Predictive mapping of mosquito distribution based on environmental and anthropogenic factors in Taita Hills, Kenya

Ruut Uusitalo[1,2,3], Mika Siljander[1], C. Lorna Culverwell[2,3,4], Noah C. Mutai[5], Kristian M. Forbes[2], Olli Vapalahti [2,3,6] Petri K.E. Pellikka[1,7,8]

[1]Department of Geosciences and Geography, P.O. Box 64, FI-00014 University of Helsinki, Finland; mika.siljander@helsinki.fi; petri.pellikka@helsinki.fi

[2]Department of Virology, Haartmaninkatu 3, P.O. Box 21, FI-00014 University of Helsinki, Finland; lorna.culverwell@helsinki.fi; kristian.forbes@helsinki.fi; olli.vapalahti@helsinki.fi

[3]Department of Veterinary Biosciences, Agnes Sjöberginkatu 2, P.O. Box 66, FI-00014 University of Helsinki, Finland

[4]Department of Life Sciences, Natural History Museum, Cromwell Road, London, SW5 7BD

[5]Department of Mathematics and Informatics, Taita Taveta University, P. O. Box 635-80300, Voi, Kenya

[6]Virology and Immunology, HUSLAB, Helsinki University Hospital

[7]Helsinki Institute of Sustainability Science, University of Helsinki, Finland

[8]Institute for Atmospheric and Earth System Research, University of Helsinki, Finland

Corresponding author:
Ruut Uusitalo
University of Helsinki, Department of Geosciences and Geography, P.O. Box 64, FI-00014
ruut.uusitalo@helsinki.fi; Tel.: +358 (0)50 5911280

Abstract

Mosquitoes are vectors for numerous pathogens, which are collectively responsible for millions of human deaths each year. As such, it is vital to be able to accurately predict their distributions, particularly in areas where species composition is unknown. Species distribution modeling was used to determine the relationship between environmental, anthropogenic and distance factors on the occurrence of two mosquito genera, *Culex* Linnaeus and *Stegomyia* Theobald (syn. *Aedes*), in the Taita Hills, southeastern Kenya. This study aims to test whether any of the statistical prediction models produced by the Biomod2 package in R can reliably estimate the distributions of mosquitoes

in these genera in the Taita Hills; and to examine which factors best explain their presence. Mosquito collections were collected from 122 locations between January–March 2016 along transects throughout the Taita Hills. Environmental-, anthropogenic- and distance-based geospatial data were acquired from the Taita Hills geo-database, satellite- and aerial imagery and processed in GIS software. The Biomod2 package in R, intended for ensemble forecasting of species distributions, was used to generate predictive models. Slope, human population density, normalized difference vegetation index, distance to roads and elevation best estimated *Culex* distributions by a generalized additive model with an area under the curve (AUC) value of 0.791. Mean radiation, human population density, normalized difference vegetation index, distance to roads and mean temperature resulted in the highest AUC (0.708) value in a random forest model for *Stegomyia* distributions. We conclude that in the process towards more detailed species-level maps, with our study results, general assumptions can be made about the distribution areas of *Culex* and *Stegomyia* mosquitoes in the Taita Hills and the factors which influence their distribution.

Keywords: species distribution modeling; vector-borne disease; GIS; predictive mapping; mosquito, biomod2

1. Introduction

Environmental and anthropogenic disturbances such as climate change, urbanization and deforestation are crucial factors in the distribution of pathogen vectors and the emergence of diseases that they transmit (Crowl *et al.*, 2008). This is already evident as the global rise in human population densities, and correlated mean temperatures leading to land-use changes, has created new suitable breeding sites for mosquitoes (Roiz *et al.*, 2011; Campbell *et al.*, 2015). Geographic information system (GIS) and species distribution modeling (SDM) approaches have been used to understand these connections and their implications for the spread of invasive species, particularly with regard to northern regions (Neteler *et al.*, 2011). Other important areas, however, remain unexplored, including biodiversity rich but increasingly fragmented parts of Africa, where infectious diseases can be highly prevalent but access to health care is limited (Bhutta *et al.*, 2014). As a result, tools such as GIS and SDM that can estimate and predict infectious disease risks have the potential to optimize the use of limited resources and improve public health outcomes.

Due to their ability to act as vectors for a suite of pathogens, mosquitoes (Diptera, Culicidae) are among the most economically and socially important taxa on the planet (WHO, 2017). Two notable

genera are *Culex* Linnaeus and *Stegomyia* Theobald (following the classification of Reinert *et al.,* (2009)). *Culex* is a large genus of mosquitoes, with 769 species in 26 subgenera (MTI, 2017), accounting for 21.6% of all mosquito species worldwide. Species of *Culex* have an almost worldwide distribution from the tropics to cool temperate regions, but do not extend into extreme northern latitudes (MTI, 2017). While species from other subgenera are known to vector pathogens, subgenus *Culex* (*Culex*) contains many of the significant human vector species, including those demonstrated to transmit West Nile fever virus, Rift Valley fever virus and Japanese Encephalitis virus (MTI, 2017; WHO, 2017). *Stegomyia* is a moderately sized mosquito genus comprising 128 species, and distributed in the Afrotropical, Oriental and Australasian regions (MTI, 2017). Following human dispersal, at least two species, *St. aegypti* and *St. albopicta*, are also present in Neotropical, Nearctic and Palearctic regions (MTI, 2017; Paupy, *et al.,* 2009). Member species are competent vectors of yellow fever virus (Huang, 1986), dengue virus serotypes 1-4, Zika virus, and Chikungunya virus, among others (Huang, 1990; MTI, 2017).

SDM helps to understand the relationship between wildlife and their environments, to estimate their distributions where empirical data is missing, and to forecast range expansions in the face of environmental change. SDM is used in decision-making for a range of global challenges, such as informing protected areas for wildlife conservation (Guisan *et al.,* 2013), the distribution potential for invasive species (Václavík & Meentemeyer, 2009), and investigating the effects of climate change on species that carry pathogens of public health concern such as mosquitoes (Dukes *et al.,* 2009). Species distribution modeling, therefore, provides a means to understand and predict vector responses to changing climate patterns in dynamic and fragmented environments.

The purpose of this study is to test whether any of the statistical prediction models produced by the Biomod2 package in R, can reliably estimate the distributions of mosquitoes in genera *Culex* and *Stegomyia* in the Taita Hills; to examine which factors best explain their presence; and to create maps of their predicted occurrence. The target area, Taita Hills in rural south-eastern Kenya, is an ecologically diverse area with little previous mosquito research, strong variability in rainfall and a rapidly growing human population. While machine-learning techniques such as maximum entropy modeling is often used to understand mosquito distributions (Mughini-Gras *et al.,* 2014; Sallam *et al.,* 2016), the employed R package, the Biomod2 (Thuiller *et al.,* 2016), has important benefits for estimating mosquito distributions. This is due to its ability to provide improved simulations across initial conditions, test multiple model classes and parameterizations, and include unlimited number of boundary conditions (Thuiller *et al.,* 2009).

2. Methods

2.1 Study area and mosquito collections

The Taita Hills (03°20’S, 38°15’E) in Taita–Taveta county, southeastern Kenya (Fig. 1.) is a range of peaks varying in altitude from 600 to 2200 m above sea level (a.s.l.) (Erdogan *et al.*, 2011). The study area of 286 km² ranges from the Taita Hills at 1800 m a.s.l. to the surrounding lowlands at 900 m a.s.l., as outlined in Fig..

