 2016; Medlock et al., 2013; Ndeffo-Mbah  
25 et al., 2019; Pandey et al., 2015).

26 Vector control has been used to mitigate damage done by AAT and HAT in east Africa since  
27 the 1960s (Bourn et al., 2001). However, population rebounds in *G. pallidipes* are thought to  
28 jeopardize the long-term success of AAT control in the region (Ilemobade, 2009; Rogers et al., 1985).  
29 Insect survival outside of the treated areas and subsequent recolonization of treated areas are thought

30 to contribute to population rebounds (Bourn et al., 2001; Okeyo et al., 2017). Knowledge of the  
31 environmental factors associated with *G. pallidipes* survival and dispersal can improve our ability to  
32 predict where tsetse flies may be able to survive vector control campaigns and potential routes of  
33 recolonization. Tsetse flies are known to be sensitive to environmental conditions (Brightwell et al.,  
34 1992; Hargrove, 2009; Rogers & Randolph, 1991). Variables such as temperature and precipitation  
35 have been shown to affect birth rates, death rates, and development of tsetse flies (Hargrove, 2009),  
36 while temperature and humidity are known to affect dispersal distance (Brightwell et al., 1992).  
37 Understanding of survival and dispersal enables strategic planning that will reduce the risk of  
38 population rebounds, and thus vector re-emergence following control efforts.

39 Advances in spatial modeling and machine learning approaches have improved predictions of  
40 species distributions and dispersal patterns by integrating ecological and genetic data (Bouyer et al.,  
41 2015; Dicko et al., 2014; Hether et al., 2012; Hirzel et al., 2008; Manel et al., 2003; Pless et al., 2021).  
42 In particular, random forest regression, a widely used machine learning method, allows for modelling  
43 of nonlinear relationships across landscapes without overfitting (Liaw & Wiener, 2002; Prasad et al.,  
44 2006; Rehfeldt et al., 2006). These advantages enable the use of correlated variables and ecological  
45 data that violate parametric assumptions (Breiman, 2001; Garzón et al., 2006; Liaw & Wiener, 2002;  
46 Murphy et al., 2010; Wagner & Fortin, 2005), contributing to the feasibility of modelling complex  
47 landscape-level factors, such as habitat suitability and genetic connectivity in vectors (Pless et al.,  
48 2021).

49 In this paper, we take advantage of such recent methodological developments in spatial  
50 modeling to achieve two goals: To (i) build and integrate models of *G. pallidipes* habitat suitability  
51 and genetic connectivity across Kenya and northern Tanzania (Fig 1), and (ii) provide novel, spatially  
52 explicit vector control recommendations. We use field records and microsatellite genotypic data from  
53 published data (Bateta et al., 2020; Cecchi, 2002; Okeyo et al., 2017, 2018) with the addition of three  
54 new sampling sites. We developed our analysis strategy in collaboration with Pless et al. (2021) to  
55 enable both the identification of environmental correlates of vector habitat suitability and genetic  
56 connectivity (from here forward referred to simply as suitability and connectivity) and mapping of  
57 these predictions across the landscape. Additionally, we integrated outputs with a novel application of  
58 bivariate mapping to identify geographic regions with distinct risks and opportunities for *G. pallidipes*

vector control. Specifically, we provide vector control recommendations that consider predicted risks of population rebounds, corridors of recolonization, and isolated populations likely to be feasibly eradicated locally and/or used in the development of novel control strategies. Although methodology for predicting vector response to climate change, especially in predicting future connectivity, has not been fully developed, our study takes a first step by demonstrating feasibility of using basic environmental predictors available under climate change scenarios to predict suitability and connectivity. We do not extend this to projecting future suitability and connectivity because of challenges with validating predictions under novel conditions, and accounting for complex biological factors such as demography (Dormann, 2017; Urban et al., 2016; Yates et al., 2018). However, we do use climate change projections of the most important predictors in our models to identify geographic areas of high priority for monitoring for changes in tsetse fly presence and movement. Results indicate strong performance of our methodology, highlighting the utility of machine learning for informing current and future vector control across Kenya and Tanzania.

72

## 73 METHODS

### 74 *Glossina pallidipes* biology and distribution in the study area

75 *Glossina pallidipes* is a member of the *G. morsitans* group and is considered a savannah species. The distribution of *G. pallidipes* is limited to savannah habitat, and extends into Ethiopia in the north, the Democratic Republic of the Congo and Uganda in central Africa, Kenya, and Tanzania in central east Africa, and Mozambique and Zambia in southern east Africa (Ford, 1971; Jordan, 1993; Rogers & Randolph, 1985; Rogers & Robinson, 2004). However, the boundaries of savannah habitat mean that the continuous distribution of *G. pallidipes* does not extend into Ethiopia or Uganda, is limited within Kenya to areas south of Mt Kenya, and is limited within Tanzania to the Serengeti ecosystem and a band of habitat along the coast of the Indian Ocean (Ford, 1971; Jordan, 1993; Pollock, 1982; Rogers & Randolph, 1985; Rogers & Robinson, 2004; Cecchi et al., 2008; Ngari et al., 2020). Previous work has shown that for *G. pallidipes*, the tsetse fly belts recognized by the Kenya Tsetse and Trypanosomiasis Eradication Council (KENTTEC) are not necessarily ecologically or evolutionarily distinct. Instead, there is a weak genetic break of recent origin with current gene flow between the Lake Victoria Basin and the Serengeti ecosystem, and a strong biogeographic break

88 caused by the expansion of the Great Rift Valley in central Kenya (Faith et al., 2016; Lehmann, 1999;  
89 Linder et al., 2012; Wilfert, 2006; Wüster et al., 2007; Fig 1) that separates populations east and west  
90 of the valley. Thus, it was suggested by Bateta et al. (2020) that all populations east of the valley  
91 should be managed together. With this in mind, the biologically relevant geographic scope for  
92 management of *G. pallidipes* in Kenya extends from the Lake Victoria Basin at the border of Uganda  
93 and Kenya east to the Indian Ocean, and south to the edge of the Serengeti ecosystem in Tanzania.

94 *Glossina pallidipes* has a generation time of approximately five per year, has generally low  
95 dispersal rates of less than ~1 km per individual/generation (Bouyer et al., 2007; Cuisance et al.,  
96 1985; Rogers, 1977), and goes through population contractions during several arid periods of the year  
97 and expansions during rainy seasons (Camberlin & Wairoto, 1997; Devisser et al., 2010; Nnko et al.,  
98 2017; Pollock, 1982; Rogers & Randolph, 1985). These population fluctuations make it difficult to  
99 identify the extent of the distribution with trapping efforts, as a negative result does not necessarily  
100 mean low density at all times of year. These challenges have prompted extensive efforts by KENTEC  
101 and others to collect across multiple seasons and years for the full distribution of *G. pallidipes* in the  
102 region (Bateta et al., 2020; Cecchi et al., 2008; Ngari et al., 2020; Okeyo et al. 2017, 2018; Opiro et  
103 al., 2017). Nonetheless, copyright of much of the sampling efforts by the Kenyan government makes  
104 this data unavailable to the scientific community (Ngari et al., 2020), leaving urgent need for a  
105 publicly available up-to-date suitability model that is based on environmental conditions and is well  
106 integrated with knowledge of tsetse dispersal patterns.  
107

## 108 Summary of data inputs

109 *A1. Field-survey occurrence data and background points* – The field data were from trapping  
110 surveys carried out from 2015 to 2019 across Kenya and northern Tanzania (Bateta et al., 2020;  
111 Okeyo et al., 2017, 2018). Bi-conical and Ngu traps were placed in the field at sampling sites in  
112 clusters of 3-5 traps separated by less than 5 km, and were left out for either 24 or 48 hours. The  
113 sampling used in this study was from a concerted effort by our research group to comprehensively  
114 sample the *G. pallidipes* distribution in Kenya, as well as the connected habitat across political  
115 boundaries (i.e. Tanzania, as the *G. pallidipes* distribution does not extend continuously into Uganda;  
116 Pollock, 1982). There is also evidence that the sampling effort was comprehensive, as there were an

117 equal number of visited sites with no fly catches as those with fly catches that were within the  
118 expected distribution (Bateta et al., 2020). Locations of traps with flies in them were used as presence  
119 points in the suitability model (A3, Fig 2), and live flies were preserved in 80% ethanol for  
120 microsatellite genotyping. Instead of absence points, we used randomly selected “background” points  
121 to characterize the full range of environmental conditions. Background points allow the model to  
122 better distinguish the conditions under which species presence is more likely from the overall  
123 environmental conditions (Elith et al., 2006; Phillips et al., 2009). Use of background points at a  
124 sample size that matches presence points (in this case ~100 once converted to a 1x1 km grid raster)  
125 has been demonstrated to maximize accuracy in species distribution models (Barbet-Massin et al.,  
126 2012; Elith et al., 2006; Phillips et al., 2009). For background points we used 10 replicates of 100  
127 randomly sampled points across the geographic scope of our study (longitude of 33.7° to 42.5°,  
128 latitude of -4.8° to 5.0°, excluding ocean) using the R package “dismo” (Hijmans et al., 2017).

129 *A2. Microsatellite data* – A total of 659 individuals from 29 sampling sites were genotyped at  
130 11 microsatellite loci, with seven to 46 individuals per sampling site. Genetic data collection included  
131 18 sampling sites in Kenya and six sampling sites in northern Tanzania (~15 flies of each sex for each  
132 sampling site; A2, Fig 2). Of these, 600 flies from 21 sampling sites were genotyped by Bateta et al.  
133 (2020), and Okeyo et al. (2017, 2018). We added 84 flies from three new sampling sites (Fig 1) and  
134 genotyped them at the same 11 loci following the protocol described by Okeyo et al. (2017, 2018).  
135 Sampling sites containing traps more than two kilometers apart were split such that all traps within  
136 sampling sites are less than two kilometers from each other. We calculated pairwise Cavalli-Sforza  
137 and Edwards’ chord (CSE) genetic distance between sampling sites (A2, Fig 3S; Cavalli-Sforza &  
138 Edwards, 1967). CSE genetic distance has been shown to perform better than other genetic distance  
139 measures when there is missing data and when the relative distances between population pairs are  
140 being measured (Bouyer et al., 2015; Pless et al., 2021). To retain only the genetic distances that  
141 reflect contemporary environmental conditions rather than more ancient divergences such as those  
142 associated with the expansion of the Great Rift Valley (Faith et al., 2016; Lehmann et al., 1999;  
143 Linder et al., 2012; Wilfert et al., 2006; Wüster et al., 2007), we only included genetic distances  
144 between sampling sites within the two major genetic clusters east and west of the Great Rift Valley

145 that were identified in previous studies (Bateta et al., 2020; Okeyo et al., 2018) and confirmed here  
146 with DAPC (File 1S; Jombart et al., 2008).

147       *A3. Remotely-sensed environmental data* – Predictor variables for both the suitability and  
148 connectivity models were based on 1-kilometer resolution environmental raster layers of 19  
149 bioclimatic variables, slope, altitude, and river density (A3, Fig 2). Although including more predictor  
150 variables (e.g. host availability, landcover) may have potential to improve the model, we chose to  
151 limit our selection to variables that are either unchanging on relevant timescales of decades and  
152 centuries (i.e. slope, altitude, and river location), or are publicly available as forecasts under four  
153 different emissions scenarios based on 36 different multiple climate change scenarios (Karger et al.,  
154 2017, i.e. 19 climatic variables reflecting temperature and precipitation, i.e. temperature and  
155 precipitation based climatic variables) allow us to visualize predicted change in climate variables  
156 important in our models.

