Transferability of tsetse habitat models between different regions in Kenya and Rwanda

Stella Gachoki [a,b,*], Anton Vrieling [b], Thomas A. Groen [b], Andrew Skidmore [b], Daniel Masiga [a]

[a] *International Centre of Insect Physiology and Ecology (icipe), P.O. Box 30772-00100 Nairobi, Kenya*
[b] *University of Twente, Faculty of Geo-information Science and Earth Observation (ITC), P.O. Box 217, 7500 AE Enschede, The Netherlands*

ABSTRACT

Accurate and reliable information on the distribution of tsetse habitats is crucial for the effective management of African Trypanosomiasis in sub-Saharan Africa. However, conducting large-scale surveillance of tsetse flies to develop distribution maps is impractical due to vast areas infested and limited resources available. To address this challenge, we evaluated the applicability of tsetse habitat models developed in the intensively sampled Shimba Hills National Reserve in Kenya for both the wet and the dry season, to two other regions in Kenya (Ruma National Park and Nguruman Conservancy) and one region in Rwanda (Akagera National Park). The models utilized satellite-based estimates of vegetation greenness, land cover, and land surface temperature, combined with tsetse occurrence data, to predict habitat suitability. An independent dataset of tsetse occurrence was used to benchmark the performance of the transferred models. The performance of the transferred models was significantly influenced by the similarity in environmental conditions between the model's development area and the transfer area. In regions with high dissimilarity, such as Nguruman Conservancy during the dry season, model transfer was unsuccessful with an F1-score of zero. In all other regions and seasons, the transferred models showed satisfactory performance, with F1-score values exceeding 0.65. Nevertheless, site-specific models out-performed (>0.8 F1-score) the transferred models, indicating that models specifically developed with data for each location can provide more accurate information on tsetse distribution. In conclusion, our study demonstrates that tsetse habitat models can be transferred with relatively good accuracies to seasons and regions that exhibit environmental similarity with the model training area. Despite the higher accuracy of site-specific models, transferring models to similar sites remains a meaningful exercise in the absence of detailed surveillance data.

1. Introduction

Tsetse flies are spatially confined to sub-Saharan Africa (Rogers and Wint, 2000), and are the sole cyclical vectors of African trypanosomiasis (AT) that cause debilitation in humans (HAT) and reduce the productivity of livestock (AAT). Since there are currently no available vaccines for African trypanosomiasis, the most promising control strategy is to disrupt disease transmission by reducing the population of tsetse flies (Cattand et al., 2006). Although various tsetse control strategies have been employed, such as the use of insecticides for spraying and insecticide-treated targets and traps (Munga and Wanga, 2014; Okello et al., 2021; Percoma et al., 2018), the effective implementation of these methods is often limited by inadequate or unreliable information about

tsetse habitats outside of monitored areas.

Accurate mapping of tsetse habitats is crucial for the effective implementation of tsetse control strategies. However, in most regions of sub-Saharan Africa, the available maps were developed at a low spatial resolution of 5 km by 5 km between 1999 and 2003 (Wint, 2003). Local research has revealed significant discrepancies between those maps and the actual presence of the fly (Cecilia et al., 2021; Esterhuizen et al., 2005; Gachoki et al., 2021), raising questions about their reliability. Large-scale tsetse surveillance using traps is costly and time-consuming, making it hard to gather detailed ground data for predicting suitable habitats for tsetse flies in Africa. To address this challenge, a promising approach is to identify the environmental conditions suitable for different tsetse species in heavily sampled areas and use this information

to map their habitats in other regions. Remote sensing of environmental variables, which can explain and predict species occurrence, provides a potential opportunity to transfer models to areas with little or no ground data, allowing assessment of likely tsetse presence over large areas (Bradley and Fleishman, 2008; He et al., 2015). Such larger-scale assessment can be valuable for conservation management to avoid the misallocation of resources (Duque-Lazo et al., 2016; Manzoor et al., 2018), or to identify priority areas for management actions.

The successful transfer of species distribution models (SDMs) depends on several critical factors such as; (1) environmental equilibrium (i.e., the species occur in all suitable areas and are absent in unsuitable ones; Araújo and Pearson 2005; Nogués-Bravo 2009), (2) environmental similarity between the training and target sites (Sequeira et al., 2018; Yates et al., 2018), (3) use of ecologically relevant predictor variables (Dormann et al., 2007; Peres-Neto and Legendre, 2010; Werkowska et al., 2017), and (4) the modeling strategy employed. Simple models, such as Generalized Linear Models (GLMs), are generally considered to transfer more effectively (Austin et al., 2006; Duque-Lazo et al., 2016; Meynard et al., 2007; Randin et al., 2006; Wenger and Olden, 2012) compared to complex models such as Random Forest (Jiménez-Valverde et al., 2008; Tsoar et al., 2007), which are more prone to overfitting.

While transferring tsetse habitat models between regions has not yet been attempted, there have been previous attempts for other plant and animal species with varying degrees of success. For example, fish distribution models transferred between two areas in southwest Finland had good Area Under Curve (AUC) accuracies (0.75–0.93) based on an independent test dataset (Sundblad et al., 2009), while transfer of models for the distribution of the fungus *Phytophthora cinnamomi* between southwest Spain and southwest Australia had marginal AUC (>0.55) values (Duque-Lazo et al., 2016). On the other hand, transfer of Douglas fir distribution models from North America to Europe (Boiffin

et al., 2017) and spotted hyena habitat models from SSA to Europe (Varela et al., 2009) were unsuccessful as the predictions showed suitable habitats for both species were not present. Recently, Gachoki et al. (2021) developed season-specific models to identify tsetse habitats for *Glossina pallidipes* species around Shimba Hills National Reserve, in Kenya. The models developed by Gachoki et al. (2021) were based on tsetse data sampled between 2017 and 2019, and incorporated ecologically-relevant predictor variables, such as landcover, which aimed to capture *G. pallidipes* habitat preferences and accounted for their movement. The GLM model outperformed other modeling techniques used, with AUC values of 0.79 in the dry season and 0.77 in the wet season (Gachoki et al., 2021).