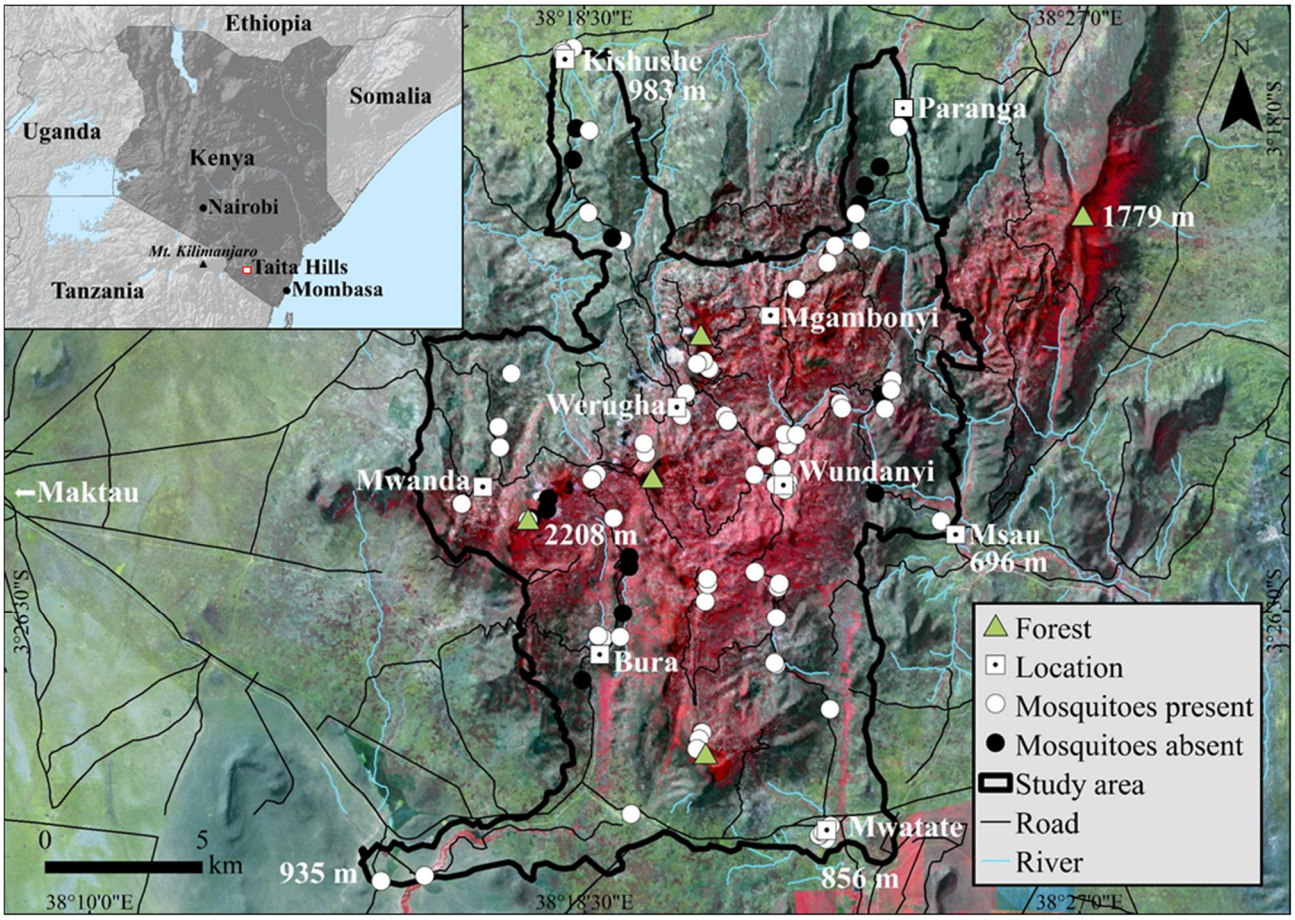


Fig. 1. The Taita Hills and sampling locations in southeastern Kenya. Sampling was conducted along the main roads, from the surrounding lowlands to mountain peaks.

Kenya has two annual wet and dry seasons, with the total rainfall varying across the country (Kaplan *et al.*, 1976; Fig. 2.). The Taita Hills region receives on average 1330–1910 mm/annually, which contributes to the formation of many suitable mosquito breeding habitats (Erdogan *et al.*, 2011).

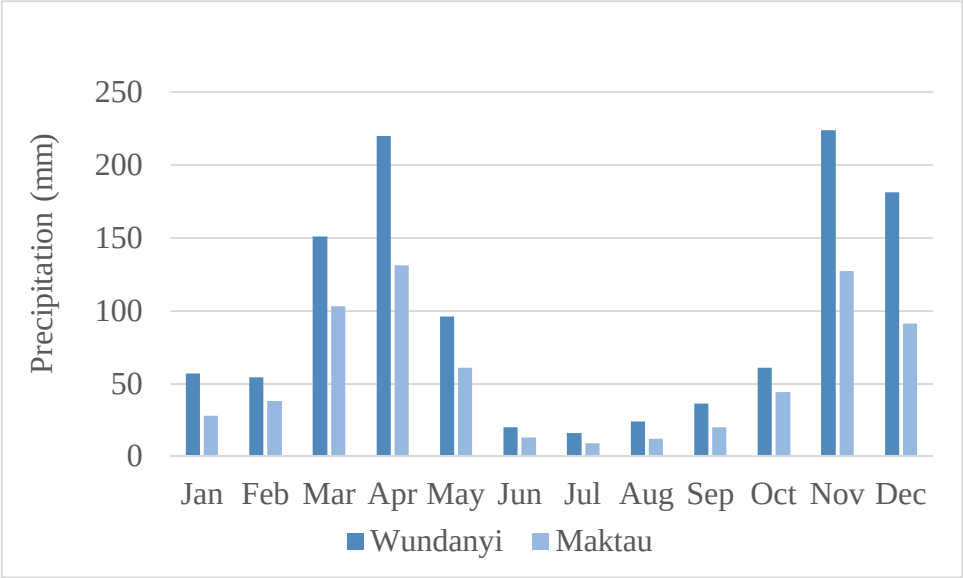


Fig. 2. Monthly average rainfall characteristics for Wundanyi town in the Taita Hills and Maktau village in the lowland plains. Weather data was collected between 1982 and 2012, with a resolution of 30 arc seconds (Climate-Data.org, 2018).

Mosquito collections were attempted at over 122 locations across the region between January and March 2016. This figure includes 'sub-locations', which fell within 500 m of the primary target location, and repeat collections from the same locations across the study time (Fig. 1.). Of these 122 attempts, mosquitoes were present in 107 collections and absent from 15. Mosquito collections started after the rainy season on January 25th, 2016 and continued until March 16th, 2016. Collections did not follow transects up to February 12th, since the goal at this time was to focus on human-biting species around human dwellings. From February 13th until the end of the collection period a stratified sampling scheme was applied based on the road network. Each main road from the lowlands to the highest reachable locations was used to collect mosquitoes on 100 m elevation intervals.