157       The 19 bioclimatic variables were temperature- and precipitation- based (Table 1S), and were  
158 calculated from raster files downloaded from Climatologies at High Resolution for the Earth’s Land  
159 Surface Areas (CHELSA; Karger et al., 2017) for the time span of 2008-2013 with the R package  
160 “dismo” (Hijmans et al., 2017). We used seasonal bioclimatic variables based on the precipitation  
161 seasonality trends observed in the study area, rather than the default quarterly estimates, to more  
162 accurately capture the seasonal variation relevant to the ecology of the region (Table 1S, Fig 1S).  
163 Slope and altitude raster files were downloaded from Geomorpho90m dataset (Amatulli et al., 2020)  
164 and Multi-Error-Removed Improved-Terrain (Yamazaki et al., 2017), respectively. Following  
165 methods described in Pless et al. (2021), we created a river density layer in the R package  
166 “KernSmooth” (Wand, 2015) based on river shapefiles downloaded from DIVA-GIS (March 2020;  
167 <http://www.diva-gis.org>). The final raster layers were clipped to the extent of Kenya and northern  
168 Tanzania (longitude of 33.7° to 42.5°, latitude of -4.8° to 5.0°) and projected to the WGS-84  
169 coordinate reference system in the R package “rgdal” (Bivand et al., 2019).

170       All spatial data, including the environmental inputs and results from the models (see below),  
171 were visualized using the R packages “raster” (Hijmans, 2019), “rgdal” (Bivand et al., 2019), “rgeos”  
172 (Bivand & Rundel, 2020), and “ggplot2” (Wickham, 2016), and figures were produced using R

173 packages “ggpubr” (Kassambara, 2019), “gridExtra” (Auguie, 2017), “patchwork” (Pedersen, 2020),  
174 and “ggrepel” (Slowikowski, 2020).

175

## 176 **Random forest model of habitat suitability**

177 *B1. Environmental point values* - For the suitability model we used environmental values  
178 extracted at the coordinates of the presence (n = 349 trap locations) and background points (n = 100  
179 per model replicate) for the 22 environmental variables using the R package “raster” (Hijmans, 2019).

180 *C1. Building and projecting the RF model* – Following methods described in Hill et al. (2017),  
181 we built, evaluated, and projected our suitability model with the R packages “biomod2” (Thuiller et  
182 al., 2019), “raster” (Hijmans, 2019), “sp” (Pebesma & Bivand, 2005), and “rgdal” (Bivand et al.,  
183 2019) using presence/background scored as 1/0, respectively, as the response variable, and 22  
184 environmental values extracted at these coordinates as the explanatory variables (B3, A3, B4, Fig 2).  
185 We treat the binary (1/0) data as a continuous response variable (i.e. ran a regression model) in order  
186 to end up with a continuous measure of suitability. Hence, we assessed model performance with the  
187 R-squared generated internally by the random forest algorithm, which is based on a bootstrapping  
188 procedure that repeatedly selects a random sample (with replacement) of training sets and compares  
189 the average predictions with the testing sets that were left out of the model (Breiman, 2001; Liaw &  
190 Wiener, 2002). We evaluated variable importance using increase in node purity, which is calculated  
191 by taking the decrease in the Residual Sum of Squares (RSS) as the result of splitting on each variable  
192 and averaging it across all trees (Liaw & Wiener, 2002). We choose to evaluate variable importance  
193 in this way rather than using percent increase in Mean Square Error from permuting each variable  
194 (another evaluation option provided by random forest) because increase in node purity is not sensitive  
195 to correlation between variables. To evaluate model performance, we used a 10-fold cross-validation  
196 procedure and calculated the true skill statistic (TSS) and the area under the receiver operating curve  
197 (AUC) (Allouche et al., 2006).

198

## 199 **Combining suitability output with previous models**

200 The existing suitability map available for *G. pallidipes* in eastern Africa (Cecchi, 2002; Cecchi  
201 et al., 2008) needed to be updated because it was based on trapping records that were more than 15

202 years old and had obvious inaccuracies. The most notable inaccuracy is the prediction of low  
203 suitability in the Serengeti ecosystem, a region known to harbor *G. pallidipes* and that had high  
204 capture rates in trapping records used in this study. However, the raw data is property of the  
205 Government of Kenya (Kenya Tsetse and Trypanosomosis Eradication Council), and we have not  
206 been granted access (Cecchi, 2002; Cecchi et al., 2008; Ngari et al., 2020). Thus, instead of building a  
207 comprehensive model, as would have been our preference, we combined our map with the existing  
208 map. We combined the maps by taking the maximum predicted suitability for each pixel from the two  
209 maps, the most conservative way possible given that for vector control, it is better to over-predict than  
210 under-predict vector presence.

211

## 212 Random forest model of genetic connectivity

213 *B2. Environmental path data and geographic distance* - We extracted the median value along  
214 straight paths ( $n = 198$  paths) between sampling sites ( $n = 29$  sampling sites) within genetic clusters  
215 for each of the 22 environmental variables (B3, Fig 2) using the R package “raster” (Hijmans, 2019).  
216 We chose to use the median value as opposed to the mean because it is not as affected by the presence  
217 of outliers. We included two additional explanatory variables, (i) mean kernel density of sampling  
218 effort and (ii) geographic distance to ensure our model accounted for spatial auto-correlation (File 1S;  
219 Shi et al., 2019; Souris et al., 2019). We created a sampling density layer in the R package  
220 “KernSmooth” (Wand, 2015; File 1S) and estimated the median value along the 198 straight paths, as  
221 was done for the 22 environmental variables. Geographic distance was estimated following Bouyer et  
222 al. (2015) by creating a uniform raster (all 1x1 km pixels were assigned a value of 1), and summing  
223 values along the 198 straight paths.

224 The inclusion of these variables was necessary because spatial auto-correlation is an almost  
225 ubiquitous confounding factor in landscape-level studies. Auto-correlation is especially pronounced in  
226 population genetic studies because genetic distance is expected to be correlated with geographic  
227 distance under neutral conditions (Rousset et al., 1997; Wright, 1943). This was of particular concern  
228 in this study because genetic and geographic distance were reported to be correlated in some subsets  
229 of this dataset (Bateta et al., 2020), a result we confirmed with Mantel tests (File 1S; Mantel, 1967;  
230 Dray & Dufour, 2007). Nonetheless, we think that the spatial modeling approach implemented is

231 appropriate because we were able to demonstrate with Anderson-Darling k-means tests (Scholz &  
232 Zhu, 2019) that the majority of variation in genetic distance remained unexplained in models that  
233 considered geographic distance alone (File 1S).

234 *C2. Building and projecting the connectivity model* – Our connectivity model was built with  
235 the full dataset (29 sampling sites, 198 Paths) using the packages “randomForest” (Liaw & Wiener,  
236 2002), “raster” (Hijmans, 2019), “spatstat” (Baddeley et al., 2005), and “sp” (Pebesma & Bivand,  
237 2005). We built a random forest model using CSE genetic distance between sampling site pairs as a  
238 proxy for connectivity (B3, C2, Fig 2). This model was projected across Kenya and Northern  
239 Tanzania to create a map of predicted connectivity using the environmental data and sample density  
240 rasters, as well as the raster with uniform values of 1 used to estimate geographic distance following  
241 Bouyer et al. (2015). This procedure essentially assigned the geographic distance between each pixel  
242 and itself to 1 km in the projections of the model. As in the suitability model, we assessed model  
243 performance with the internally generated R-squared and variable importance using increase in node  
244 purity.

245 *C2a. Model evaluation* – To allow for evaluation of the connectivity model’s performance in  
246 different subsets of the data, we used leave-one-out cross-validation. For each run of the cross-  
247 validation the Root Mean Square Error (RMSE) was calculated based on testing data not included in  
248 the training of the model. We assessed the accuracy of our models by generating a null distribution of  
249 100 RMSE values (i.e. values expected by chance for this type of modeling) from models trained on  
250 randomly shuffled data, and used this as a benchmark against which to compare our observed RMSE  
251 distribution using Welch’s t-tests (File 1S).

252 *C2b. Spatial evaluation* – We estimated the accuracy of the projections for each run of the  
253 leave-one-out cross validation by extracting the median CSE genetic distance along straight paths  
254 between sampling sites from the testing data. Comparing these spatially predicted CSE values to the  
255 observed CSE values allowed us to estimate RMSE values that reflected the accuracy of the projected  
256 connectivity map. As we did for the model evaluation, we compared the observed spatial RMSE  
257 values to null distributions generated with shuffled data (see paragraph above, File 1S).

258

## 259 **Integrating and interpreting outputs to inform vector control**

260           C3. *Integrating habitat suitability and genetic connectivity models* – We created a bivariate  
261 map of predicted suitability and connectivity (C3, Fig 2; File 2S) using R packages “raster” (Hijmans,  
262 2019), “rgdal” (Bivand et al., 2019), “classInt” (Bivand, 2018), “XML” (Lang et al., 2019). We  
263 masked all probability of presence values less than ten percent in the suitability model projection such  
264 that comparisons were not made where tsetse flies were expected to be absent. More information  
265 about the creation of this bivariate map can be found in File 1S and File 2S.

266           C4a. *Post-hoc visualization of local correlations* – The random forest approach we use in this  
267 study has several advantages over other standard modelling approaches, such as simple linear  
268 regression, including greater flexibility and higher predictive power when modeling complex, non-  
269 linear relationships (File 1S). However, as is the case with many machine learning methods, the trade-  
270 off for this superior performance is more complexity and less interpretability. Thus, to gain a better  
271 understanding of the environmental drivers of suitability and connectivity, we used the corLocal()  
272 function in the R package “raster” (Hijmans, 2019) to calculate the Pearson’s correlation coefficient  
273 between projections of the response variables of interest (i.e. suitability or connectivity (1 –  
274 scaled genetic distance)) and the top predictor variables identified by our random forest  
275 models.

276           C4b. *Post-hoc visualization of predicted environmental change* – Global warming is expected  
277 to affect tsetse fly distribution and connectivity (Bourn et al., 2001), making knowledge of the  
278 environmental drivers of tsetse fly distribution and connectivity under current and future conditions a  
279 valuable part of planning vector control strategy. For short term planning, the bivariate maps we built  
280 can provide specific vector control recommendations for different categories of landscape in Kenya  
281 and northern Tanzania (see above). Long term planning is more difficult and is influenced by more  
282 uncertainties. Although it would be ideal to project our models under future conditions, the  
283 methodology for this is not fully developed. There are outstanding challenges in transferring models  
284 to novel conditions, such as accounting for the effects of biological mechanisms (i.e. demography,  
285 species interactions, and evolutionary change), quantifying uncertainty, and assessing transferability  
286 (Dormann, 2017; Urban et al., 2016; Yates et al., 2018). Instead, we take an alternative approach that  
287 avoids unrealistic assumptions about the effects of biological mechanisms as well as problems with  
288 model validation and transferability: We provide estimates of predicted change in the most important

289 environmental variables from our models of *G. pallidipes* suitability and connectivity. In this way, our  
290 approach informs which geographic regions will experience environmental change that may affect *G.*  
291 *pallidipes* vectoring capacity, and we interpret these as the regions that should be monitored for  
292 changes in vector presence and dispersal. Even though we cannot presently define the magnitude or  
293 direction of future changes in connectivity and suitability given the limitations of our data and  
294 models, knowing where to expect relevant environmental change could be used to optimize future  
295 monitoring efforts. We estimated the predicted change of the most important environmental variables  
296 from the suitability and connectivity models under the NASA RCP 4.5 climate change model for  
297 2041-2060, calculated by subtracting the present environmental layer (an average across 2008-2013)  
298 from the future environmental layer. Both present and future environmental layers for each variable  
299 were sourced from CHELSA (Karger et al., 2017).

300

## 301 RESULTS

### 302 Habitat suitability model

303 *Full model results* - The mean R-squared for the 10 suitability models built using all presence  
304 points and each of the 10 sets of background points, was 0.80 (SD = 0.02), indicating that on average  
305 80% of the variance in suitability was explained by the predictor variables. The most important  
306 variable for six of the 10 models, based on the increase in node purity, was the maximum temperature  
307 of the warmest month (Fig 5A, Fig 6SA), and for the remaining four models the most important  
308 variable was the temperature annual range (Fig 5A, Fig 6SA). These variables suggest that  
309 temperature was the most predictive climatic variable of *G. pallidipes* presence in tsetse fly traps.