The objective of this study was to evaluate whether *G. pallidipes* habitat models, previously developed by Gachoki et al. (2021), could effectively identify tsetse habitats in three other regions where this tsetse species occur (Akagera National Park in Rwanda and Nguruman Conservancy and Ruma National Park in Kenya). We hypothesized that the transfer of the tsetse habitat model to areas exhibiting similar environmental conditions to Shimba Hills National Reserve will result in comparable accuracies while dissimilar areas will have lower accuracies.

2. Material and methods

2.1. Study area and tsetse data

To test the transferability of the habitat suitability model for Shimba Hills National Reserve (reference site), three study areas (target sites) were selected: Ruma National Park (RNP; Fig. 1b) and Nguruman Conservancy (NC; Fig.1c) in Kenya, and Akagera National Park (ANP; Fig. 1d) in Rwanda. *Glossina pallidipes* is the most abundant tsetse
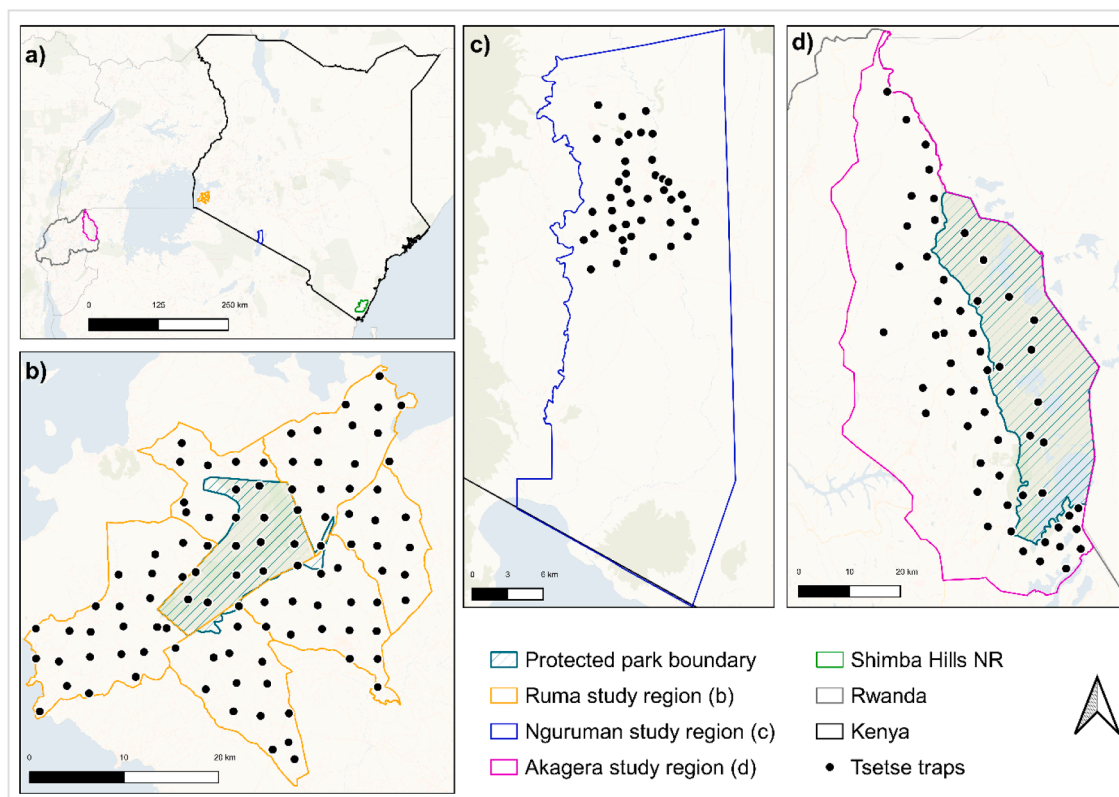
Fig. 1. Location of the study areas. a) Overview map of all the study sites; b) Study extent around Ruma National Park in Kenya; c) study extent around Nguruman Conservancy (no official boundary for the extent of the conservancy) in Kenya; d) study extent around Akagera National Park in Rwanda. The background image is the Voyager (no labels) map as provided in QuickMapServices plugin in Quantum-GIS (QGIS) software.

species in these four areas. These regions are also home to a variety of wild animals, including warthogs and buffaloes, which are known to be frequently fed on by *G. pallidipes*. Nguruman Conservancy is the only region where wild and domesticated animals intermingle freely; the other three regions are fenced and have limited livestock grazing inside the parks. Based on worldClim V2 climatic data (Fick and Hijmans, 2017), annual precipitation shows similar spatial variability in three of the four sites: ANP, RNP, and SHNR. The landscape structure of these three sites is mainly woodlands with patches of grasslands within the park and crop farming outside the park. On the other hand, Nguruman Conservancy exhibits contrasting characteristics. It consists of extensive bare lands with limited crop farming and experiences lower annual precipitation ranging from 300 to 700 mm. Akagera and Ruma National Park share relatively similar average temperature ranges (18–23 ℃), while Nguruman Conservancy and Shimba Hills National Reserve have higher temperatures (21–26 ℃). Table 1 provides a summary of the environmental and climatic conditions across the four study sites.