Collection locations and sub-locations were mainly around human dwellings close to roads, but collections were also made in forest fragments and in croplands. Adults were collected using commercially available Prokopacks (The John W. Hock Company, Gainesville, USA) or by using CDC Miniature Light Traps (The John W. Hock Company, Gainesville, USA). Immature life stages (larvae, pupae and eggs) were collected from stagnant water using a 1 litre plastic dipper, a fine meshed aquarium net or a turkey baster. All water sources were considered, including septic tanks, discarded tires, tree cavities, leaf axils and other items filled with rain water, such as discarded artificial containers or water tanks. At each collection location, the GPS point was recorded using a Garmin Map 64S handset. Collections were initially sorted in a field laboratory and stored in ethanol

141 or RNA-later to preserve the RNA and DNA for future studies. Genus-level identifications were made
142 in Finland using suitable identification keys (Service, 1991; Huang, 2000; Huang, 2004).
143
144 2.2 Explanatory environmental, anthropogenic and distance variables
145 The selection of explanatory variables to predict mosquito distributions was based on those used in
146 existing literature. Environmental and anthropogenic data for the Taita Hills region were obtained
147 directly from satellite imagery and airborne digital data, or by deriving them from the satellite
148 imagery in ArcGIS (ESRI, Redlands, CA, USA, v.10.3.1) (Table 1). Distance variables were created
149 from an existing Taita Hills geo-database. For the distance to buildings calculations, existing airborne
150 laser scanning (ALS) data was used. In this data set, buildings were classified from point cloud data
151 that were derived from 2014-2015 Taita Hills ALS flights (Adhikari *et al.,* 2017). LAStools software
152 (rapidlasso GmbH) was used to classify ground, building and vegetation returns.
153
154 Normalized difference vegetation index (NDVI) is a measure of plant greenness, ranging between
155 -1 to 1 (Tucker, 1979), and was used to approximate land cover type. Mean precipitation, mean
156 temperature and mean relative humidity from January - March was calculated for each collection
157 location in ArcGIS. Mean monthly precipitation was derived from long-term precipitation grids
158 between 1987 and 2005 (Hutchinson, 1991; Erdogan *et al.,* 2011). A digital elevation model (DEM)
159 was used to calculate the elevation, slope and mean monthly solar radiation in ArcGIS, and the values
160 were extracted to each mosquito collection location.
161
162 Table 1. Value ranges and sources for explanatory variables used in species distribution modeling.

Environmental, anthropogenic or distance factor	Min.	Max.	Avg.	Data source
Distance to houses (m)	0	1270	52	Distance to houses was determined from the building data computed from airborne laser scanning (ALS) data (Adhikari *et al.,* 2017), modified by digitizing more houses in the study area using QGIS, and calculating Mean Euclidean distance to houses.
Elevation (m)	694	2079	1330	Mean elevation was derived from digital elevation model (DEM). DEM was obtained from scanned Survey of Kenya 1:50 000 scale topographic map from 1991, in which a 20 m planimetric resolution DEM was interpolated from 50 feet interval contours (Clark & Pellikka., 2005).
Distance to roads (m)	0	927	127	Road data was digitized in 2004 from Survey of Kenya 1:50 000 scale topographic map from 1991.

				(Broberg *et al.*, 2004). Modified by calculating Mean Euclidean distance to roads using ArcGIS.
Mean precipitation (mm)	20	113	47	Mean precipitation was obtained from long-term mean precipitation grids, which were interpolated on to a 20 m resolution grid from monthly available meteorological data and surrounding areas between 1987 and 2005, obtained from Kenya Meteorological Department using ANUSPLIN (Australian National University Splines) software (Hutchinson, 1991; Erdogan *et al.*, 2011).
Mean radiation (kWh/m²)	176	228	216	Mean radiation was received on a given surface area in a given time (kWh/m²), was calculated from the DEM using ArcGIS Area Solar Radiation tool (Esri, Redlands, CA, USA).
Mean relative humidity (%)	71	94	77	Mean relative humidity data was created based on data logger observations between April 2013 and May 2014 (Virtanen 2015).
Mean temperature (C°)	15	25	21	Mean temperature data was created based on data logger observations between April 2013 and May 2014 (Virtanen 2015).
Slope (°)	0	43	11	Slope degree was derived from DEM using Slope function in ArcGIS.
Human population density (persons/km²)	0	9090	1260	Human population density was estimated from ALS buildings data and from a random non-stratified household survey (n=100) carried out in October 2006, which found an average of 6 persons per dwelling in the area (Siljander *et al.*, 2011). The data was modified by digitizing more houses in the study area using QGIS.
Normalized Difference Vegetation Index (NDVI)	-0.4	0.2	-0.2	NVDI was derived from a Sentinel-2A MSI Level-1C satellite image from 8 October 2016, downloaded from the Sentinel's Scientific DataHub (ESA, 2015).

2.3. Data analysis and modeling

All the GIS datasets including environmental and other attributes were set to the same spatial extent, geographic coordinate system (WGS 1984 UTM Zone 37S) and resolutions (20 m x 20 m). The compiled data consisted of presence-absence data and explanatory variables for each sampling location. We predicted the distribution of *Culex* and *Stegomyia* mosquitoes using the Biomod2–platform (version 3.3–7) in R software (version 3.0.3; R Development Core Team 2013). Prior to prediction modelling, multicollinearity of the variables was tested using Pearson correlation

coefficients in R, which resulted in the exclusion of four variables. Models were processed mainly using the default settings of the Biomod2. The following eight predictive modeling techniques were used: generalized linear models (GLM), generalized additive models (GAM), classification tree analysis (CTA), artificial neural networks (ANNs), multivariate adaptive regression splines (MARS), general boosting method (GBM), random forest (RF) and maximum entropy (Maxent). Flexible discriminant analysis (FDA) and surface range envelope (SRE) were excluded due to their methodological weaknesses (Hastie *et al.*, 1994).

The original data set of 122 collection locations was randomly divided into model training (70%, n= 85) and model evaluation sets (30%, n=37) (split-sample approach: Guisan & Zimmermann 2000). Models were built using the training set, and the models were validated using the model evaluation set. The area under the curve (AUC) of a receiver operating characteristic (ROC) plot was produced based on each model to estimate the predictive power of the model (Fielding & Bell, 1997), to assess the agreement between the presence-absence records and the predictions. The model types with highest AUC values ($0.7<AUC>1.0$) and statistically significant variables ($p\leq0.05$) were used to conduct the predictive modeling (Drew *et al.* 2011). In order to receive the highest possible prediction accuracy values, we compared the different sets and orders of predictor variables. Depending on their strength of relationship to the response variable, we selected the best combination of predictors to include in a model. We compared the variable contributions in the models to define the most powerful variables, and their relative magnitude. Spatial autocorrelation (SAC) of the predictor variables was measured using Moran´s Index (Moran, 1950).

3. Results

A total of 3130 mosquitoes were collected from 107 locations across the Taita Hills region. The majority (~2600 mosquitoes from 73 locations), belonged to genus *Culex* and included at least three subgenera: *Culex* (*Culex*), *Cx*. (*Culiciomyia*) and *Cx*. (*Eumelanomyia*). *Culex* mosquitoes were common across the whole study area, including around human settlements and in forests. *Stegomyia* mosquitoes were the second most abundant genus in the collections, (~180 individuals from 28 locations). Larvae of both genera were found in water tanks and small ponds, and adults were collected from houses along the roadsides and in villages. *Stegomyia aegypti* larvae were especially common in water tanks in the villages of Paranga and Kishushe in the Taita Hills north, and also in

car tires in Mwatate village. Both genera were present in lowland and upland areas, including elevations up to 1900 m.