310 *Model evaluation* - The random forest suitability models demonstrated high accuracy across  
311 all 10 folds of the cross-validation and all 10 sets of randomly selected background points. The mean  
312 AUC of all sets and folds was 0.99 (SD = 0.01) and the AUC never fell below 0.92, indicating an  
313 overall favorable ratio between sensitivity (low false negatives) and specificity (low false positives)  
314 across all thresholds. The mean true skill statistic (TSS) of all sets and folds was 0.96 (SD = 0.02) and  
315 the TSS never fell below 0.80, indicating that the models were both sensitive and specific when  
316 discerning presence and absence points based on the threshold that optimizes the TSS as determined  
317 in “biomod2” (Thuiller et al., 2019).

318

319 **Genetic connectivity model**

320       *Full model results* – The full model of connectivity (Fig 5SB) performed well with a R-squared of 0.67, indicating that on average 67% of the variance in genetic distance was explained by the predictor variables. Results from the increase in node purity analysis indicated that precipitation of the driest season was the most important variable in the final model of connectivity (Fig 5B, Fig 6SB). Increase in node purity measures how well the variable of interest can be used to split the data, suggesting that precipitation may be an important environmental driver of tsetse fly movement and/or survival and reproduction after relocating.

327       *Model evaluation* – The mean RMSE from the leave-one-out cross validation was 0.07 (SD = 0.03) across all 29 runs (all 29 sampling sites; Fig 3A). The mean RMSE for testing sampling sites from the east was 0.06 (SD = 0.03) and from the west was 0.08 (SD = 0.02) and this difference was not significant ( $t(20.799) = -1.18$ ,  $p = 0.25$ ). Based on t-tests, the RMSE values from our model were significantly lower ( $p\text{-value} < 0.05$ ) than the RMSE values from the null models (mean = 0.11, SD = 0.02; File S1).

333       *Spatial evaluation* – Spatial evaluations were calculated by comparing the median genetic distances from straight paths between sampling sites along the projected model surface to the observed genetic distances between sampling sites. The mean RMSE from the spatial evaluation of the model projections was 0.08 (SD = 0.03) across all 29 leave-one-out cross-validation runs (Fig 3B). The mean spatial RMSE for testing sampling sites from the east was 0.07 (SD = 0.03) and from the west was 0.09 (SD = 0.03), but this difference was not significant ( $t(24.261) = -2.04$ ,  $p = 0.05$ ). Based on t-tests, the spatial RMSE values from our model were significantly lower ( $p\text{-value} < 0.05$ ) than the spatial RMSE values from the null models (mean = 0.11, SD = 0.02; File S1).

341

342 **Integrating and interpreting outputs to inform vector control**

343       *Integrating habitat suitability and genetic connectivity* – The bivariate map of the final 344 suitability and connectivity models, showed heterogeneous spatial patterns in suitability and 345 connectivity (Fig 4). Low suitability was predicted primarily in the Chalbi desert (Fig 1) and around 346 the center of the Great Rift Valley in Kenya (Fig 4A). Regions of high connectivity and high

347 suitability included the northeastern part of Tanzania (around the Serengeti area), central Kenya  
348 (along the Aberdare mountain range, Fig 1), and a small section of the eastern coast of Kenya (Fig  
349 4C). In Kenya, the southern tip (extending into Tanzania) and the area to the west of the Great Rift  
350 Valley (around Lake Victoria, Fig 1) had high predicted suitability, but low connectivity (Fig 4C).

351 *Post-hoc visualization of local correlations* – The maps of the Pearson's correlations between  
352 the most important predictor variables and the response variables (i.e. suitability and connectivity,  
353 respectively) showed spatial variation in the direction and magnitude of associations (Fig 5C). The  
354 correlation between maximum temperature of the warmest month (i.e. the most important variable  
355 from the suitability model) and suitability was generally positive in the eastern part of Kenya, around  
356 the Lake Victoria basin and following the Great Rift Valley into Tanzania (Fig 5C). In the western  
357 part of Kenya, the spatial pattern of correlation was much more patchy, with interspersed areas of  
358 positive and negative associations (Fig 5C). The map of correlation between precipitation of the driest  
359 season (i.e. the most important variable from the connectivity model) and connectivity had positive  
360 patches in eastern Kenya, primarily along rivers, as well as around the Serengeti (Fig 5C).  
361 Precipitation of the driest season had a strong, negative correlation with connectivity around the Lake  
362 Victoria basin in western Kenya (Fig 5C).

363 *Post-hoc visualization of predicted environmental change* – To inform understanding of the  
364 impact of climate change on *G. pallidipes* connectivity and suitability, we estimated the predicted  
365 change over the next 20-40 years (NASA RCP 4.5 climate change model for 2041-2060) of the most  
366 important variables from our models (Fig 5, Fig 6S). Predicted change in the maximum temperature  
367 of the warmest month, the most important variable from the suitability model, indicated that changes  
368 in temperature are expected across most of Kenya, with a general increase in temperature in the north  
369 and a decrease in temperature in the south and coastal habitats (Fig 5C). Precipitation of the driest  
370 season, the most important variable from the connectivity model, is predicted to change fairly  
371 homogeneously across the landscape (Fig 5C). A notable deviation from this uniform change is a  
372 concentrated patch of predicted decreased precipitation along the eastern shore of Lake Victoria  
373 (southwest corner of Kenya; Fig 5C).

374

## 375 **DISCUSSION**

376        The goals of this paper were to: (i) Build and integrate models of *G. pallidipes* suitability and  
377 connectivity, and (ii) provide spatially explicit vector control recommendations. Both our models  
378 demonstrated strong performance, and were able to explain a large portion of the variance in  
379 suitability and connectivity. Bivariate maps of suitability and connectivity provide evidence that these  
380 factors vary independently across the landscape, and indicate that the Serengeti comprises an area of  
381 high suitability and high connectivity while both the Lake Victoria basin and southeastern Kenya  
382 constitute areas of high suitability and low connectivity. These results suggest that vector control  
383 campaigns are likely to be less successful in the Serengeti, and more successful in the Lake Victoria  
384 basin and southeastern Kenya. We further recommend that future monitoring efforts should focus on  
385 tracking potential changes in vector presence and dispersal around the Serengeti and the Lake Victoria  
386 basin based on projected local climatic shifts.

387

### 388        **Habitat suitability model**

389        We were able to explain approximately 80% of the variance in suitability with our suitability  
390 model, which also demonstrated strong performance based on the 10-fold cross-validation for each of  
391 the 10 background point replicates. The standard evaluation statistics were close to the best score  
392 possible of one (AUC = 0.99, and TSS = 0.96), indicating that the models accurately predicted the  
393 testing data during cross-validation. The suitability model predicted a patchy distribution of habitat  
394 concentrated in the southeast of Kenya and around the Lake Victoria basin. There is a possibility that  
395 the model was overfit to our sampling locations, so to be as conservative as possible we combined our  
396 final suitability model with the existing FAO model (Cecchi, 2002). The existing FAO model was  
397 built from data collected before 2002, making it out of date, and also shows indications of overfitting  
398 since there was a gap in sampling that coincided with low predicted suitability in the Serengeti  
399 ecosystem despite this region being known to harbor tsetse flies (Cecchi, 2002). Although the best  
400 solution to this problem would have been to include all known presence points from both data sources  
401 in this study, this was not possible because of copyright restrictions (Cecchi, 2002; Ngari et al., 2020),  
402 so we combined the models to err on the side of over-predicting vector presence.

403        The most important variable based on increase in node purity, a random forest variable  
404 importance measure, was maximum temperature of the warmest month (Fig 5A, Fig 6SA). Based on

405 the map of local correlations, maximum temperature of the warmest season generally had a positive  
406 effect on suitability across Kenya and Tanzania (Fig 5C). Temperature is known to affect tsetse fly  
407 birth rates, mortality, and development (Brightwell et al., 1992; Hargrove, 2009), suggesting that  
408 thermal tolerance may be an important driver of *G. pallidipes* habitat use.

409

#### 410 **Genetic connectivity model**

411 The final random forest model of connectivity explained 67% of the variance in genetic  
412 distance and performed well based on both direct evaluation of the model predictions and spatial  
413 evaluation of the projected map (Fig 3). There were no notable differences in model performance  
414 between the two genetic clusters. Two sampling sites (SHT in the east and NGU in the west) had  
415 substantially high error values in comparison to the other sites and the null values (Fig 3; File 1S).  
416 The site in the east (SHT) was an outlier in the genetic distance distribution from the east. These  
417 differences are likely the result of the smaller sampling size for this sampling site ( $n = 7$ ) compared to  
418 the average sampling size of 23 individuals. The site in the west (NGU) may have low accuracy  
419 because it's assignment to the eastern genetic lineage was not fully supported in all analyses (Bateta et  
420 al., 2020), implying that genetic divergence from current landscape features could have been masked  
421 by the stronger signal of divergence from past vicariance events (i.e. expansion of the Great Rift  
422 Valley ~2-5 mya; Faith et al., 2016; Lehmann et al., 1999; Linder et al., 2012; Wilfert et al., 2006;  
423 Wüster et al., 2007).

424 The most important variable for the connectivity model was precipitation of the driest season  
425 (Fig 5B, Fig 6SB). While it is not possible to discern direct causal relationships between  
426 environmental variables and connectivity using this methodology, the importance of precipitation of  
427 the driest season may be related to the sensitivity of tsetse fly immature life stages to desiccation  
428 (Hargrove, 2009). The risk of desiccation in immature stages may limit successful offspring survival  
429 until reproduction in migrants. If true, this suggests that migration often occurs over several  
430 generations along corridors of high connectivity. This suggestion has been made to explain the much  
431 longer migration distances retrieved in genetic studies that consider several generations than  
432 migration distances found in ecological field studies that track a single individual (Bateta et al., 2020;  
433 Okeyo et al., 2018; Opiro et al., 2017).

434        The local correlations between precipitation of the driest season and connectivity exhibit  
435 variation spatially (Fig 5C). In the west, connectivity generally has a negative association with  
436 precipitation during the driest season, especially around the Lake Victoria basin and parts of the Great  
437 Rift Valley (Fig 5C). One possible explanation for this negative association is that flies have to  
438 migrate further to find water in regions where there is low precipitation during the dry season,  
439 however it is not possible to distinguish causality using these models.

440        In eastern Kenya and parts of Tanzania there are several discontinuous regions, primarily  
441 along rivers and part of the Great Rift Valley, where higher connectivity is associated with higher  
442 precipitation during the driest season. This difference in the direction of the correlation between  
443 connectivity and precipitation suggests that the ecological mechanisms affecting connectivity may  
444 vary across Kenya and Tanzania. Adaptive differences between populations could also play a role in  
445 establishing different associations between connectivity and climatic variables, something that could  
446 be explored in the future using landscape genomics methods to identify adaptive variation in *G.*  
447 *pallidipes* associated with climatic variables such as temperature and precipitation. Although  
448 valuable, this is outside of the goals of this paper since the microsatellites used target neutral genetic  
449 variation.

450

#### 451        **Integrating habitat suitability and genetic connectivity models**

452        The bivariate map indicates that suitability and connectivity (Fig 4) are not strongly correlated  
453 with each other. A large fraction of the study area with high predicted suitability has low predicted  
454 connectivity (blue, Fig 4), contradicting the expectation from landscape ecology that suitability  
455 facilitates connectivity (Zeller et al., 2012). This may be due to the limitations of the habitat  
456 suitability model, which only takes into account abiotic factors (e.g. ignores ecological interactions)  
457 and may overpredict suitability (Broennimann et al., 2012; De Araújo et al., 2014). However, it is also  
458 possible that the pattern we observe reflects the biological reality that suitability does not always  
459 facilitate connectivity in this system and that different ecological constraints are responsible for  
460 shaping habitat use and connectivity in *G. pallidipes*. For example, habitat use may be more strongly  
461 influenced by the risk of thermal stress while migration over multiple generations that results in gene  
462 flow may be more strongly influenced by the risk of desiccation in juveniles.