2.2. Predictor variables and similarity analysis

To ensure consistency with the reference model (Gachoki et al., 2021), we generated identical predictor variables for the target sites for every season during which tsetse data were available. These variables included:

  i the Normalized Difference Vegetation Index (NDVI) and the Modified Normalized Difference Water Index (MNDWI) as estimated from spectral reflectance data obtained from the Sentinel-2 satellites,
 ii Land Surface Temperature (LST) derived from Landsat 8 Optical Land Imager
iii the Topographic Wetness Index (TWI; Mattivi et al. (2019)) and slope as estimated from a digital elevation model obtained from the Shuttle Radar Topography Mission (SRTM)
 iv land cover fractions extracted from 10 m resolution land cover maps that we developed using random forest classification of Sentinel-2 satellites spectral bands whereby training data was obtained from visual interpretation of very high-resolution Google Earth imagery. In order to account for the changes in leafy vegetation between the dry and wet seasons, we generated seasonal land cover maps.
  v silt content obtained from the soil property maps obtained from Hengl et al. (2021) at 30 m resolution.

The data sources used were similar to those utilized by Gachoki et al. (2021), except for silt content which in prior study was estimated based on interpolation of our own in-situ point measurements of silt content. Gachoki et al. (2021) provided a detailed explanation of why each predictor variable was included and how they were derived.

To assess the environmental similarity between the target and reference sites, we utilized the Mahalanobis distance, which measures the distance between a point and a distribution in a multi-dimensional "environmental space" using the environmental variable values as coordinates (McLachlan, 1999). To apply this to our study, we first extracted the predictor variables for each target site at each trapping location across different seasons. Afterwards we calculated the Mahalanobis distance of each target site using the reference site variables to define the mean and standard deviation from which the trap locations of the target sites could deviate from. A lower Mahalanobis distance indicates a greater similarity between the target and the reference site. To gain a deeper understanding of each predictor variable, we performed a one-way Analysis of Variance (ANOVA) and used the Tukey's Honest Significant Difference (Tukey HSD) test to determine if the variables obtained from the target sites were significantly different from those obtained in the reference site. This analysis was conducted using the *stats* package in the R programming language (R Core Team, 2020).

2.3. Model transfer and performance evaluation

The tsetse habitat models developed by Gachoki et al. (2021) were applied to predictor variables obtained from target sites. To assess the performance of these transferred models, we utilized *in-situ* tsetse data from each target site and calculated the F1-score (Eq. (1)) values for each of the resulting model predictions using the maximum sum of sensitivity and specificity value (0.55) from Gachoki et al. (2021) as the cut off point for suitable and unsuitable sites.

$$F1\ score = \frac{2*True\ positive}{2*True\ positive + false\ positive + false\ negative} \quad \text{(Eq 1)}$$

In model transfer, it is assumed that the relationships between the predictors and the target species presence are consistent across different sites. To confirm whether this is the case for *G. pallidipes*, besides transferring the model to the target sites, we also created site-specific GLM models for each target site following the same procedures used by Gachoki et al. (2021). For the site-specific models we also calculated the F1-scores and compared them to those calculated from the transferred models. Due to the imbalanced nature of our target sites data, with varying numbers of presence and absence points, we further evaluated model performance by generating sensitivity and specificity curves. These curves allowed us to examine the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) at different classification thresholds. Additionally, we generated variable importance plots based on AUC tests to measure the contribution of each variable in explaining the distribution of tsetse flies. We also generated response curves for the site-specific models and compared them to those obtained from the reference model. In this way, we

Table 1
Summary of the environmental and climatic conditions for the four sites.

	Akagera	Nguruman	Ruma	SHNR
Annual precipitation (mm)	800–1100	300–700	900–1800	800–1400
Average temperature (℃)	19–29	28–32	20–28	24–26
Altitude	1220 - 4507	600–2300	1200–1600	120–450
Landcover	Woodlands, grasslands, bushlands, swamp	Woodlands, grasslands, bare land, patches of forest	Woodlands, grasslands	Woodlands, grasslands, patches of forest
Tsetse species	*G. pallidipes, G. morsitans*	*G. pallidipes, G. longipenis*	*G. pallidipes, G. fuscipes*	*G. pallipides, G. brevipalpis G. austeni*
Size of the study area (km²)	3409	647	207	1173
Tsetse traps	55 biconical traps (Gashururu et al., 2021)	40 Ngu traps	111 Ngu and biconical traps	260 biconical traps (Gachoki et al., 2021)
Temporal availability	June-August 2028 (dry) March-May 2019 (wet)	May 2021 (wet) November 2021 (dry)	November 2021 (wet)	Various months (2017–2019) January -April (dry) and May-December (wet)

analyzed how different the transferred model performed from a model directly trained with site data.

3. Results

3.1. Environmental similarity

Predictor variables observed in Akagera National Park during the dry season, were the most comparable environmental conditions to Shimba Hills National Reserve (lowest Mahalanobis values) while Nguruman exhibited the greatest dissimilarity in both seasons (Fig. 2n). The predictor variable values for the Nguruman Conservancy for the dry season were significantly different from those in Shimba Hills while in Akagera National Park only the grassland fractions and median LST significantly differed (Fig. 2c and i). In the wet season and for all target sites, the woodland fractions significantly differed from those observed in Shimba Hills. These findings show that some predictor variables from the target site display significant differences with those of the reference site and this may affect the performance of the transferred model.