The environmental predictors, elevation, mean precipitation, mean temperature and mean relative humidity, were positively or negatively correlated with each other ($r \geq 0.9$ or $r \geq -0.9$), as were slope and mean radiation ($r \geq -0.75$); therefore, only one of each variable group was retained in the final model. Normalized difference vegetation index, distance to roads, distance to houses, and human population density did not result in high correlations, and were retained in the model process. The combination of non-correlating variables, which resulted in highest evaluation values, was run to obtain reliable estimations for *Culex* and *Stegomyia* (Appendix 1). Mean temperature, mean radiation, elevation, slope and NDVI were influential factors in the models (Appendix 2). Human population density had the greatest effect on the distribution of both *Culex* and *Stegomyia* in all models, but its importance in each model varied (Appendix 2). Moderately explaining variables - NDVI, slope, distance to roads and elevation - were not ranked consistently in the models. The model that best explained *Culex* mosquito distribution included the variables slope, population density, NDVI, distance to roads, and elevation. Mean radiation, NDVI, human population density, distance to roads, and mean temperature were most influential in explaining the distribution of *Stegomyia* mosquitoes.

Spatial autocorrelation (SAC) of the predictor variables was measured using Moran´s Index (Moran, 1950) (Appendix 3). For *Culex*, human population density and elevation were highly spatially autocorrelated (Moran´s $I \geq 0.8$) for short distances ($p < 0.05$) but not at longer distances. Human population density and mean temperature were highly spatially autocorrelated at short distances, but not at longer distances for *Stegomyia*. The models with highest AUC value and significant p-values for *Culex*, were GAM (AUC=0.791) and MARS (AUC=0.809) (Table 2). Altogether, six of the eight models provided reliable estimates for *Culex* mosquito distribution (AUC>0.7). For *Stegomyia*, only two of the eight models resulted in evaluation values considered reliable (AUC>0.7); GBM (AUC=0.708) and RF models (AUC=0.708).

Table 2. Area under the curve **(**AUC) values of all models used to estimate the presence of *Culex* and *Stegomyia* mosquitoes.

Culex		*Stegomyia*	
Model	AUC	Model	AUC
GLM	0.730	GLM	-

GAM	0.791	GAM	0.643
GBM	0.750	GBM	0.708
CTA	0.620	CTA	0.616
ANN	0.764	ANN	0.612
MARS	0.806	MARS	-
RF	0.729	RF	0.708
Maxent	0.585	Maxent	0.690

For the predictive models (described below), we focus on the GAM model for *Culex* and the RF model for *Stegomyia* due to their ability to explain the data reliably and produce the best explaining variable contributions for this study. According to the GAM model, the probability of *Culex* presence was high (≥80%) in locations with NDVI- values ranging from -0.4 to 0.2 (Fig. 3.). The likelihood of *Culex* presence remained high when roads were within 500 m, and at elevations between 800 and 2000 m, in locations with moderate slope angles (0°–35°), and with population densities of 500–6000 people/ km².

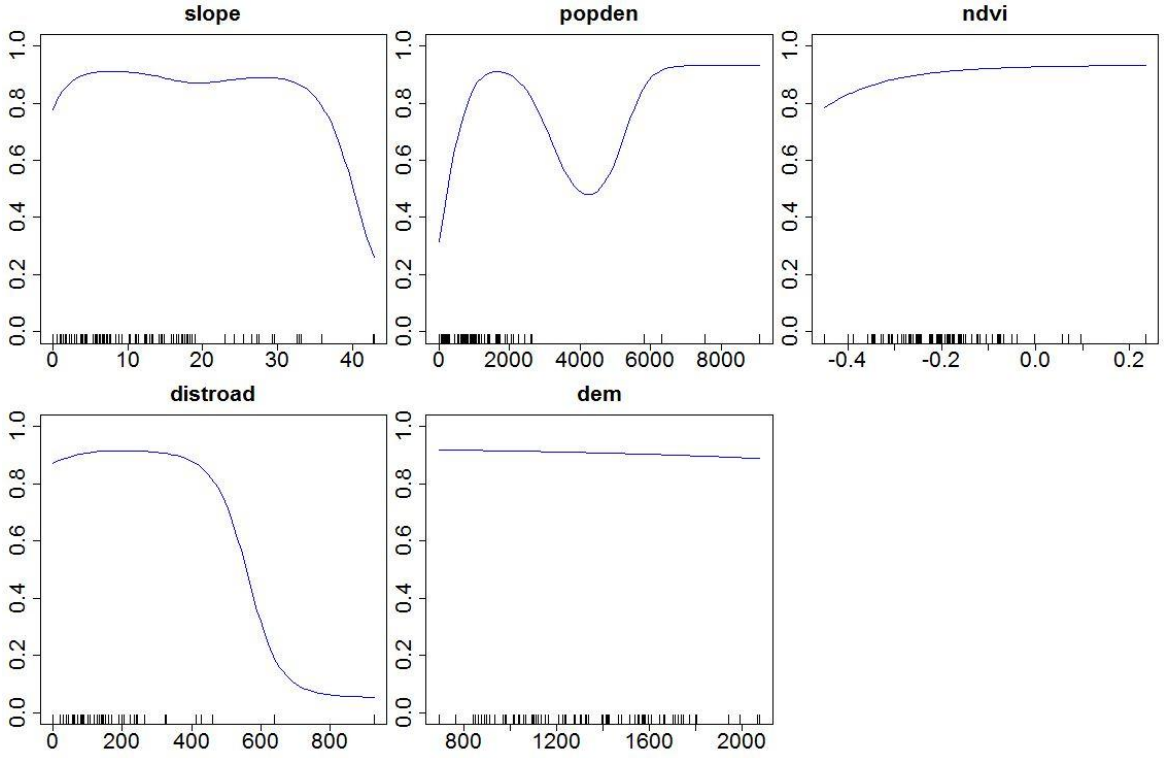


Fig. 3. Response curves for *Culex* mosquito estimations based on the GAM model. The Y-axis represents the probability of presence, and the black marks along the x-axis represent empirical observations.

According to the RF model, the presence of *Stegomyia* was highest in locations with solar radiation levels ≥230 kWh/m² (Fig. 4.). They favored temperatures between 15ºC and 20º C, and over 23ºC,

and locations with poor (NDVI ≤ -0.2) or moderate vegetation (NDVI ≥0.2). All distances between 0 and 600 m from roads resulted in moderate probabilities of *Stegomyia* presence.