463        Regardless of the mechanisms controlling heterogeneity in suitability and connectivity, the  
464 bivariate map can be used to identify three categories of landscape that will likely require different  
465 vector control strategies: areas of (a) high connectivity and high suitability (red, Fig 4), (b) high  
466 connectivity and low suitability (yellow, Fig 4), and (c) low connectivity and high suitability (blue,  
467 Fig 4).

468        Areas of (a) high connectivity and high suitability are found primarily in patches centered in  
469 the Serengeti ecosystem and central Kenya (Fig 1, Fig 4). Our models suggest that these regions could  
470 support healthy tsetse populations with high dispersal. High recolonization potential within these  
471 regions could render internal control efforts ineffective. Instead, it may be more effective to focus on  
472 isolating these areas from neighboring habitat by establishing vector control along their perimeters.

473        Areas of (b) high connectivity and low suitability are found at the margins of the *G. pallidipes*  
474 distribution (Fig 4). Our models suggest that these regions support high dispersal and could facilitate  
475 reinvasion and seasonal migration. Although these areas may not support year-round tsetse  
476 populations that require targeted treatment, they could act as dispersal corridors. Knowledge of these  
477 dispersal corridors can help identify areas with low risk of reinvasion when planning spatially explicit  
478 eradication efforts, and can also inform placement of treatment technology to block dispersal from  
479 outside areas.

480        Areas of (c) low connectivity and high suitability are found in two large patches, one in  
481 western Kenya in the Lake Victoria basin (Fig 4), and another in southeastern Kenya (Fig 4). Our  
482 models suggest that these regions could support large tsetse populations, but that there is low  
483 connectivity so these populations are therefore likely to be isolated. The presence of isolated  
484 populations in these regions could present an opportunity for testing of novel vector control methods  
485 as well as local eradication of tsetse flies. The identification of isolated tsetse fly populations using  
486 suitability modeling and population genetics has been previously used to plan successful vector  
487 control efforts in Senegal that lead to the local eradication of tsetse flies opening new areas for  
488 agriculture (Dicko et al., 2014; Solano et al., 2010).

489

## 490 Applications to vector control

491        Results from the bivariate map can be used to provide regionally-specific recommendations  
492 for vector control. In the west, there is a noticeable divide between the region of high suitability and  
493 low connectivity in the Lake Victoria basin (Fig 4) and the region of high suitability and high  
494 connectivity within the serengeti ecosystem. This suggests an effective vector control strategy could  
495 be a “rolling carpet” approach, moving from the western part of Kenya towards the Serengeti to  
496 minimize re-invasions. Vector control in the west is particularly important because this region  
497 includes a tsetse belt that has been found to have high rates of AAT infection in cattle in addition to a  
498 significantly high prevalence of AAT related disability in human populations (Grady et al., 2011). In  
499 the east, a large area of low connectivity and high suitability overlaps with three KENTECC identified  
500 tsetse belts (the Mbeere-Meru fly belt, the Central Kenya fly belt, and the Coastal fly belt). Bateta et  
501 al. (2020) argued that the eastern belts should be treated as one *G. pallidipes* population based on the  
502 results of their population genetic analysis. Our modeling approach detected continuous highly  
503 suitable habitat with no notable breaks in connectivity in these eastern belts, thus generally supporting  
504 the conclusion of Bateta et al. (2020) that the eastern belts should be managed as a single unit.

505        Results from our post-hoc analysis can also be applied to future vector control planning. Post-  
506 hoc analysis from the suitability model indicates that the top predictor, temperature of the warmest  
507 month, is projected to change the most in north central Kenya (north of the Tana River), and northern  
508 Tanzania in the Serengeti region (Fig 5C). We suggest that these regions should be monitored for  
509 changes in tsetse fly presence and abundance (Fig 4) to provide early warning if there are increases in  
510 tsetse fly abundance that could extend the region impacted by AAT. For example, a useful  
511 experimental approach could be to set up traps along the perimeters of these regions (e.g. along the  
512 Serengeti National Park boundaries in Tanzania and range limits north of the Tana River in Kenya)  
513 and monitor annually for changes in tsetse fly density approximated by the number of flies caught in  
514 traps using a standard trapping protocol (e.g. those of Bateta et al., 2020; Okeyo et al., 2017, 2018).

515        Post-hoc analysis from the connectivity model indicates that the top predictor, precipitation of  
516 the driest season, is expected to change uniformly across Kenya (Fig 5C). An exception occurs in a  
517 discrete patch along the eastern shore of Lake Victoria (southwest corner of Kenya) which is expected  
518 to experience a substantial decrease in precipitation (Fig 5C). We recommend that future studies are  
519 designed to detect changes in connectivity across this patch to provide early warning of increased risk

520 of HAT spreading from the Uganda/Kenya border where the most recent HAT cases were detected  
521 (World Health Organization, 2020). Alternatively, a decrease in connectivity over time could present  
522 an opportunity to efficiently fortify the barrier to HAT spread eastward with minimal vector control  
523 effort. A useful experimental set up in this case would be to place traps throughout the region  
524 bounded by the Nzoia river, the eastern shore of Lake Victoria, and the Great Rift Valley (east of the  
525 Uganda/Kenya border), an area which has not been well sampled in this or previous studies (Figure 1;  
526 Bateta et al., 2020; Okeyo et al., 2017, 2018; Ouma et al., 2006). Time series samples should be  
527 collected from the same georeferenced localities every 5 years to monitor for changes in dispersal  
528 patterns, approximated by changes in genetic distance and population structure. Previous studies have  
529 documented temporal genetic differentiation in *G. pallidipes* in eastern Africa at this time scale  
530 (Okeyo et al., 2017).

531 Finally, although we did not directly forecast suitability and connectivity in this study, our  
532 results represent a first-step towards this goal. Our models, built using only environmental predictors  
533 that are available for 36 different climate change models under four different emissions scenarios  
534 (Karger et al., 2017), or are expected to remain constant in the future (e.g. slope and altitude),  
535 performed very well, suggesting that these variables can, at least in theory, provide enough  
536 environmental information to allow for projections of both suitability and connectivity models under  
537 climate change. However, we refrain from projecting our models in this study due to our current  
538 inability to validate projections through time and perform adequate sensitivity analyses to explore  
539 how robust our predictions would be to uncertainty in the climate projections. As new data and  
540 methods become available, we plan to build on these results and use future projections to evaluate  
541 climate change risks impacting the spread of AAT and HAT by tsetse flies.

542

### 543 Conclusion

544 We identified regions that may host resilient tsetse fly populations, potential routes of  
545 recolonization, and candidate isolated locations for local eradication and/or development of novel  
546 vector control strategies. Our findings suggest that our machine learning approach can accurately  
547 predict tsetse habitat use and connectivity, and has great potential to improve understanding of animal  
548 habitat use and movement in a changing climate. In this study, our choice of environmental variables

549 that are available as future projections are a first step towards making climate change projections. In  
550 this study, we did not make future projections of suitability and connectivity because of the  
551 unresolved challenges of transferring models to novel future climatic conditions (Dormann, 2017;  
552 Urban et al., 2016; Yates et al., 2018). Future studies should work towards developing and evaluating  
553 such projections of suitability and connectivity with respect to the uncertainty of climate change  
554 forecasts. Beyond utility for vector control for AAT and HAT in Kenya and Tanzania, the methods  
555 we develop can inform management of biological resources in a variety of contexts, from the control  
556 of unwanted species to the conservation of threatened and endangered biodiversity.

557

## 558 **Acknowledgements**

559 This work was funded by the Foundation for the National Institutes of Health  
560 (<https://fnih.org/>) grant number U01 AI115648, awarded to Serap Aksoy and Adalgisa Caccone, the  
561 Foundation for the National Institutes of Health (<https://fnih.org/>) Fogarty Global Infectious Diseases  
562 Training Grant number D43TW007391, awarded to Serap Aksoy, and the Rosenfeld Science Scholars  
563 Program Fellowship, awarded to Anusha Bishop.

564

## 565 **Data Archiving Statement**

566 All data for this study including tsetse fly genotypes, tsetse fly trapping localities, and  
567 landscape/environmental parameters are available at the Dryad Digital Repository:  
568 <https://doi.org/10.6078/D1B715>.

569    Literature Cited

- 570    Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models:  
571    Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*.  
572    <https://doi.org/10.1111/j.1365-2664.2006.01214.x>
- 573    Amatulli, G., McInerney, D., Sethi, T., Strobl, P., & Domisch, S. (2020). Geomorpho90m, empirical  
574    evaluation and accuracy assessment of global high-resolution geomorphometric layers. *Scientific  
575    Data*, 7(1), 162. <https://doi.org/10.1038/s41597-020-0479-6>
- 576    Auguie, B. (2017). gridExtra: Miscellaneous Functions for “Grid” Graphics. Retrieved from  
577    <https://cran.r-project.org/package=gridExtra>
- 578    Baddeley, A., & Turner, R. (2005). {spatstat}: An {R} Package for Analyzing Spatial Point Patterns.  
579    *Journal of Statistical Software*, 12(6), 1–42. Retrieved from <http://www.jstatsoft.org/v12/i06/>
- 580    Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for  
581    species distribution models: How, where and how many? *Methods in Ecology and Evolution*.  
582    <https://doi.org/10.1111/j.2041-210X.2011.00172.x>
- 583    Bateta, R., Saarman, N. P., Okeyo, W. A., Dion, K., Johnson, T., Mireji, P. O., ... Caccone, A.  
584    (2020). Phylogeography and population structure of the tsetse fly *Glossina pallidipes* in Kenya  
585    and the Serengeti ecosystem. *PLoS Neglected Tropical Diseases*, 14(2), 1–26.  
586    <https://doi.org/10.1371/journal.pntd.0007855>
- 587    Bivand, R. (2018). classInt: Choose Univariate Class Intervals. Retrieved from <https://cran.r-project.org/package=classInt>
- 588
- 589    Bivand, R., Keitt, T., & Rowlingson, B. (2019). rgdal: Bindings for the “Geospatial” Data Abstraction  
590    Library. Retrieved from <https://cran.r-project.org/package=rgdal>
- 591    Bivand, R., & Rundel, C. (2020). rgeos: Interface to Geometry Engine - Open Source ('GEOS').  
592    Retrieved from <https://cran.r-project.org/package=rgeos>