3.2. Transferred model accuracy

The transferred models predicted a greater extent to be suitable for tsetse fly occurrence in Akagera National Park and Nguruman Conservancy during the wet season, as opposed to the dry season (Fig. 3). The presence of abundant green leaves in woodlands during the wet season,

which provide more shade required by tsetse flies, may explain the abundance of suitable habitats during this season. In Ruma National Park, the transferred models tended to overestimate tsetse habitats beyond the known areas (Fig. 3; within the park and along the shores of Lake Victoria). However, habitat suitability is not synonymous with the actual presence of the fly but rather suggests that environmental conditions can support its existence.

The F1 score analysis revealed that the transfer of Shimba Hills habitat models to Nguruman Conservancy during the dry season (most dissimilar site; Fig. 2n) was unsuccessful, resulting in a F1 score of 0 (Fig. 4). For all other target sites and periods, the F1 score values were above 0.6. Site-specific models outperformed the transferred models (Fig. 4). These results imply that developing local models for a particular site is likely to provide more accurate information compared to models transferred from a different region.

The sensitivity and specificity analysis showed that using the optimal threshold of 0.55 derived from Shimba Hills National Reserve models did not consistently yield optimal results when applied to the target sites (Fig. 5k and l; dotted black line). For the transferred models, this optimal threshold yielded reasonable accuracy only for Ruma National Park during the wet season (Fig. 5j), whereas for other target sites, optimal performance was mostly achieved at lower thresholds (e.g., Nguruman Conservancy in the dry season; Fig. 5d). However, it is important to acknowledge that threshold optimization is not feasible in areas lacking sampled locations. Therefore, in this study, the sensitivity and specificity curves for the transferred model are solely utilized to
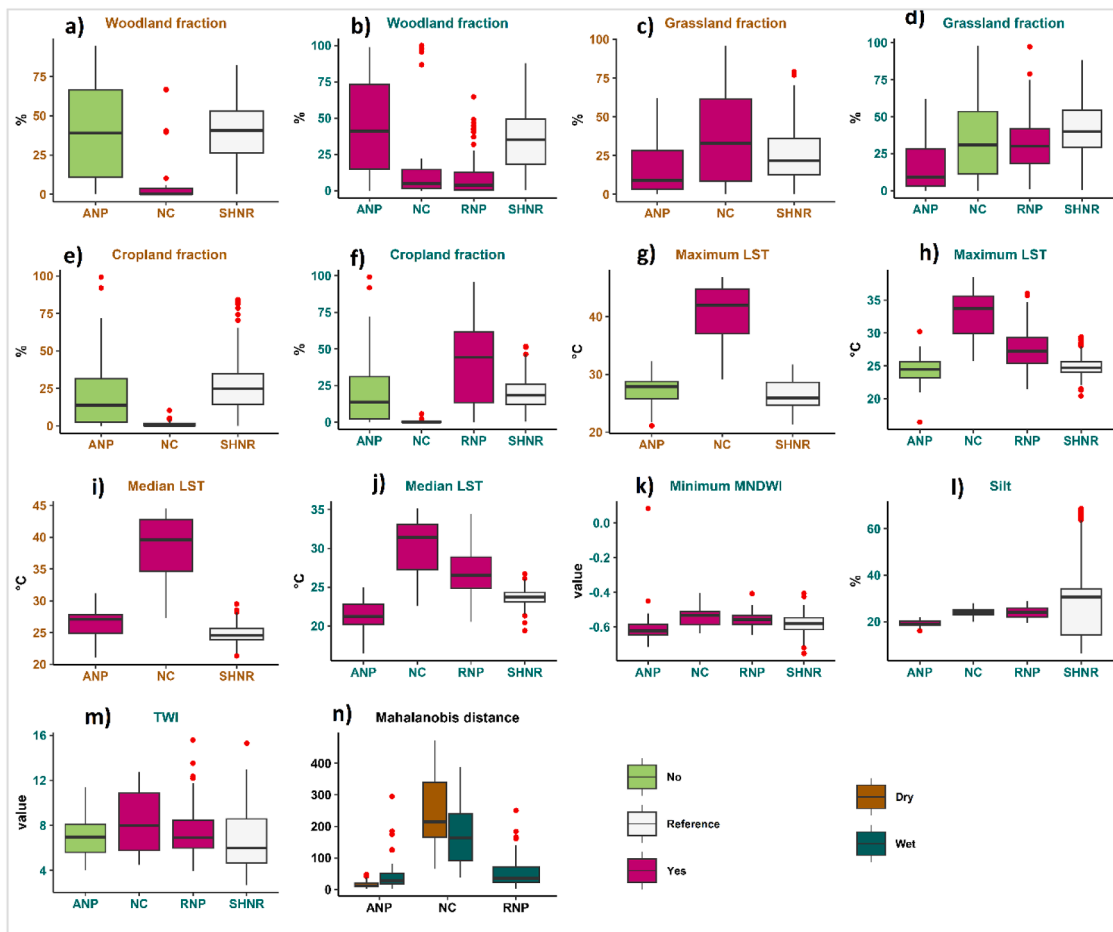
Fig. 2. Environmental similarity analysis between the target sites and the reference site. ANP = Akagera National Park, NC = Nguruman Conservancy, RNP = Ruma National Park. SHNR = Shimba Hills National Reserve. The colored text on the y and x axes represents different seasons; brown = dry season and dark green = wet season. Boxplots with a light green color indicate that the predictor variable was not significantly different from the corresponding variable observed in SHNR ($p >$ 0.05) while boxplots in pink indicate that these variables were significantly different from those observed in SHNR ($p \leq 0.05$) according to Tukey HSD test.
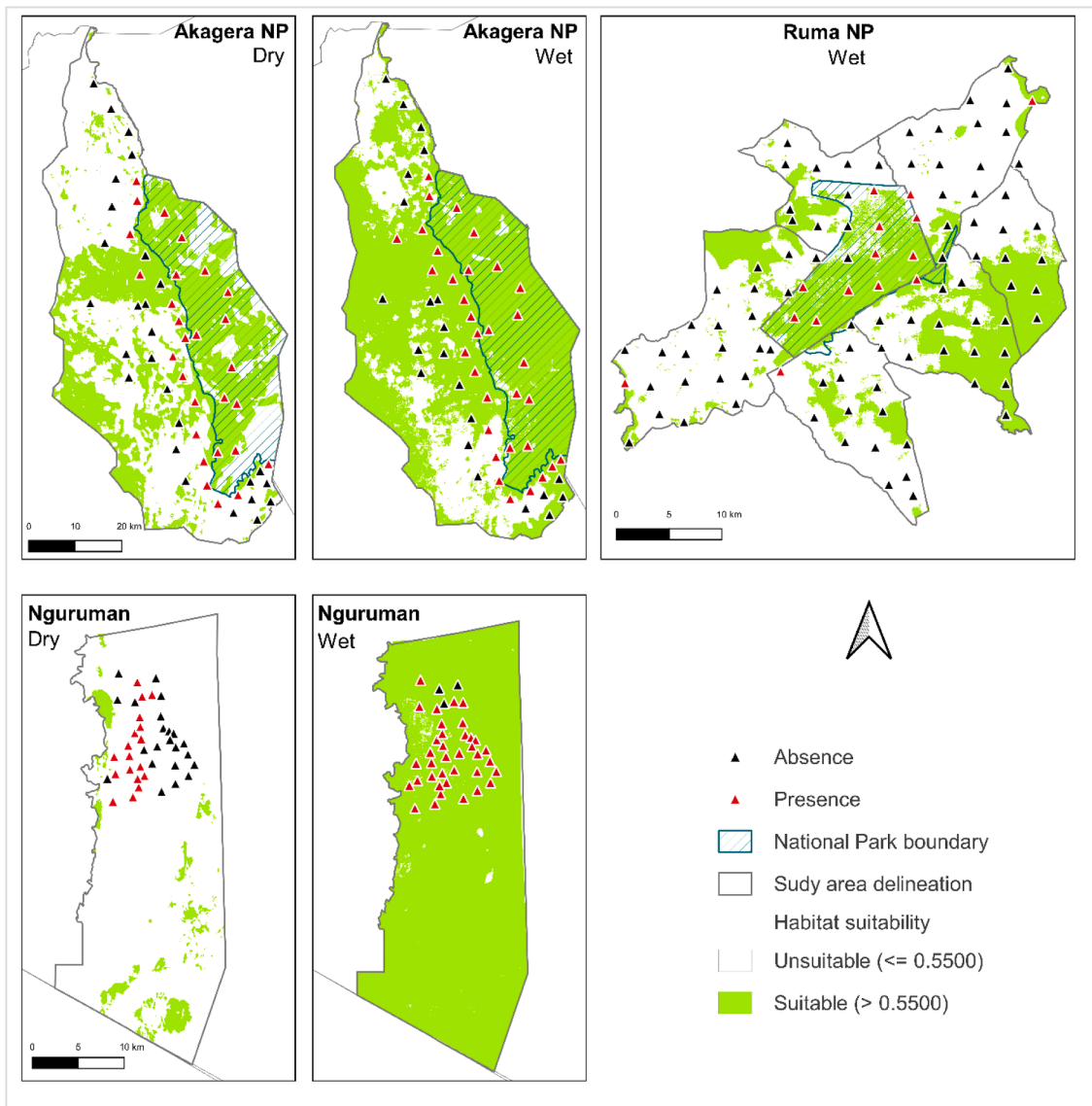
Fig. 3. Predicted tsetse suitability maps of the target sites. The classification of the suitable and unsuitable sites was based on the optimal maximum sum of sensitivity and specificity thresholds as calculated from the reference model and this value was 0.55 for both seasons.