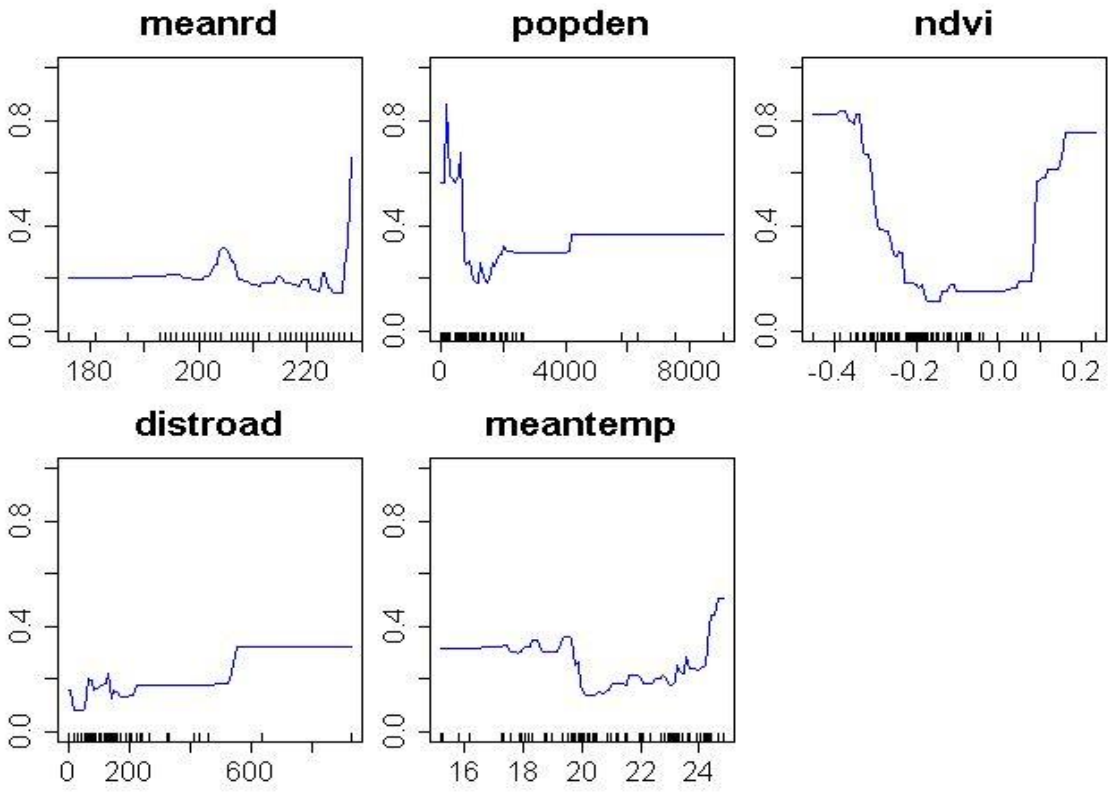


Fig. 4. The response curves of predictors for *Stegomyia* mosquito estimations in RF model. The Y-axis represent the probability of presence, and the black ticks along the x-axis represent empirical observations.

The prediction model risk maps (Figs. 5A and 5B) show that village areas and forests, including elevations above 1500 meters in the Taita Hills, were most suitable for *Culex* mosquitoes (80–100%), whereas the likelihood of their presence was reduced in lowland plateaus, apart from Mwatate village in the South (0–20%). Conversely, the likelihood of *Stegomyia* mosquitoes was high in villages on the plateaus and close to roads (Fig. 5.B.). Areas with high *Stegomyia* mosquito probability were sporadically dispersed.

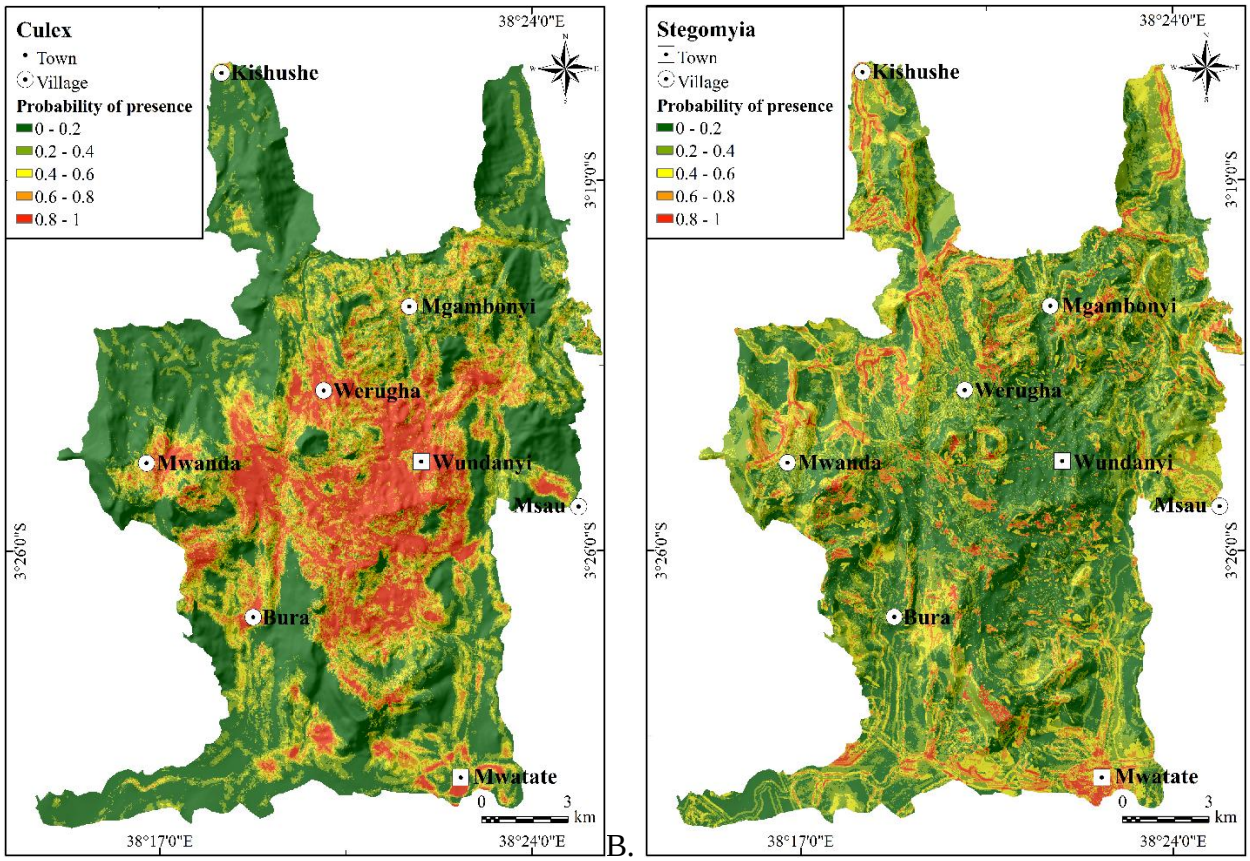


Fig. 5. Predictive risk maps for *Culex* and *Stegomyia* mosquito presence over the study area in the Taita Hills. A. The probability of *Culex* presence was highest in the central part based on the GAM model. B. The probability of *Stegomyia* presence was highest in plateau villages and fragmented areas based on a RF model.

4. Discussion

We show that the Biomod2 ensemble platform in R, was able to predict the distribution of *Culex* and *Stegomyia* mosquitoes in the Taita Hills in Kenya. These two genera were found to overlap in geographical range, but also to be common in different areas based on environmental and anthropogenic factors. *Culex* and *Stegomyia* mosquitoes contain member species that are well-known vectors of significant human pathogens (Lounibos 2002). This approach can be applied to better understand how these vectors, and the infections they carry, vary across space and time, and later, this may be extrapolated to species-level studies and to other understudied areas in rural Africa.

The majority of the models utilized herein produced high AUC values for *Culex* mosquitoes, indicating reliable estimations. We concentrated on the GAM model because it resulted in high AUC values with relevant variable contributions having an automated approach to identify non-linear

relationships, and it may result in more accurate predictions than other model types (Yee *et al.*, 1991). For *Stegomyia* estimations, we focused on the RF model because of its advantages over other models, including high classification accuracy and the ability to model interactions among predictors (Cutler *et al.*, 2007). However, some uncertainties exist concerning *Stegomyia* estimations due to a higher number of absences (94 collections) than presences (28 collections) in the sampling data. Environmental, anthropogenic and distance variables were spatially autocorrelated to some extent, which may, among other things, affect the precision of coefficients (Diniz- Filho *et al.*, 2003).