- 593 Bourn, D., Reid, R., Rogers, D., Snow, B., & Wint, W. (2001). *Environmental change and the*  
594 *autonomous control of tsetse and trypanosomosis in sub-Saharan Africa: case histories from*  
595 *Ethiopia, The Gambia, Kenya, Nigeria and Zimbabwe.*
- 596 Bouyer, J., Ravel, S., Dujardin, J. P., De Meeus, T., Vial, L., Thévenon, S., ... Solano, P. (2007).  
597 Population structuring of *Glossina palpalis gambiensis* (Diptera: Glossinidae) according to  
598 landscape fragmentation in the Mouhoun River, Burkina Faso. *Journal of Medical Entomology*.  
599 [https://doi.org/10.1603/0022-2585\(2007\)44\[788:PSOGPG\]2.0.CO;2](https://doi.org/10.1603/0022-2585(2007)44[788:PSOGPG]2.0.CO;2)
- 600 Bouyer, J., Dicko, A. H., Cecchi, G., Ravel, S., Guerrini, L., Solano, P., ... Lancelot, R. (2015).  
601 Mapping landscape friction to locate isolated tsetse populations that are candidates for  
602 elimination. *Proceedings of the National Academy of Sciences*, 112(47), 14575 LP-14580.  
603 <https://doi.org/10.1073/pnas.1516778112>
- 604 Breiman, L. (2001). Random forests. *Machine Learning*. <https://doi.org/10.1023/A:1010933404324>
- 605 Brightwell, R., Dransfield, R. D., & Williams, B. G. (1992). Factors affecting seasonal dispersal of  
606 the tsetse flies *Glossina pallidipes* and *G. longipennis* (Diptera: Glossinidae) at Nguruman, south-  
607 west Kenya. *Bulletin of Entomological Research*. <https://doi.org/10.1017/S0007485300051695>
- 608 Broennimann, O., Fitzpatrick, M. C., Pearman, P. B., Petitpierre, B., Pellissier, L., Yoccoz, N. G., ...  
609 Guisan, A. (2012). Measuring ecological niche overlap from occurrence and spatial environmental  
610 data. *Global Ecology and Biogeography*. <https://doi.org/10.1111/j.1466-8238.2011.00698.x>
- 611 Camberlin, P., & Wairoto, J. G. (1997). Intraseasonal wind anomalies related to wet and dry spells  
612 during the “long” and “short” rainy seasons in Kenya. *Theoretical and Applied Climatology*.  
613 <https://doi.org/10.1007/BF00867432>
- 614 Cavalli-Sforza, L. L., & Edwards, A. W. (1967). Phylogenetic analysis. Models and estimation  
615 procedures. *American Journal of Human Genetics*. <https://doi.org/10.2307/2406616>
- 616 Cecchi, G. (2002). Predicted areas of suitability for tsetse flies (*Glossina pallidipes*) [Raster].  
617 FAO/IAEA. Retrieved from

- 618        <http://www.fao.org/geonetwork/srv/en/graphover.show?id=12758&fname=kenyalarge.gif&access>  
619        =public.
- 620        Cecchi, G., Mattioli, R. C., Slingenbergh, J., & De La Rocque, S. (2008). Land cover and tsetse fly  
621        distributions in sub-Saharan Africa. *Medical and Veterinary Entomology*.  
622        <https://doi.org/10.1111/j.1365-2915.2008.00747.x>
- 623        Chanie, M., Adula, D., & Bogale, B. (2013). Socio-Economic Assessment of the Impacts of  
624        Trypanosomiasis on Cattle in Girja District , Southern Oromia Region, Southern Ethiopia. *Acta*  
625        *Parasitologica Globalis*.
- 626        Cuisance, D., Fevrier, J., Dujardin, J. P., & Filledier, J. (1985). Dispersion linéaire de Glossina  
627        palpalis gambiensis et de Glossina tachinoides dans une galerie forestière en zone soudano-  
628        guinéenne (Burkina-Faso). *Livestock and Veterinary Medicine Journal of Tropical Countries*.
- 629        Davis, S., Aksoy, S., & Galvani, A. (2011). A global sensitivity analysis for African sleeping  
630        sickness. *Parasitology*. <https://doi.org/10.1017/S0031182010001496>
- 631        De Araújo, C. B., Marcondes-Machado, L. O., & Costa, G. C. (2014). The importance of biotic  
632        interactions in species distribution models: A test of the Eltonian noise hypothesis using parrots.  
633        *Journal of Biogeography*. <https://doi.org/10.1111/jbi.12234>
- 634        Devisser, M. H., Messina, J. P., Moore, N. J., Lusch, D. P., & Maitima, J. (2010). A dynamic species  
635        distribution model of Glossina subgenus Morsitans: The identification of tsetse reservoirs and  
636        refugia. *Ecosphere*. <https://doi.org/10.1890/ES10-00006.1>
- 637        Dicko, A. H., Lancelot, R., Seck, M. T., Guerrini, L., Sall, B., Lo, M., ... Bouyer, J. (2014). Using  
638        species distribution models to optimize vector control in the framework of the tsetse eradication  
639        campaign in Senegal. *Proceedings of the National Academy of Sciences of the United States of*  
640        *America*. <https://doi.org/10.1073/pnas.1407773111>
- 641        Dormann, C. F. (2007). Promising the future? Global change projections of species distributions.  
642        *Basic and Applied Ecology*. <https://doi.org/10.1016/j.baae.2006.11.001>

- 643 Dray, S., & Dufour, A. B. (2007). The ade4 package: Implementing the duality diagram for  
644 ecologists. *Journal of Statistical Software*. <https://doi.org/10.18637/jss.v022.i04>
- 645 Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., ... E. Zimmermann, N.  
646 (2006). Novel methods improve prediction of species' distributions from occurrence data.  
647 *Ecography*. <https://doi.org/10.1111/j.2006.0906-7590.04596.x>
- 648 Faith, J. T., Tryon, C. A., & Peppe, D. J. (2016). Environmental change, ungulate biogeography, and  
649 their implications for early human dispersals in equatorial East Africa. *Vertebrate Paleobiology*  
650 and *Paleoanthropology*. [https://doi.org/10.1007/978-94-017-7520-5\\_13](https://doi.org/10.1007/978-94-017-7520-5_13)
- 651 Ford, J. (1971). *The role of the trypanosomiases in African ecology. A study of the tsetse fly problem.*  
652 London: Clarendon Press, Oxford University Press.
- 653 Franco, J. R., Simarro, P. P., Diarra, A., & Jannin, J. G. (2014). Epidemiology of human African  
654 trypanosomiasis. *Clinical Epidemiology*. <https://doi.org/10.2147/CLEP.S39728>
- 655 Franco, J. R., Cecchi, G., Priotto, G., Paone, M., Diarra, A., Grout, L., ... Argaw, D. (2020).  
656 Monitoring the elimination of human African trypanosomiasis at continental and country level:  
657 Update to 2018. *PLoS Neglected Tropical Diseases*. <https://doi.org/10.1371/journal.pntd.0008261>
- 658 Garzón, M. B., Blazek, R., Neteler, M., Dios, R. S. de, Ollero, H. S., & Furlanello, C. (2006).  
659 Predicting habitat suitability with machine learning models: The potential area of Pinus sylvestris  
660 L. in the Iberian Peninsula. *Ecological Modelling*.  
661 <https://doi.org/10.1016/j.ecolmodel.2006.03.015>
- 662 Gilbert, J. A., Medlock, J., Townsend, J. P., Aksoy, S., Ndeffo Mbah, M., & Galvani, A. P. (2016).  
663 Determinants of Human African Trypanosomiasis Elimination via Paratransgenesis. *PLoS*  
664 *Neglected Tropical Diseases*. <https://doi.org/10.1371/journal.pntd.0004465>
- 665 Grady, S. C., Messina, J. P., & McCord, P. F. (2011). Population vulnerability and disability in  
666 Kenya's tsetse fly habitats. *PLoS Neglected Tropical Diseases*.  
667 <https://doi.org/10.1371/journal.pntd.0000957>

- 668 Hargrove, J. W. (2009). *Tsetse population dynamics. The Trypanosomiases*.  
669 <https://doi.org/10.1079/9780851994758.0113>
- 670 Hether, T. D., & Hoffman, E. A. (2012). Machine learning identifies specific habitats associated with  
671 genetic connectivity in *Hyla squirella*. *Journal of Evolutionary Biology*.  
672 <https://doi.org/10.1111/j.1420-9101.2012.02497.x>
- 673 Hijmans, R. J. (2019). raster: Geographic Data Analysis and Modeling. Retrieved from <https://cran.r-project.org/package=raster>  
674
- 675 Hijmans, R. J., Phillips, S., Leathwick, J., & Elith, J. (2017). dismo: Species Distribution Modeling.  
676 Retrieved from <https://cran.r-project.org/package=dismo>
- 677 Hill, L., Hector, A., Hemery, G., Smart, S., Tanadini, M., & Brown, N. (2017). Abundance  
678 distributions for tree species in Great Britain: A two-stage approach to modeling abundance using  
679 species distribution modeling and random forest. *Ecology and Evolution*, 7(4), 1043–1056.  
680 <https://doi.org/10.1002/ece3.2661>
- 681 Hirzel, A. H., & Le Lay, G. (2008). Habitat suitability modelling and niche theory. *Journal of Applied  
682 Ecology*. <https://doi.org/10.1111/j.1365-2664.2008.01524.x>
- 683 Illemobade, A. A. (2009). Tsetse and trypanosomosis in Africa: The challenges, the opportunities. In  
684 *Onderstepoort Journal of Veterinary Research*. <https://doi.org/10.4102/ojvr.v76i1.59>
- 685 Jombart, T. (2008). Adegenet: A R package for the multivariate analysis of genetic markers.  
686 *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btn129>
- 687 Jordan, A. M. (1993). Tsetse flies (Glossinidae). *Medical Insects and Arachnids*, 333–388.
- 688 Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., ... Kessler, M.  
689 (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific Data*.  
690 <https://doi.org/10.1038/sdata.2017.122>

- 691 Kassambara, A. (2019). *ggpubr*: “*ggplot2*” Based Publication Ready Plots. Retrieved from  
692 <https://cran.r-project.org/package=ggpubr>
- 693 Lang, D. T., & the CRAN Team. (2019). *XML*: Tools for Parsing and Generating XML Within R and  
694 S-Plus. Retrieved from <https://cran.r-project.org/package=XML>
- 695 Lehmann, T., Hawley, W. A., Grebert, H., Danga, M., Atieli, F., & Collins, F. H. (1999). The Rift  
696 Valley complex as a barrier to gene flow for *Anopheles gambiae* in Kenya. *Journal of Heredity*.  
697 <https://doi.org/10.1093/jhered/90.6.613>
- 698 Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3), 18–22.  
699 Retrieved from <https://cran.r-project.org/doc/Rnews/>
- 700 Linder, H. P., de Klerk, H. M., Born, J., Burgess, N. D., Fjeldså, J., & Rahbek, C. (2012). The  
701 partitioning of Africa: Statistically defined biogeographical regions in sub-Saharan Africa.  
702 *Journal of Biogeography*. <https://doi.org/10.1111/j.1365-2699.2012.02728.x>
- 703 Manel, S., Schwartz, M. K., Luikart, G., & Taberlet, P. (2003). Landscape genetics: Combining  
704 landscape ecology and population genetics. *Trends in Ecology and Evolution*.  
705 [https://doi.org/10.1016/S0169-5347\(03\)00008-9](https://doi.org/10.1016/S0169-5347(03)00008-9)
- 706 Manel, S., Schwartz, M. K., Luikart, G., & Taberlet, P. (2003). Landscape genetics: Combining  
707 landscape ecology and population genetics. *Trends in Ecology and Evolution*.  
708 [https://doi.org/10.1016/S0169-5347\(03\)00008-9](https://doi.org/10.1016/S0169-5347(03)00008-9)
- 709 Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach.  
710 *Cancer Research*.
- 711 Medlock, J., Atkins, K. E., Thomas, D. N., Aksoy, S., & Galvani, A. P. (2013). Evaluating  
712 Paratransgenesis as a Potential Control Strategy for African Trypanosomiasis. *PLoS Neglected  
713 Tropical Diseases*. <https://doi.org/10.1371/journal.pntd.0002374>