elucidate the underlying reasons for the observed results. Overall, the accuracy of the site-specific models consistently surpassed that of the transferred models in all regions.

Significant variations were observed in the key factors influencing tsetse occurrence across the four study locations (Fig. 6). During the dry season, tsetse occurrence in Shimba Hills was mainly influenced by woodland and cropland fractions, while for Akagera National Park woodland fraction was most important, and for the Nguruman Conservancy it was median LST (Fig. 6**A**). During the wet season, the primary factors influencing tsetse numbers were grassland and cropland fractions in Shimba Hills, woodland fraction in Ruma National Park, and maximum land surface temperature (LST) in both Nguruman Conservancy and Akagera National Park (Fig. 6**B**).

The croplands (Fig. 7**a**) and woodlands (Fig. 7**b**) fractions in the dry season consistently showed positive and negative correlations, respectively, with the occurrence of tsetse flies in all locations. However, the association between the other variables and tsetse flies differed across various sites in both seasons. In contrast to what was expected based on the known behavior of tsetse flies, cropland fraction (Fig. 7**f**) in Akagera National Park showed a surprising positive relationship with tsetse flies

during the wet season.

4. Discussion

The purpose of this study was to investigate the transferability of seasonal tsetse habitat models developed in the intensively-sampled Shimba Hills National Reserve (SHNR) in Kenya to other sparsely sampled regions in Kenya (Ruma National Park; RNP and Nguruman Conservancy; NP) and Rwanda (Akagera National Park; ANP). Visually, the predicted maps appeared to over-predict tsetse habitats during the wet season and underpredict similar habitats in the dry season (Fig. 3). The F1 scores achieved by the transferred models had F1-score values of greater than >0.65 for all target sites and seasons, except for Nguruman Conservancy during the dry season where the transfer was unsuccessful (F1 score = 0). Notably, Nguruman Conservancy during the dry season displayed the highest dissimilarity when compared to the reference site (SHNR), confirming our hypothesis that model transfer accuracy tends to decrease as environmental dissimilarity increases. The site-specific models consistently demonstrated higher F1 scores (>0.80) compared to the transferred models (Fig. 4) suggesting that location-specific
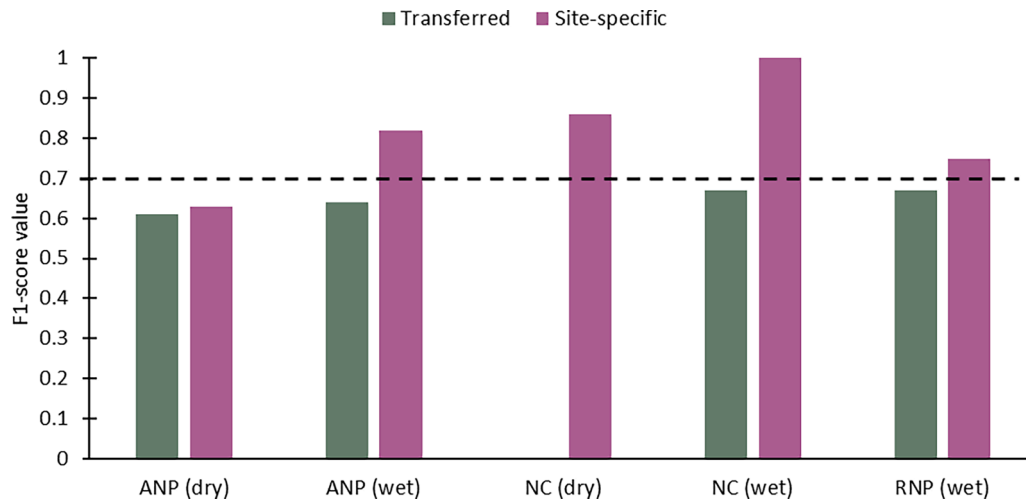
Fig. 4. The F1 scores for the transferred and the site-specific models. The dotted black line indicates the F1 score for Shimba Hills National Reserve in the dry and wet season (for both 0.7). ANP = Akagera National Park, NC = Nguruman Conservancy, RNP = Ruma National Park.
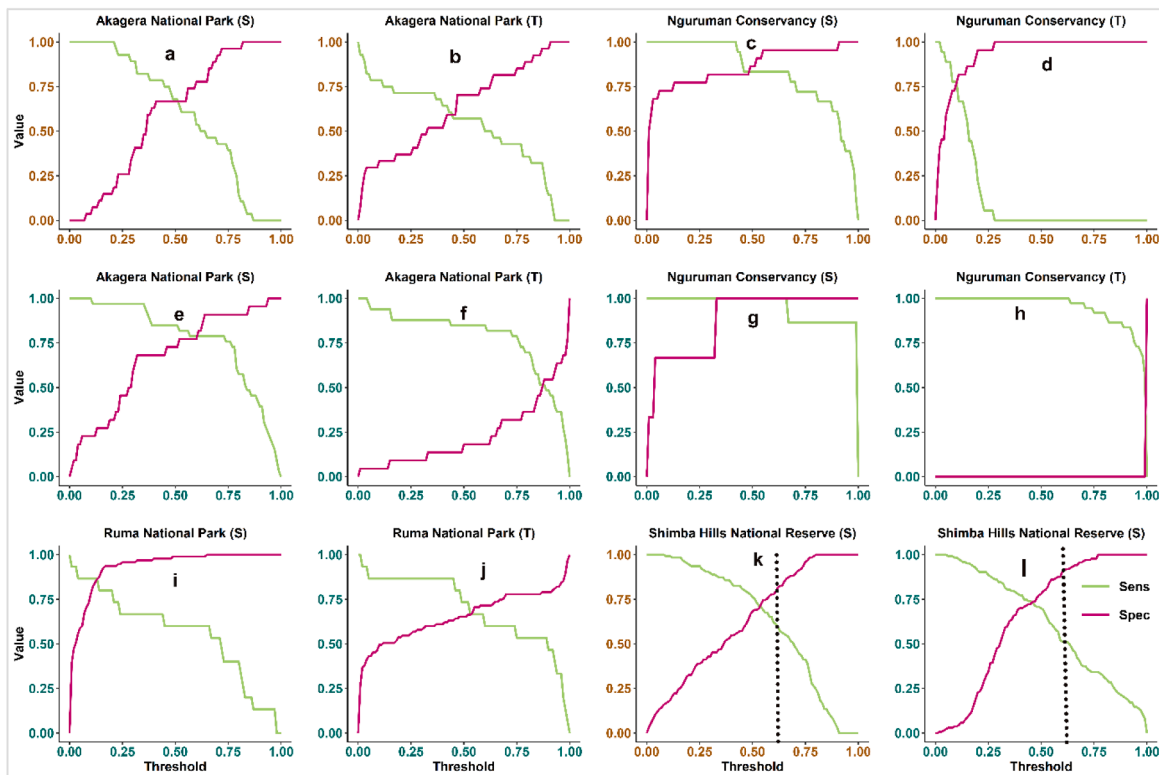
Fig. 5. Sensitivity and specificity curves for both transferred and site-specific models for the dry (brown) and wet (dark green) season. The site-specific models are denoted with (S) while the transferred models are denoted with (T). The dotted black line is the 0.55 optimal threshold used to categorize the suitable and unsuitable predicted tsetse habitats.

models are more reliable in explaining the distribution patterns of tsetse flies compared to transferred models.