In our study, the spatial resolution of environmental data was 20 meters. This resolution may have produced some bias to the study results because mosquitoes are also breeding e.g. in tree holes and water tanks in different microclimates. The lack of very high-resolution environmental data (such as the resolutions of 2 meters and 5 meters) is a general problem and not limited just to the Taita Hills. In addition, increasing the resolution of environmental data by deriving them from a high-resolution environmental data does not always improve the species distribution models (Pradervand *et al.*, 2014). Instead, very high-resolution environmental predictors should be produced by taking into account more local field measurements such as fine environmental mapping or in-situ measurements (Pradervand *et al.*, 2014; Lembrechts *et al.*, 2018). The potential solution for the lack of microclimate data could be the use of unmanned aerial vehicles (UAVs) in order to map the sampling areas in real time, and to produce higher resolution environmental data of even 3 centimeters (Kimberly *et al.*, 2014; Anderson *et al.*, 2013). Furthermore, there certainly exist other explanatory factors affecting the mosquito distributions which were not included in the modelling process including landscape fragmentation indicators. Fragmentation variables, such as distance from forest patch, patch size, distance from patch edge and the landscape metric of PPU (patches per unit) have resulted significant statistical relations with mosquito distributions (Richman *et al.*, 2018; Reiter & LaPointe 2007). For example, the use of FRAGSTATS- a spatial pattern analysis program may add value for our future mosquito habitat studies (McGarical *et al.*, 2002).

Consistent with previous research from other global locations (Sallam *et al.*, 2017; Ding *et al.*, 2017), environmental and anthropogenic variables were important determinants of *Culex* and *Stegomyia* mosquito distribution in the Taita Hills. For example the response curve for slope (Fig 3.) shows high probability for *Culex* presence for slopes less than 35°. In the Taita Hills, dwelling units are mainly absent on slopes greater than 35° (Siljander *et al.*, 2011). High probability for *Culex* presence were found with high NDVI values caused by strongly reflecting orchard trees and croplands adjacent to dwelling units in the Taita Hills. Specifically, we found that locations with high population densities,

short distances to roads, and gentle slope angles positively influenced *Culex* presence. Residential and urban areas are often recognized as important for *Culex* distributions (Reiter & LaPointe, 2007; Conley *et al.*, 2014). Rich and poor vegetation and all elevations were suitable for *Culex* mosquito presence, strengthening previous findings of its widespread distribution (MTI, 2017). However, we cannot rule out some effects of potential sample biases (e.g. time of the day and traps involved) which may have affected the mosquito spectrum captured. Sampling bias may also have been the reason for an unexpected response curve for population density (*popden*) at values from ca. 4000 to 5000 (Fig. 3).

Our study suggests, that *Stegomyia* mosquitoes prefer locations with lower human population densities, higher temperatures and solar radiation, and poor or moderate vegetation. This finding somewhat contradicts the notion that *St. aegypti* distribution is linked to growing human population (Fatima *et al.*, 2016), but is consistent with historic records of the variety of *Stegomyia* species suggesting that most members of the *Stegomyia* genus are forest dwelling, with only some adapted to breeding close to human habitation (Powell & Tabachnick 2013). Since our collections included several different *Stegomyia* species - with *St. aegypti* being the only species collected from around human habitation, and other species being restricted to the forest - the model findings are consistent with these historical reports (Powell & Tabachnick 2013). *Stegomyia* favored locations with intermediate and high temperatures, confirming the argument that *Stegomyia* mosquitoes have temperature-based limits to survival (Brady *et al.*, 2013). Overall, anthropogenic and distance factors appeared to be more important than environmental drivers for the distribution of both *Culex* and *Stegomyia* mosquitoes. We note that other variables, not considered in this study, may also affect the distribution of *Culex* and *Stegomyia* mosquitoes and that the genus-level pooling of the species will mask species distribution determinants, which will be targeted in future work with larger and more detailed datasets.

5. Conclusions

Our results affirm the utility and reliability of the Biomod2 package in R as a valid modeling method for species distributions, resulting in insights into vector-ecological interactions. Such work will be further refined with species-level identification and screening for arboviruses. With our study results, general assumptions can be made about the distribution of mosquitoes belonging to *Culex* and *Stegomyia* genera in the Taita Hills and the factors that influence their distribution. High population densities, short distance to roads and gentle slope angles were associated with the occurrence of *Culex* mosquitoes. For *Stegomyia* mosquitoes, on the other hand, low human population densities, high

temperatures and solar radiations as well as poor or moderate vegetation were suitable factors. Together, these results have implications for applying the approach for identifying risk areas and optimizing the use of limited resources for mitigation strategies.

Acknowledgements

We are thankful to Essi Korhonen, Joni Uusitalo, Masika Moses and Peter Mwazi for assistance with the fieldwork, and Miska Luoto, Juha Aalto, Sakari Äärilä and Ninna Malinen for thoughtful discussion on the topic. We also acknowledge the logistical support provided by the Taita Research Station of the University of Helsinki. Funding for this work was provided by the Jane and Aatos Erkko Foundation, the Ministry for Foreign Affairs of Finland TAITAGIS project (http://www.cimo.fi/programmes/hei_ici_index/programmes/hei_ici/projects/taitagis), and the Academy of Finland. The work was carried out under a research permit for Taita Research Station from the National Council for Science and Technology, Kenya (permit NCST/RCD/17/012/33).

References

Adhikari, H., Heiskanen, J., Siljander, M., Maeda, E., Heikinheimo V., & Pellikka, P.K.E. (2017). Determinants of Aboveground Biomass across an Afromontane Landscape Mosaic in Kenya. *Remote Sensing, 9*(8), 827.

Anderson, K., & Gaston, K. (2013). Lightweight unmanned aerial vehicles will revolutionize spatial ecology. *Frontiers in Ecology and the Environment, 11*(3), 138-146.

Bhutta, Z.A., Sommerfeld, J., Lassi, Z.S., Salam, R.A., & Das, J.K. (2014). Global burden, distribution, and interventions for infectious diseases of poverty. *Infectious Diseases of Poverty, 3*, 21.

Broberg, A., Salminen, H., Tolvanen, R., & Ylhäisi, J. (2004). Werugha - village in the heart of the Taita Hills. In Pellikka, P., Ylhäisi, J., & B. Clark (Eds.), *Taita Hills and Kenya, 2004–seminar, reports and journal of a field excursion to Kenya (*pp.108-113). Expedition reports of the Department of Geography. University of Helsinki.

Campbell, L. P., Luther, C., Moo-Llanes, D., Ramsey, J. M., Danis-Lozano, R., & Peterson, A. T. (2015). Climate change influences on global distributions of dengue and chikungunya virus vectors. *Philosophical Transactions of the Royal Society B: Biological Sciences, 370*(1665).

Clark, B., & Pellikka, P. (2005, June). The development of a land use change detection methodology for mapping the Taita Hills, south-east Kenya. CD Rom publication at the meeting of the Proceedings of the 31st International Symposium of Remote Sensing of the Environment, St. Petersburg, Russia.

Climate-Data.org (2018). Monthly average rainfall in Maktau. https://en.climate-data.org/location/103844/ Accessed 30 April 2018.

Conley, A., Fuller, D., Haddad, N., Hassan, A., Gad, A., & Beier, J. (2014). Modeling the distribution of the West Nile and Rift Valley fever vector *Culex pipiens* in arid and semi-arid regions of the Middle East and North Africa. *Parasites & Vectors, 7,* (289).