- 714 Murphy, M. A., Evans, J. S., & Storfer, A. (2010). Quantifying *Bufo boreas* connectivity in  
715 Yellowstone National Park with landscape genetics. *Ecology*. <https://doi.org/10.1890/08-0879.1>
- 716 Narladkar, B. W. (2018). Projected economic losses due to vector and vector-borne parasitic diseases  
717 in livestock of india and its significance in implementing the concept of integrated practices for  
718 vector management. *Veterinary World*. <https://doi.org/10.14202/vetworld.2018.151-160>
- 719 Ngari, N. N., Gamba, D. O., Olet, P. A., Zhao, W., Paone, M., & Cecchi, G. (2020). Developing a  
720 national atlas to support the progressive control of tsetse-transmitted animal trypanosomosis in  
721 Kenya. *Parasites and Vectors*. <https://doi.org/10.1186/s13071-020-04156-5>
- 722 Ndeffo-Mbah, M. L., Pandey, A., Atkins, K. E., Aksoy, S., & Galvani, A. P. (2019). The impact of  
723 vector migration on the effectiveness of strategies to control gambiense human African  
724 trypanosomiasis. *PLOS Neglected Tropical Diseases*, 13(12), 1–15.  
725 <https://doi.org/10.1371/journal.pntd.0007903>
- 726 Nnko, H. J., Ngonyoka, A., Salekwa, L., Estes, A. B., Hudson, P. J., Gwakisa, P. S., & Cattadori, I.  
727 M. (2017). Seasonal variation of tsetse fly species abundance and prevalence of trypanosomes in  
728 the Maasai Steppe, Tanzania. *Journal of Vector Ecology*. <https://doi.org/10.1111/jvec.12236>
- 729 Okeyo, W. A., Saarman, N. P., Bateta, R., Dion, K., Mengual, M., Mireji, P. O., ... Caccone, A.  
730 (2018). Genetic differentiation of *Glossina pallidipes* tsetse flies in Southern Kenya. *American  
731 Journal of Tropical Medicine and Hygiene*, 99(4), 945–953. [https://doi.org/10.4269/ajtmh.18-0154](https://doi.org/10.4269/ajtmh.18-<br/>732 0154)
- 733 Okeyo, W. A., Saarman, N. P., Mengual, M., Dion, K., Bateta, R., Mireji, P. O., ... Caccone, A.  
734 (2017). Temporal genetic differentiation in *Glossina pallidipes* tsetse fly populations in Kenya.  
735 *Parasites and Vectors*, 10(1), 1–13. <https://doi.org/10.1186/s13071-017-2415-y>
- 736 Opiro, R., Saarman, N. P., Echodu, R., Opiyo, E. A., Dion, K., Halyard, A., ... Caccone, A. (2017).  
737 Genetic diversity and population structure of the tsetse fly *Glossina fuscipes fuscipes* (Diptera:

- 738        Glossinidae) in Northern Uganda: Implications for vector control. *PLoS Neglected Tropical*  
739        *Diseases*. <https://doi.org/10.1371/journal.pntd.0005485>
- 740        Ouma, J. O., Marquez, J. G., & Krafur, E. S. (2006). Microgeographical breeding structure of the  
741        tsetse fly, *Glossina pallidipes* in south-western Kenya. *Medical and veterinary entomology*, 20(1),  
742        138–149. <https://doi.org/10.1111/j.1365-2915.2006.00609.x>
- 743        Pandey, A., Atkins, K. E., Bucheton, B., Camara, M., Aksoy, S., Galvani, A. P., & Ndeffo-Mbah, M.  
744        L. (2015). Evaluating long-term effectiveness of sleeping sickness control measures in Guinea.  
745        *Parasites and Vectors*. <https://doi.org/10.1186/s13071-015-1121-x>
- 746        Pebesma, E. J., & Bivand, R. S. (2005). Classes and methods for spatial data in {R}. *R News*, 5(2), 9–  
747        13. Retrieved from <https://cran.r-project.org/doc/Rnews/>
- 748        Pedersen, T. L. (2020). patchwork: The Composer of Plots. Retrieved from <https://cran.r-project.org/package=patchwork>
- 750        Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009).  
751        Sample selection bias and presence-only distribution models: Implications for background and  
752        pseudo-absence data. *Ecological Applications*. <https://doi.org/10.1890/07-2153.1>
- 753        Pless, E., Saarman, N. P., Powell, J. R., Caccone, A., & Amatulli, G. (2021). A machine-learning  
754        approach to map landscape connectivity in *Aedes aegypti* with genetic and environmental data.  
755        *Proceedings of the National Academy of Sciences*, 118(9).  
756        <https://doi.org/10.1073/pnas.2003201118>
- 757        Pollock, J. N. (1982). Training Manual for Tsetse Control Personnel. Vol.1: Tsetse biology,  
758        systematics and distribution, techniques. *FAO*.
- 759        Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques:  
760        Bagging and random forests for ecological prediction. *Ecosystems*.  
761        <https://doi.org/10.1007/s10021-005-0054-1>

- 762 Ram, K., & Wickham, H. (2018). wesanderson: A Wes Anderson Palette Generator. Retrieved from  
763 <https://cran.r-project.org/package=wesanderson>
- 764 Rehfeldt, G. E., Crookston, N. L., Warwell, M. V., & Evans, J. S. (2006). Empirical analyses of plant-  
765 climate relationships for the Western United States. *International Journal of Plant Sciences*.  
766 <https://doi.org/10.1086/507711>
- 767 Rogers, D. (1977). Study of a Natural Population of Glossina fuscipes fuscipes Newstead and a Model  
768 of Fly Movement. *The Journal of Animal Ecology*. <https://doi.org/10.2307/3962>
- 769 Rogers, D. J., & Randolph, S. E. (1991). Mortality rates and population density of tsetse flies  
770 correlated with satellite imagery. *Nature*. <https://doi.org/10.1038/351739a0>
- 771 Rogers, D. J., & Randolph, S. E. (1985). Population ecology of tsetse. *Annual Review of Entomology*.  
772 Vol. 30.
- 773 Rogers, D. J., & Robinson, T. P. (2004). Tsetse distribution. In I. Maudlin, P. H. Holmes, & M. A.  
774 Miles (Eds.), *The trypanosomiases* (pp. 139–179). Wallingford: CABI.  
775 <https://doi.org/10.1079/9780851994758.0139>
- 776 Rohr, J. R., Barrett, C. B., Civitello, D. J., Craft, M. E., Delius, B., DeLeo, G. A., ... Tilman, D.  
777 (2019). Emerging human infectious diseases and the links to global food production. *Nature  
778 Sustainability*. <https://doi.org/10.1038/s41893-019-0293-3>
- 779 Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under  
780 isolation by distance. *Genetics*.
- 781 Scholz, F., & Zhu, A. (2019). kSamples: K-Sample Rank Tests and their Combinations. Retrieved  
782 from <https://cran.r-project.org/package=kSamples>
- 783 Shi, X., Li, M., Hunter, O., Guetti, B., Andrew, A., Stommel, E., ... Karagas, M. (2019). Estimation  
784 of environmental exposure: interpolation, kernel density estimation or snapshotting. *Annals of  
785 GIS*. <https://doi.org/10.1080/19475683.2018.1555188>

- 786 Slowikowski, K. (2020). ggrepel: Automatically Position Non-Overlapping Text Labels with  
787 “ggplot2.” Retrieved from <https://cran.r-project.org/package=ggrepel>
- 788 Solano, P., Kaba, D., Ravel, S., Dyer, N. A., Sall, B., Vreysen, M. J. B., ... Bouyer, J. (2010).  
789 Population genetics as a tool to select tsetse control strategies: Suppression or eradication of  
790 *Glossina palpalis gambiensis* in the niayes of senegal. *PLoS Neglected Tropical Diseases*.  
791 <https://doi.org/10.1371/journal.pntd.0000692>
- 792 Souris, M., & Demoraes, F. (2019). Improvement of spatial autocorrelation, kernel estimation, and  
793 modeling methods by spatial standardization on distance. *ISPRS International Journal of Geo-*  
794 *Information*. <https://doi.org/10.3390/ijgi8040199>
- 795 Thuiller, W., Georges, D., Engler, R., & Breiner, F. (2019). biomod2: Ensemble Platform for Species  
796 Distribution Modeling. Retrieved from <https://cran.r-project.org/package=biomod2>
- 797 Urban, M. C., Bocedi, G., Hendry, A. P., Mihoub, J.-B., Pe'er, G., Singer, A., ... Travis, J. M. J.  
798 (2016). Improving the forecast for biodiversity under climate change. *Science*, 353(6304),  
799 aad8466. <https://doi.org/10.1126/science.aad8466>
- 800 Wagner, H. H., & Fortin, M. J. (2005). Spatial analysis of landscapes: Concepts and statistics.  
801 *Ecology*. <https://doi.org/10.1890/04-0914>
- 802 Wand, M. (2015). KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones 1995.  
803 Retrieved from <https://cran.r-project.org/package=KernSmooth>
- 804 Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*.  
805 <https://doi.org/10.1007/978-0-387-98141-3>
- 806 Wilfert, L., Kaib, M., Durka, W., & Brandl, R. (2006). Differentiation between populations of a  
807 termite in eastern Africa: Implications for biogeography. *Journal of Biogeography*.  
808 <https://doi.org/10.1111/j.1365-2699.2006.01556.x>

- 809 World Health Organization. (2020). Trypanosomiasis, human African (sleeping sickness). Retrieved  
810 from <https://www.who.int/news-room/fact-sheets/detail/trypanosomiasis-human-african-sleeping-sickness>
- 811
- 812 Wright, S. (1943). Isolation by Distance. *Genetics*, 28(2), 114–138. Retrieved from  
813 <https://www.genetics.org/content/28/2/114>
- 814 Wüster, W., Crookes, S., Ineich, I., Mané, Y., Pook, C. E., Trape, J. F., & Broadley, D. G. (2007).  
815 The phylogeny of cobras inferred from mitochondrial DNA sequences: Evolution of venom  
816 spitting and the phylogeography of the African spitting cobras (Serpentes: Elapidae: Naja  
817 nigriceps complex). *Molecular Phylogenetics and Evolution*.  
818 <https://doi.org/10.1016/j.ympev.2007.07.021>
- 819 Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., ... Bates, P. D.  
820 (2017). A high-accuracy map of global terrain elevations. *Geophysical Research Letters*.  
821 <https://doi.org/10.1002/2017GL072874>
- 822 Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., ... Sequeira, A.  
823 M. M. (2018). Outstanding Challenges in the Transferability of Ecological Models. *Trends in  
824 Ecology & Evolution*, 33(10), 790–802. <https://doi.org/https://doi.org/10.1016/j.tree.2018.08.001>
- 825 Zeller, K. A., McGarigal, K., & Whiteley, A. R. (2012). Estimating landscape resistance to  
826 movement: A review. *Landscape Ecology*. <https://doi.org/10.1007/s10980-012-9737-0>

827 **Figure Captions**

828 **Figure 1.** Map of sampling sites in Kenya and Tanzania, color coded by genetic cluster. The boxed  
829 area of detail is the location of the study region in Africa. The approximate area of the Serengeti  
830 ecosystem is shaded in green (combination of the Maasai Mara National Reserve and the Serengeti  
831 National Park), and the approximate outline of the Great Rift Valley is shaded in purple. The three  
832 new sampling sites for this study (OTT, CNP, and AMR) are labeled. CNP was split into CNPa and  
833 CNPb for our analysis as some trap locations from this sampling site were found to be further than  
834 two kilometers apart (see methods). This map was created using the R packages “ggplot2” (Wickham,  
835 2016), “raster” (Hijmans, 2019), and “rgdal” (Bivand et al., 2019) with publicly available data from  
836 DIVA-GIS (March 2020; <http://www.diva-gis.org>), Map Library (March 2020;  
837 <http://www.maplibrary.org>), World Map (March 2020; <https://worldmap.harvard.edu>) and MaMaSe  
838 (March 2020; <http://maps.mamase.org>).

839

840 **Figure 2.** Diagram of simplified methods. Light gray shaded boxes indicate the separate pipelines for  
841 the suitability (A1, C1) and connectivity (A2, C2) models. The original data inputs are presence-  
842 background data (A1) and microsatellite data (A2) from flies caught during trapping surveys in Kenya  
843 and northern Tanzania as well as remotely sensed data from CHELSA, MERIT, and DIVA-GIS  
844 repositories (A3). See methods for more details on calculation of genetic distances (A2), manipulation  
845 of environmental data (B1, B2), and selection of background points (A1). Dark grey outlined boxes  
846 (C1, C2, C3, C4) illustrate the final outputs of the pipelines (C1, C2), the bivariate map of  
847 connectivity and suitability (C3), and post-hoc analyses (C4).