To gain a deeper understanding of the observed results in tsetse habitat transfer, the sensitivity and specificity curves of these transferred models showed that using the optimal threshold from the reference site did not consistently yield accurate models. In some cases, it was found that using a lower threshold resulted in better-performing models (e.g., Nguruman conservancy in the dry season; Fig. 5d). This suggests that while the spatial patterns in the prediction maps in these areas do contain valuable information for identifying tsetse habitats in a relative

sense, the absolute values of predicted probabilities within them may not accurately reflect the actual presence or absence of tsetse habitats. Similar to the F1 scores, the evaluation of sensitivity and specificity curves confirmed that the site-specific models exhibited higher accuracies compared to the transferred models. In addition, the factors identified as significant in explaining tsetse occurrence (Fig. 6) as well as their direction of relationship (Fig. 7) varied across different locations for these site-specific models. However, it is important to note that optimizing thresholds or generating variable importance plots and response curves relies on having sampled locations and thus cannot be
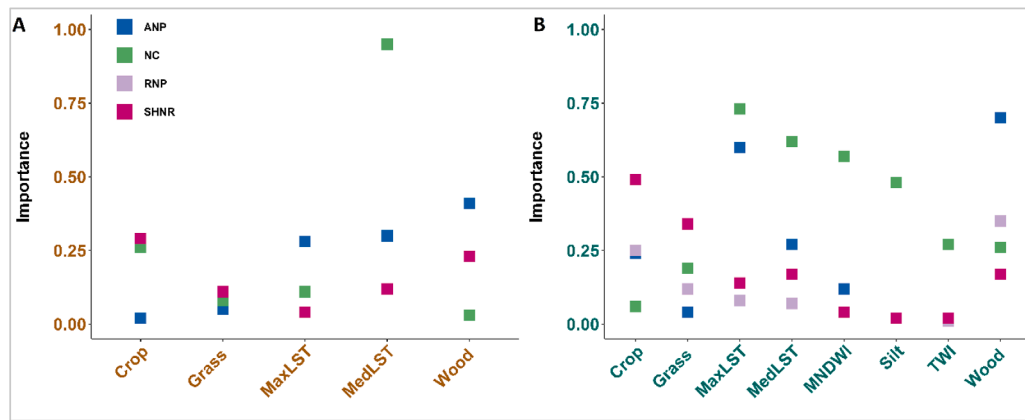
Fig. 6. Site-specific variable importance plots for the four study locations in the dry (a) and wet (b) season. The importance is based on the Area Under Curve (AUC) tests and higher values of variable importance mean that the variable plays a more prominent role in the (site-specific) models for each site.

performed when models are transferred to regions that lack *in-situ* tsetse trap data.

The observed differences in variable importance (Fig. 6) and direction of influence (Fig. 7) in site-specific models confirm the role of microhabitats and local adaptations in shaping tsetse fly habitat suitability. For example, while tsetse flies are typically negatively correlated with human-induced changes like intense land cultivation (Kuzoe and Schofield, 2004), in Akagera National Park tsetse occurrence was positively correlated with cropland fractions in the wet season. A possible explanation could be the fact that there is intensive livestock farming (Gashururu et al., 2021) outside of Akagera National Park providing host availability for tsetse to feed on despite the presence of scattered croplands. Furthermore, the finding that maximum and median LST were identified as more important variables in Akagera National Park and Nguruman Conservancy during the dry season suggests that tsetse flies are not consistently influenced by the same environmental factors across different geographic areas. In fact, even the models developed by Gachoki et al. (2021) showed that tsetse flies have locally adapted in the wet season, where their occurrence was negatively correlated with woodland fraction, contrary to expectations for *G. pallidipes*, which are considered woodland species.

When transferring models, it is expected that models are likely to transfer better in areas exhibiting environmental similarity. However, the threshold of how similar the area of model development should be to the target area is unknown (Sequeira et al., 2018; Yates et al., 2018). In our study, the models failed to transfer to the site and season that exhibited the highest dissimilarity, i.e., Nguruman Conservancy in the dry season. One possible explanation for this could be the presence of significant ecological variability at a small scale, as well as local factors that were not fully accounted for in our reference model. For instance, in Shimba Hills National Reserve, even though the cover by green wooded vegetation decreases slightly during the dry season, it still provides enough shade for tsetse flies to rest and breed. In contrast, in Nguruman Conservancy, the dry season experiences a stronger reduction in green cover, resulting in only small pockets of woodlands that remain green. As a result, tsetse flies in this area may have adapted to these specific microhabitats, which may have not been fully captured in the reference model, leading to poor transferability. Hence, although the transferability of tsetse habitat models is generally viable, it is crucial to recognize their inherent limitations that arise from the variability in tsetse flies' response to diverse environmental factors across regions. Therefore, during model development, it is crucial to make appropriate adjustments that consider the unique characteristics of all areas where tsetse flies are found.

Our study highlights the potential for transferring tsetse habitat models across geographically separated regions but with certain

limitations. A successful model transfer is more likely when the environmental conditions that determine the presence or absence of tsetse flies are generalizable. However, if the species is found in specific and localized microhabitats that are not incorporated in the original model, the transfer of the model is likely to fail (Liu et al., 2020). Therefore, the training models should be developed using comprehensive information that covers the entire geographic range where the species is likely to occur. One approach could be to combine tsetse data from various sites to create a "global" model that can be applied to other locations. For example, Bamford et al. (2009) found that models that included nesting habitats data for two vultures from different regions transferred better than single models. Additionally, incorporating biological traits of the species, such as information on animal host availability, as predictor variables in species distribution models, are likely to improve transferability to other regions compared to models that solely rely on environmental variables. Future research should focus on improving training models by using larger and more comprehensive tsetse datasets that capture the entire geographic range of the species and incorporating spatial predictors that account for the distribution of animal hosts. This could lead to more accurate and reliable models for predicting tsetse fly distribution at both local and at transferred sites.