Crowl, T., Crist, T., Parmenter, R., Belovsky, G., & Lugo, A. (2008). The spread of invasive species and infectious disease as drivers of ecosystem change. *Frontiers in Ecology and the Environment, 6,* 238–246.

Cutler, D.R., Edwards, T.C.Jr., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., & Lawler, J.J. (2007). Random forests for classification in ecology. *Ecology, 88,* 2783–2792.

Ding, F., Fu, J., Jiang, D., Hao, M., & Lin, G. (2017). Mapping the spatial distribution of *Aedes aegypti* and *Aedes albopictus*. *Acta Tropica, 178,* 155–162.

Diniz- Filho, J., Bini, L., & Hawkins, B. (2003). Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology & Biogeography, 12,* 53–64.

Drew, C., Wiersma, Y., & Huettmann, F. (2011). *Predictive species and habitat modeling in landscape ecology*. 319p. The United States.

Dukes, J., Pontius, J., Orwig, D., Garnas, J., Rodgers, V., Brazee, N., Cooke, B., Theoharides, K., Stange, E., Harrington, R., Ehrenfeld, J., Gurevitch, J., Lerdau, M., Stinson, K., Wick, R. & M. Ayres (2009). Responses of insect pests, pathogens, and invasive plant species to climate change in the forests of northeastern North America: What can we predict? *Canadian Journal of Forest Research, 39,* 231–248.

Environment Systems Research Institute (ESRI). (1991). ARC/INFO User's guide. Cell-based modelling with GRID. Analysis, display and management (Redlands, CA: ESRI). European Spatial Agency (ESA). (2015). Sentinel-2 user handbook. *ESA Standard Document*, *64*.

Erdogan, E.H., Pellikka, P., & Clark, B. (2011). Modelling the impact of land-cover change on potential soil loss in the Taita Hills, Kenya, between 1987 and 2003 using remote-sensing and geospatial data. *International Journal of Remote Sensing*, *32*(21), 5919–5945.

Fatima, S., Atif, S., Rasheed, S., Zaidi, F., & Hussain, E. (2016). Species distribution modeling of *Stegomyia aegypti* in two dengue-endemic regions of Pakistan. *Tropical Medicine and International Health, 21,* 427–436.

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

Fornace, K., Drakeley, C., William, T., Espino, F. & Cox, J. (2014). Mapping infectious disease landscapes: unmanned aerial vehicles and epidemiology. *Trends in Parasitology, 30*(11), 514-519.

Franklin, J., & Miller, J. (2010). Statistical methods – modern regression. In: Franklin, J.(ed.) *Mapping species distribution: spatial inference and prediction*. Cambridge: Cambridge University Press, (pp. 340).

Guisan, A., & Zimmermann, N. (2000). Predictive habitat distribution models in ecology. *Ecological Modeling, 135*(2–3), 147–186.

Guisan, A., Tingley, R., Baumgartner, J., Naujokaitis-Lewis, I., Sutcliffe, P., Tulloch, A., Regan, T., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T., Rhodes, J., Maggini, R., Setterfield, S., Elith, J., Schwartz, M., Wintle, B., Broennimann, O., Austin, M., Ferrier, S., Kearney, M., Possingham, H., & Buckley, Y. (2013). Predicting species distributions for conservation decisions. *Ecology Letters, 16*(12), 1424–1435.

Hastie, T., Tibshirani, R., & Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association, 89*(428), 1255–1270.

Huang, Y-M. (1986). *Aedes* (*Stegomyia*) *bromeliae* the yellow fever virus vector in East Africa. *Journal of Medical Entomology*, 23(2), 196–200.

Huang, Y-M. (1990). The subgenus *Stegomyia* of *Aedes* in the Afrotropical Region I. The Africanus group of species (Diptera: Culicidae). *Contributions of the American Entomological Institute, 26*(1), 1–90.

Huang, Y-M. (2001). A pictorial key for the identification of the subfamilies of Culicidae, genera of Culicinae, and subgenera of *Aedes* of the Afrotropical Region (Diptera: Culicidae). *Proceedings of the Entomological Society of Washington, 103*, 1–53.

Huang, Y-M. (2004). The subgenus *Stegomyia* of *Aedes* in the Afrotropical Region with keys to the species (Diptera: Culicidae). *Zootaxa, 700*(1), 1–120.

Hutchinson, M.F. (1991). The application of thin plate splines to continent-wide data assimilation. In Data Assimilation Systems. J.D. Jasper (Ed.). Melbourne: Bureau of Meteorology. *BMRC Research Report, 27*, 104–113.

Kaplan, I., Dobert, M.K., Marvin, B.J., McLaughlin, J.L., & Whitaker, D.P. (1976). *Area Handbook for Kenya*. Foreign Area Studies (FAS). The American University, Washington D.C., U.S.A, (pp. 472).

Lembrechts, J., Nijs, I., & Lenoir, J. (2018). Incorporating microclimate into species distribution models. *Ecography*.

Lounibos, L.P. (2002). Invasions by insect vectors of human disease. *Annual Review of Entemology, 47*, 233–266.

McGarigal, K., Cushman, S., Neel, M., & Ene, E. (2002). FRAGSTATS: Spatial pattern analysis program for categorical maps. Computer software program produced at the University of Massachusetts, Amherst. https://www.umass.edu/landeco/research/fragstats/fragstats.html/ Accessed 7 November 2018.

Moran, P (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37, 17–23.

Mosquito taxonomic inventory (MTI). (2017). *Anopheles*; *Culex*; *Stegomyia*. http://mosquito-taxonomic-inventory.info/simpletaxonomy/term/8577/ Accessed 5 April 2017.

Mughini-Gras, L., Mulatti, P., Severini, F., Boccolini, D., Romi, R., Gioia, B., Khoury, C., Bianchi, R., Montarsi, F., Patregnani, T., Bonfanti, L., Rezza, G., Capelli, G., & Busani, L. (2014). Ecological niche modelling of potential West Nile virus vector mosquito species and their geographical association with equine epizootics in Italy. *EcoHealth, 11*, 120.

Neteler, M., Roiz, D., Rocchini, D., Castellani, C., & Rizzoli, A. (2011). Terra and Aqua satellites track tiger mosquito invasion: modeling the potential distribution of *Aedes albopictus* in north-eastern Italy. *International Journal of Health Geographics, 10* (49).

Paupy, C., Delatte, H., Bagny, L., Corbel, V. & Fontenille, D. (2009). *Aedes albopictus*, an arbovirus vector: from the darkness to the light. *Microbes and Infection*, *11*, 1177–1185.

Pellikka, P. K. E., Heikinheimo, V., Hietanen, J., Schäfer, E., Siljander, M. & Heiskanen, J. (2018). Impact of land cover change on aboveground carbon stocks in Afromontane landscape in Kenya. *Applied Geography*, *94*, 178–189.

Powell, J. R., & Tabachnick, W. J. (2013). History of domestication and spread of *Aedes aegypti* - A Review. *Memórias Do Instituto Oswaldo Cruz*, *108*(1), 11–17.

Pradervand, J., Dubuis, A., Pellissier, L., Guisan, A., & Randin, C. (2014). Very high resolution environmental predictors in species distribution models: Moving beyond topography? *Progress in Physical Geography*, *38*, 79-96.

R Core Team. (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. URL http://www.R-project.org/.

Reinert, J. F., Harbach, R. E., & Kitching, I. J. (2009). Phylogeny and classification of tribe Aedini (Diptera: Culicidae). *Zoological Journal of the Linnean Society*, *157*, 700–794.