848

849 **Figure 3.** Maps of RMSE values for each sampling site from the leave-one-out cross-validation  
850 results. Sampling sites are color coded by genetic cluster: **(A)** RMSE values from external validation  
851 of the genetic connectivity model, **(B)** RMSE values from the spatial evaluation of the genetic  
852 connectivity map (the projection of the genetic connectivity model). Sites with high error compared to  
853 other sites and to the null models are labeled (File 1S).

854

855   **Figure 4.** Predicted genetic connectivity and habitat suitability based on machine learning (random  
856 forest) models. White areas in all three maps are regions where the predicted probability of *G.*  
857 *pallidipes* presence is less than ten percent, based on the habitat suitability map. (**A**) scaled map of  
858 habitat suitability (combination of our final model and the FAO model), (**B**) scaled and transformed (  
859  $1 - \text{scaled genetic distance}$ ) map of genetic connectivity, (**C**) bivariate map of genetic connectivity  
860 versus habitat suitability. The bivariate legend in the bottom left-hand corner shows the corresponding  
861 colors for the different percentiles of genetic connectivity and habitat suitability (dark red: high  
862 genetic connectivity/high habitat suitability, yellow: high genetic connectivity/low habitat suitability,  
863 blue: low genetic connectivity/high habitat suitability, gray: low genetic connectivity/low habitat  
864 suitability).

865

866   **Figure 5.** Variable importance plots for (A) the 10 replicate habitat suitability models and (B) the  
867 final genetic connectivity model. Only the top 10 most important variables are shown, for the full  
868 variable importance plots see supplemental Figure 6S. The R package “randomForest” measures  
869 importance based on the increase in node purity (IncNodePurity). Variables correspond to those  
870 described in Table 1S. (C) Post-hoc analyses of the most important predictor variable for habitat  
871 suitability (left column) and genetic connectivity (right column). The first row of maps show the  
872 current environmental conditions (color palette from the “wesanderson” package; Ram &  
873 Wickham, 2018). The second row of maps shows the local Pearson's correlations between the top  
874 predictor variables and response variables of interest (i.e. maximum temperature of the warmest  
875 month vs suitability (probability of presence) and precipitation of the dries season vs connectivity (  
876  $1 - \text{scaled genetic distance}$ )). The local correlation coefficients were calculated with the corLocal()  
877 function from the R package “raster” (neighborhood size = 21; Hijmans, 2019). The third row shows  
878 maps of the predicted future change in the top predictor variables under the NASA RCP 4.5 climate  
879 change model for 2041 - 2060. White areas in all maps are regions where the predicted probability of  
880 *G. pallidipes* presence is less than ten percent, based on the habitat suitability model. Abbreviations:  
881 Precipitation (Prec), Temperature (Temp), Maximum (Max), Correlation (Corr), Month (Mo).

882    **Supporting Information Captions**

883    **Table 1S: Environmental variables included as predictors in machine learning models.**

884    Complete list of 22 environmental variables. All “BIO” (Bioclimatic) variables were created using  
885    CHELSA data and the R package “biomod2” based on the bioclimatic variables from Worldclim  
886    (Thuiller et al., 2019; Karger et al., 2017). Bioclimatic variables ending in “S” are seasonal  
887    calculations of synonymous quarterly bioclimatic variables based on the precipitation cycles of Kenya  
888    (Figure 1S).

889    **Table 2S: Comparison of observed and predicted distributions of genetic distance.** Table of  
890    results from Anderson Darling k-means tests comparing the (a) observed Cavalli-Sforza and Edwards’  
891    chord (CSE) genetic distance to predicted distributions based on models with environmental  
892    predictors of increasing complexity: (b) geographic distance only, (c) sampling density only, (d)  
893    geographic distance and sampling density, (e) environmental variables only, (f) environmental  
894    variables and geographic distance, (g) environmental variables and sampling density, and (h) the final  
895    full model with all environmental variables, geographic distance, and sampling density. Values in the  
896    lower triangle of this table are p-values and values in the upper triangle are Anderson-Darling  
897    Criterion values (with Anderson Darling standardized test statistics in parenthesis).

898    **Figure 1S. Characterization of seasons based on mean precipitation.** Monthly precipitation at  
899    sampling sites, grouped by season.

900    **Figure 2S. Genetic clustering results.** Cluster membership assignments for each site based on a  
901    Discriminate Analysis of Principal Components (DAPC) results. Size of each box is proportional to  
902    the number of individuals assigned to that group (i.e. cluster).

903    **Figure 3S.** Spatial distributions of Cavalli-Sforza and Edwards’ chord (CSE) genetic distance in the  
904    two major genetic clusters east and west of the Great Rift Valley. Density plots depict the distribution  
905    of CSE values for each genetic cluster. On the map, paths between sites within genetic clusters are  
906    colored according to their corresponding CSE value (darker values indicate higher genetic distance).

907    **Figure 4S. Mantel tests for correlation of geographic and genetic distance.** Results of the mantel  
908    tests for the **(A)** eastern and **(B)** western major genetic clusters, and **(C)** the western Serengeti sub-  
909    cluster (Bateta et al., 2020). The Lake Victoria sub-cluster (Bateta et al., 2020) was not included

910 because of insufficient sample size. Plotted red lines are based on a linear model of Cavalli-Sforza  
911 and Edwards' chord (CSE) genetic distance and geographic distance (km). Simulated p-values are  
912 based on 999 replicates.

913 **Figure 5S. Projections of final models of habitat suitability and genetic connectivity.** Raw  
914 projections of (A) the combined habitat suitability model and (B) the genetic connectivity model. The  
915 R-squared of the habitat suitability model (A) is based on the average R-squared of the 10 model  
916 replicates built using different sets of randomly sampled background points and all presence points.  
917 The R-squared of the genetic connectivity model (B) is the R-squared of the final model created with  
918 all of the data.

919 **Figure 6S. Variable importance for models of habitat suitability and genetic connectivity.**  
920 Variable importance plots for (A) the 10 replicate habitat suitability models and (B) the final genetic  
921 connectivity model. The R package "randomForest" measures importance based on the increase in  
922 node purity (IncNodePurity), which is calculated by taking the decrease in the Residual Sum of  
923 Squares (RSS) as the result of splitting on each variable and averaging it across all trees (Liaw &  
924 Wiener, 2002). Variables correspond to those described in Table 1S. Abbreviations: Precipitation  
925 (Prec), Temperature (Temp).

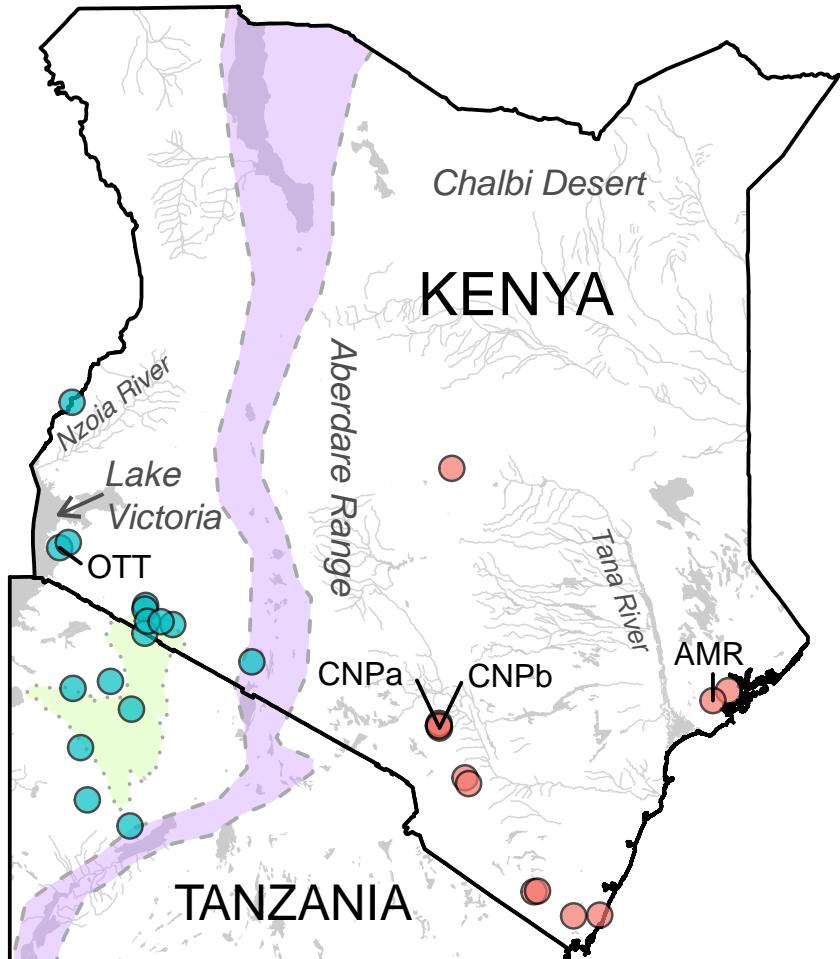
926 **Figure 7S. Comparison of observed and null RMSE distributions.** Density plots of the observed  
927 distribution of RMSE values (red) from the connectivity model compared to 100 null distributions of  
928 RMSE values from models built using shuffled data (black). (A) Comparison of observed and null  
929 RMSE values from the model evaluation, (B) Comparison of the observed and null RMSE values  
930 from the spatial evaluation.

931 **Figure 8S. Distributions of observed and predicted genetic distances.** Density plots depicting the  
932 distribution of Cavalli-Sforza and Edwards' chord (CSE) genetic distance values from the observed  
933 data (first plot) and from predictions of models with different variable combinations. The R squared  
934 values (RSQ) displayed are from random forest models created using the full dataset and the selected  
935 variables (as described in the plot titles). The p-values (p) in red are from Anderson-Darling k-sample  
936 tests used to compare the predicted distributions to the observed distribution (graphed in red).  
937 "Environmental" is abbreviated as "Env".

938 **Figure 9S.** Comparison of random forest and simple linear regression model projections. Both models  
939 were built using the same response (CSE) and predictor variables. The left column of graphs are  
940 projections of the linear model. The right column of graphs are projections of the random forest  
941 model. The top row of graphs are maps set to the default scales (range of each projection). The bottom  
942 row of graphs are maps set to the scale of the observed data (range of observed CSE values).

943 **File 1S. Supplemental methods and results.** Description of methods and associated results that were  
944 not part of the central data flow of our pipeline, but were important in evaluation and ensuring  
945 repeatability of our study. We include details from (I) the habitat suitability modeling on selection of  
946 background points, (II) the genetic connectivity modeling on population structure, accounting for  
947 spatial auto-correlation, model evaluation with leave-one-out cross-validation, and comparison of the  
948 random forest method we use and linear methods, and (III) creation of the bivariate map.

949 **File 2S. Code for the bivariate map that summarizes predicted habitat use and connectivity.**  
950 Code developed to create bivariate map of genetic connectivity and habitat suitability (C3, Fig 4)  
951 using R packages “raster” (Hijmans, 2019), “rgdal” (Bivand et al., 2019), “classInt” (Bivand, 2018),  
952 “XML” (Lang et al., 2019).