5. Conclusion

Our research shows that tsetse habitat models can be transferred between geographically separated regions. However, the success of these transferred models depends on how similar the environmental conditions of the target sites are to those of the model training area. When applied to highly dissimilar regions, the transferred models fail. From this, we conclude that transfer of tsetse habitat models should be done with care and only for areas that have similar environmental characteristics as the location for which models were initially developed. Nonetheless, it is reassuring that fair accuracies can be achieved even for a tsetse species that is versatile and can occur in a wide range of environmental conditions. Evaluating model transferability represents a novel research direction in tsetse habitat mapping. To enhance habitat models further, incorporating data on the presence of host species is necessary. In addition, further assessment of models for other tsetse species could result in accurate and operational maps that enhance resource allocation for tsetse elimination efforts.
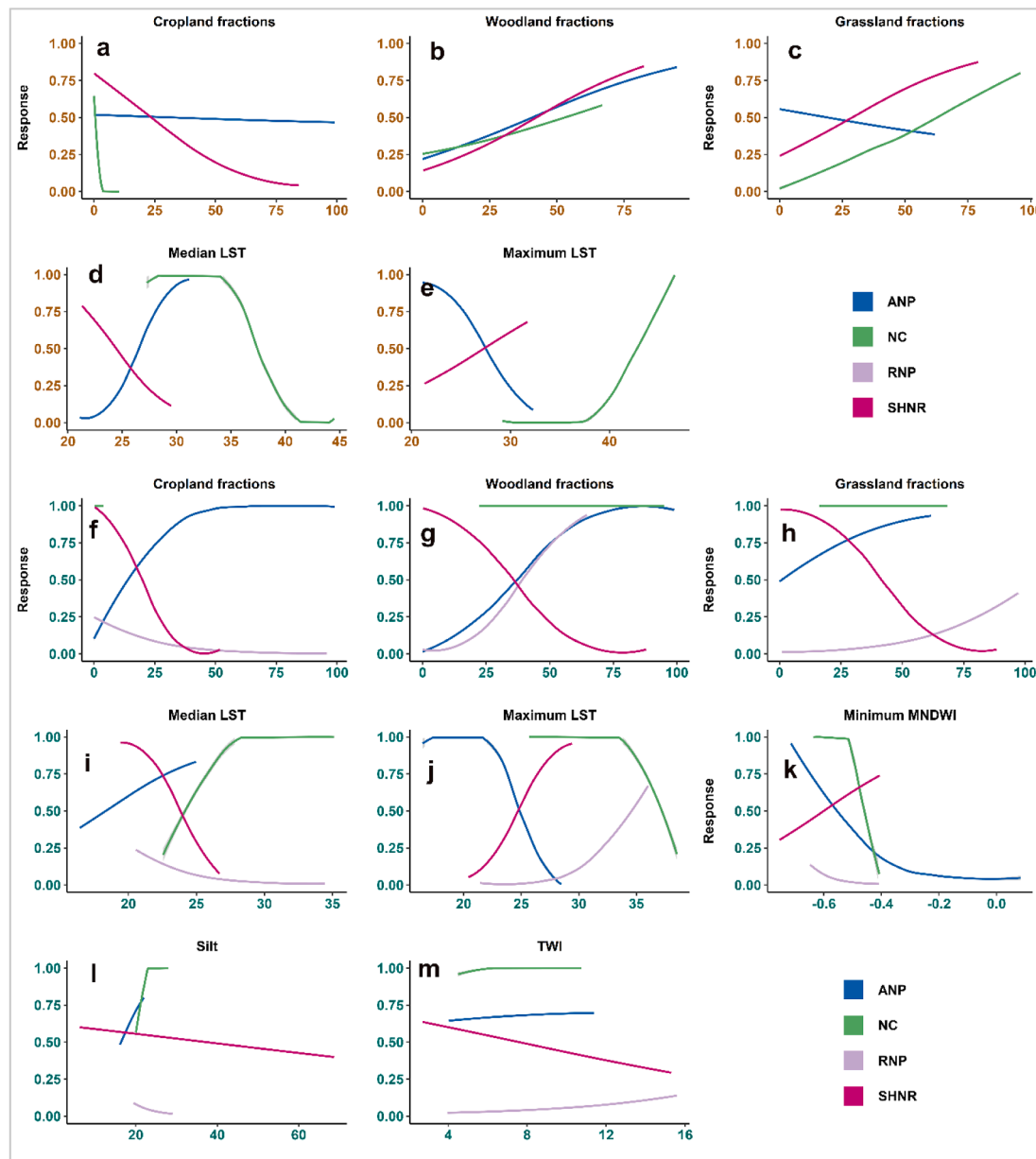
Fig. 7. Response curves for the site-specific model for the dry (brown) and wet season (dark green). Different colors illustrate different sites. ANP = Akagera National Park, NC = Nguruman Conservancy, RNP = Ruma National Park, SHNR = Shimba Hills National Reserve.

CRediT authorship contribution statement

Stella Gachoki: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Writing – original draft. **Anton Vrieling:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Thomas A. Groen:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Andrew Skidmore:** Conceptualization, Supervision, Writing – review & editing. **Daniel Masiga:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Tsetse data for Akagera National Park are publicly available at Gashururu et al. (2021). All other data are available at Data Archiving and Networked Services (DANS) repository (https://doi.org/10.17026/dans-253-37dq).