Reiter, M., & LaPointe, D. (2007). Landscape factors influencing the spatial distribution and abundance of mosquito vector *Culex quinquefasciatus* (Diptera: Culicidae) in a mixed residential-agricultural community in Hawai ‘i. *Journal of Medical Entomology*, *44*(5), 861–868.

480 Richman, R., Diallo, D., Diallo, M., Sall, A., Faye, O., Diagne, C., Dia, I., Weaver, S., Hanley, K., &
481 Buenemann, M. (2018). Ecological niche modeling of Aedes mosquito vectors of chikungunya
482 virus in southeastern Senegal. *Parasites & vectors*. *11*(1), 255.
483 Roiz, D., Neteler, M., Castellani, C., Arnoldi, D., & Rizzoli, A. (2011). Climatic factors driving
484 invasion of the tiger mosquito (*Aedes albopictus*) into new areas of Trentino, Northern Italy.
485 *PLoS ONE*, *6*(4), e1480.
486 Sallam, M., Xue, R., Pereira, R., & Koehler, P. (2016). Ecological niche modeling of mosquito
487 vectors of West Nile virus in St. John's County, Florida, USA. *Parasites & Vectors, 9*, 371.
488 Sallam, M.F., Michaels, S.R., Riegel, C., Pereira, R.M., Zipperer, W., Lockaby, B.G., & Koehler,
489 P.G. Spatio-Temporal Distribution of Vector-Host Contact (VHC) Ratios and Ecological Niche
490 Modeling of the West Nile Virus Mosquito Vector, *Culex quinquefasciatus,* in the City of New
491 Orleans, LA, USA. *International Journal of Environmental Research and Public Health*, *14*(8),
492 892.
493 Service, M. (1991). *Mosquito ecology: Field sampling methods.* (2nd edition). 988 p. Elsevier.
494 Siljander, M., Clark, B., & Pellikka, P. (2011). A predictive modelling technique for human
495 population distribution and abundance estimation using remote sensing and geospatial data in a
496 rural mountainous area in Kenya. *International Journal of Remote Sensing, 32(21),* 5997–6023.
497 Thuiller, W., Georges, D., Engler, R., & Breiner, F. (2016). Package "biomod2". Ensemble platform
498 for species distribution modeling. Version 3.3-7. https://cran.r-
499 project.org/web/packages/biomod2/ biomod2.pdf/ Accessed 5 November 2016.
500 Thuiller, W., Lafourcade, B., Engler, R., & Araújo, M. (2009). BIOMOD - A platform for ensemble
501 forecasting of species distributions. *Ecography*, *32*, 369–373.
502 Tucker, C. (1979). Red and photographic infrared linear combinations for monitoring vegetation.
503 *Remote Sensing of Environment, 8*, 127–150. University of Helsinki (2003-2009). The Taita
504 Hills. http://www.helsinki.fi/science/taita/taitahills.html/ Accessed 22 April 2016.
505 Virtanen, E. (2015). *Fine-resolution climate grids for species studies in data-poor regions.* Master´s
506 Thesis. University of Helsinki.
507 The World Health Organization (WHO). (2017). Vector-borne diseases. October 2017. Fact sheets.
508 http://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases/ Accessed 6 June 2018.
509 Yee, T., & Mitchell, N. (1991). Generalized additive models in plant ecology. *Journal of Vegetation*
510 *Science, 2*, 587–602.
511
512
513

Appendix 1. Multicollinearity of the variables included in the models. All statistical significance values (p-values), are marked as follows: statistically highly significant *** = <0.001, statistically significant ** = <0.01, statistically significant *= <0.05 and ns = not statistically significant.

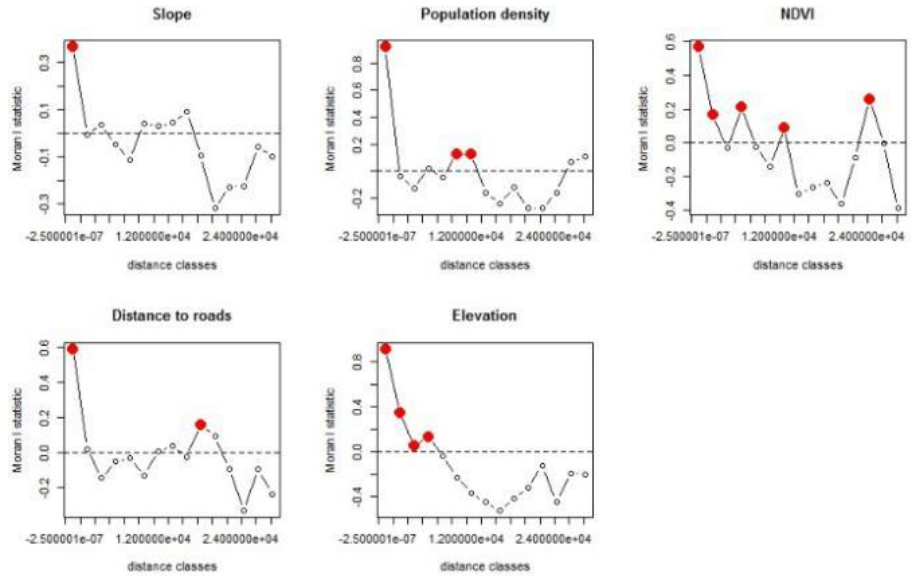
Culex					
Variable	Slope	NDVI	Population density	Distance to roads	Elevation
Slope		0.115ns	-0.294**	0.172ns	0.279**
NDVI	0.115ns		-0.226*	0.391***	0.557***
Population density	-0.294**	-0.226*		-0.26**	-0.313***
Distance to roads	0.172ns	0.391***	-0.26**		0.143ns
Elevation	0.279**	0.557***	-0.313***	0.143ns	
Stegomyia					
Variable	Mean radiation	NDVI	Population density	Distance to roads	Mean temperature
Mean radiation		-0.04ns	0.312***	-0.12ns	0.10ns
NDVI	-0.04ns		-0.226*	0.391***	-0.543***
Population density	0.313***	-0.226*		-0.26**	0.286**
Distance to roads	-0.12ns	0.391***	-0.26**		-0.12ns
Mean temperature	0.10ns	-0.543***	0.286**	-0.12ns	

Appendix 2. Variable importance presented in each model for *Culex* and *Stegomyia* estimates.

Culex								
Variable	GLM	GAM	GBM	CTA	ANN	MARS	RF	Maxent
Slope	0	0.147	0.078	0	0.147	0.385	0.200	0.432
Population density	0.658	0.764	0.678	1	0.773	0.702	0.429	0.513
NDVI	0.261	0.165	0.088	0	0	0	0.114	0.382
Distance to roads	0	0.179	0.039	0	0.092	0	0.064	0.453
Elevation	0	0.028	0.006	0	0.097	0	0.030	0.436
Stegomyia								
Variable	GLM	GAM	GBM	CTA	ANN	MARS	RF	Maxent
Mean radiation	-	0.041	0	0	0.289	-	0.059	0.176
Population density	-	0.575	0.680	0.966	1	-	0.206	0.357
NDVI	-	0.388	0	0	0	-	0.210	0.291
Distance to roads	-	0.337	0	0	0.215	-	0.087	0.028
Mean temperature	-	0	0.358	0	0.081	-	0.089	0.026

Appendix 3. A. SAC of *Culex* predictors. B. SAC of *Stegomyia* predictors. Red circles indicated a significant p-value ($p < 0.05$) and indicate distances where the variable was autocorrelated.

A.



B.