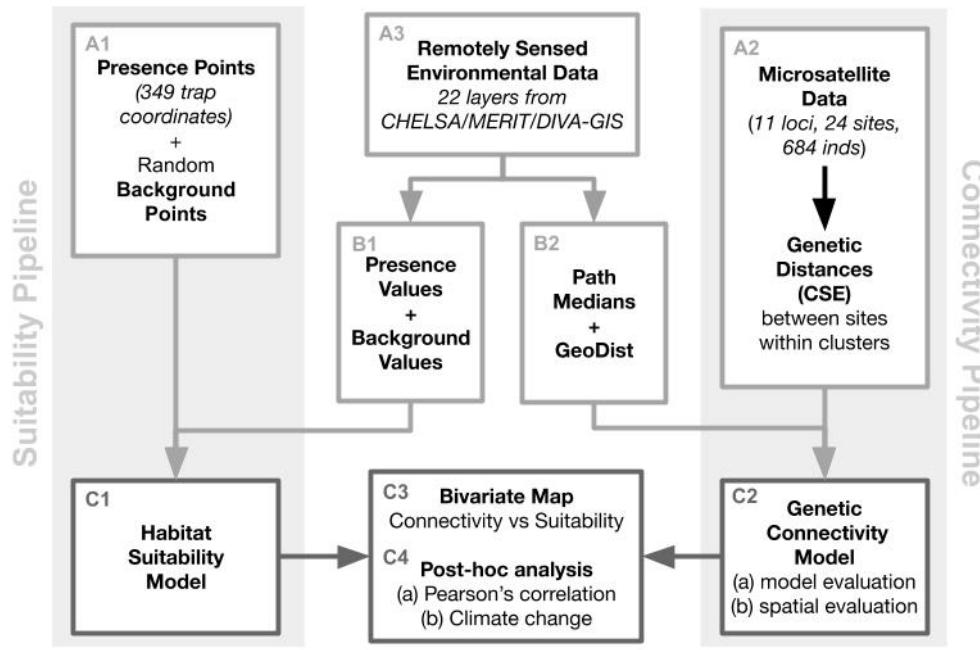
#### Key

- Country Borders
- Inland Water
- Great Rift Valley
- SNP + MMNR

#### Cluster

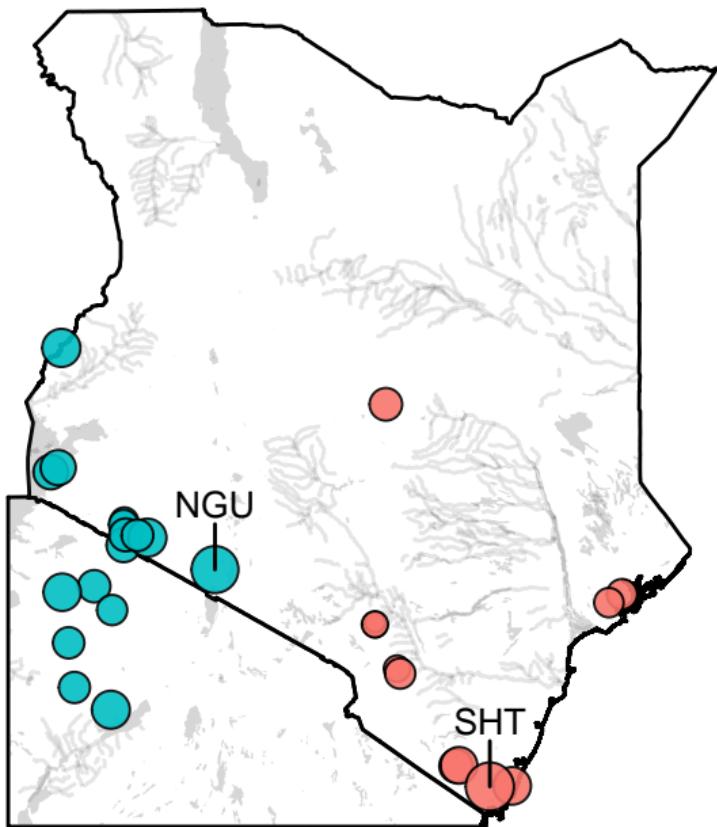
- east
- west

# Accepted Article

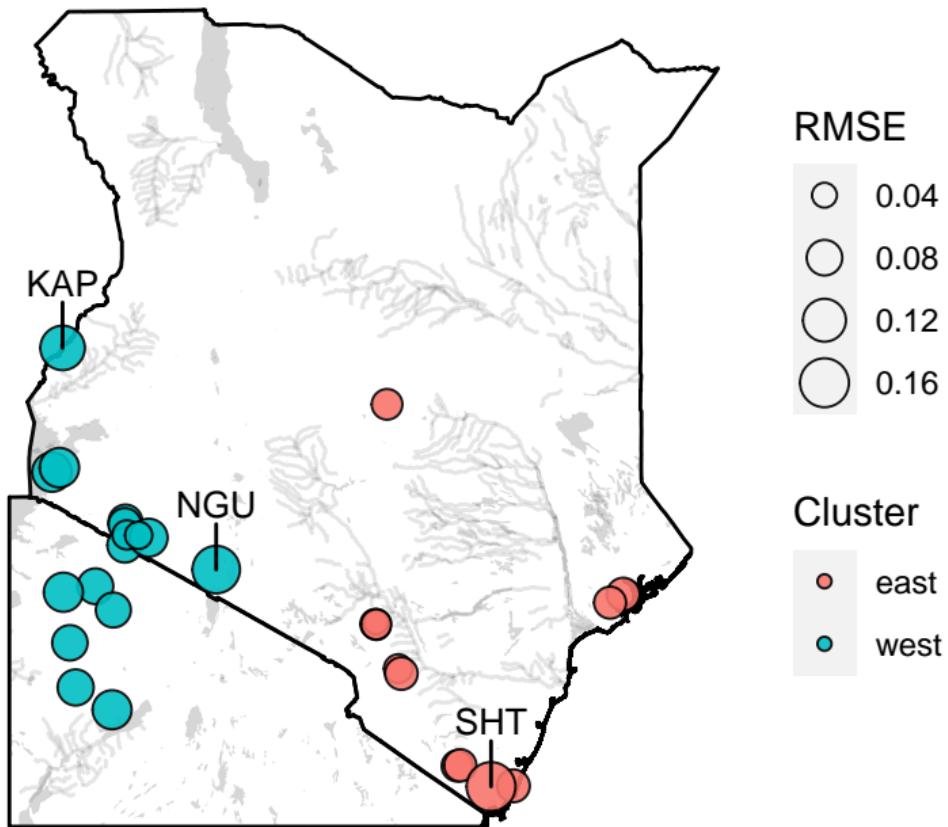


eva\_13237\_f2.jpg

A. Model Evaluation



B. Spatial Evaluation

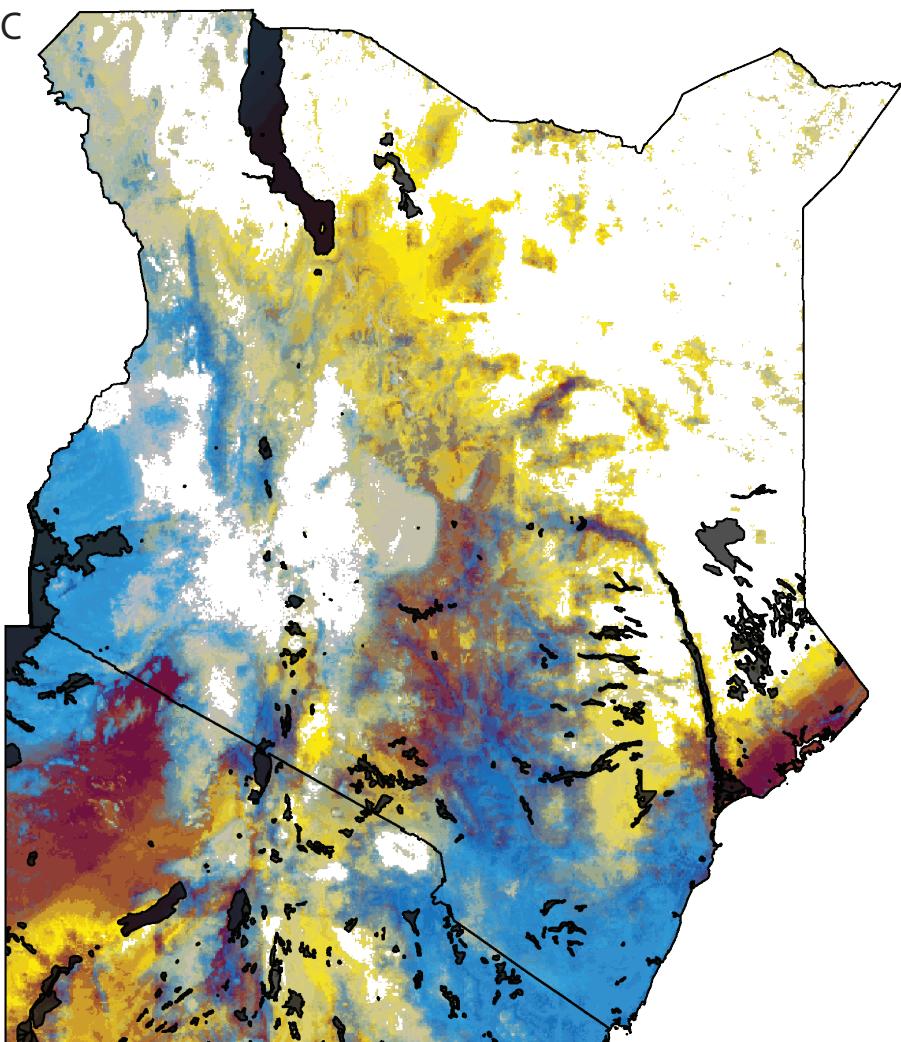
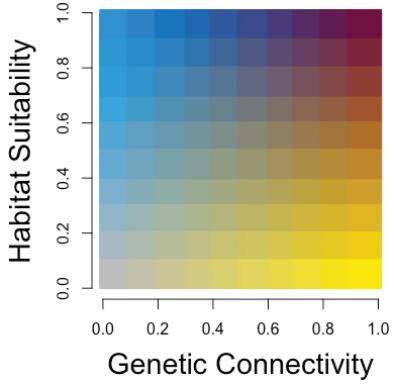
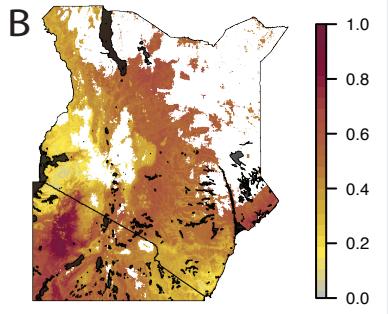
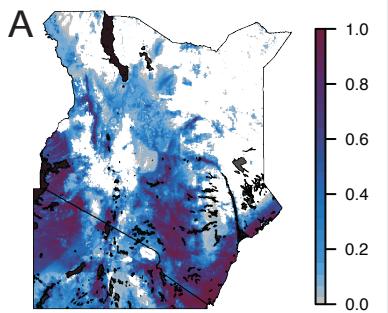


RMSE

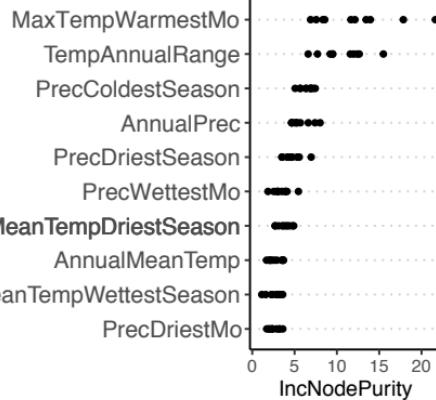
- 0.04
- 0.08
- 0.12
- 0.16

Cluster

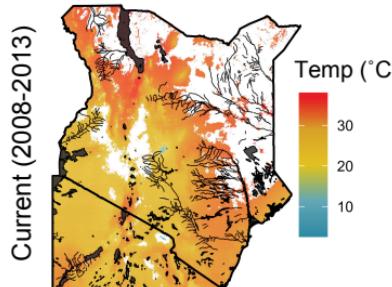
- east
- west



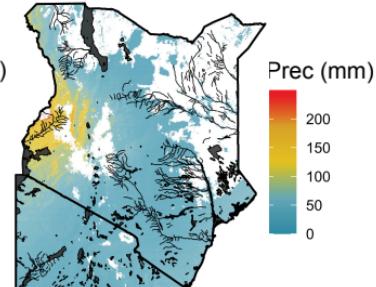
### A. Habitat Suitability Variable Importance



### C. Max Temp of Warmest Month Top Predictor of Suitability



### Prec of Driest Season Top Predictor of Connectivity



### B. Genetic Connectivity Variable Importance

