Journal of Public Health in Developing Countries Vol. 2, No. 2, pp. 150-161
http://www.jphdc.org/

ISSN 2059-5409

Original Contribution **Open Access**

Predictive Mapping of Transmission Risk of a Soil-Transmitted Helminth across East Africa: Findings from Community Prevalence Surveys

Jodi Gentry[1], Belinda Sturm[1], A. Townsend Peterson[1]

[1]University of Kansas, Lawrence, Kansas, USA

Correspondence to: A. Townsend Peterson, Biodiversity Institute, University of Kansas, 1345 Jayhawk Blvd, Lawrence, Kansas 66045-7593, USA. Email: town@ku.edu

ARTICLE INFO

Article history:

Received: 16 Feb 2016
Accepted: 23 Apr 2016
Published: 6 Jun 2016

Keywords:

- *Ascaris lumbricoides*
- Helminth
- Parasitic worm
- Ecological niche modeling
- Mapping
- Transmission risk
- Africa

ABSTRACT

Background: Despite being identified as a major health concern and neglected tropical disease, Ascariasis, caused by *Ascaris lumbricoides*, a soil-transmitted helminth, ranks among the most common infections worldwide.

Methods: Based on prevalence data from the community surveys across East Africa for 1975-2009, we used ecological niche modeling to summarize and predict the spatial distribution of *A. lumbricoides*' transmission risk.

Results: Projection of this model across East Africa identified 51.4% of the region as suitable for *A. lumbricoides* transmission, with least suitable areas in the Sahara and the Somali-Chalbi deserts. Partial area under the curve (AUC) tests based on independent data showed that our model predictions were better than random expectations in predicting infection risk of *A. lumbricoides*; the model successfully identified areas of high and low infection risk in Ethiopia and Eritrea.

Conclusions: Although preliminary, this occurrence-environment mapping approach provides predictions that can guide education and mitigation efforts in the broader East African region.

Citation: Gentry J, Sturm B, Peterson AT. Predictive Mapping of Transmission Risk of a Soil-Transmitted Helminth across East Africa: Findings from Community Prevalence Surveys. J Public Health Dev Ctries. 2016; 2(2): 150-161.

INTRODUCTION

Soil-transmitted helminths (STHs) are probably the most common infective agents worldwide. Recent prevalence maps show their distributions spanning three continents, most frequently in marginalized human populations [1,2]. Although reports vary, over one billion people are likely infected with one or more STH, most commonly *Ascaris lumbricoides* [2-4]. While STH-related mortality is low, morbidity includes diarrhea, malnutrition, stunted growth, and reduced cognitive development, particularly for children 5-15 years of age [1,5]. *Ascaris* infections are also associated with increased fecundity in human populations [6].

STHs spend portions of their life cycles outside of human hosts. Adult worms inhabit intestines of humans, reproduce sexually, and produce eggs that are released to the environment through defecation. Once in the environment, eggs are not infective until they undergo embryonation, and reach the appropriate larval stage. In *Ascaris lumbricoides*, eggs do not hatch until they are (re)ingested, generally in environments contaminated via practices such as open defecation and agricultural application of human waste [7].

Despite their public health importance, STH infections remain listed by the World Health Organization as one of 17 neglected tropical diseases [1]. Although extensive antihelminthics drug administration campaigns and sanitation/hygiene education programs have reduced STH prevalences in some areas, potential for reinfection is high, considering their ubiquitous presence and resiliency in natural environments, as well as the underlying need for improved infrastructure. Because transmission involves complex interactions between pathogens, hosts, and their environments [8], many challenges remain for eradication. By integrating known occurrences with environmental conditions under which a species can survive through its entire life cycle, correlational ecological niche models (ENMs) offer insights into disease transmission geography, even when factors that determine the distribution, prevalence, and intensity of infection are not fully understood [9].

The purpose of this study was to assess the utility of ENMs in anticipating the spatial distribution of *Ascaris lumbricoides* transmission risk across East Africa, and to evaluate how those predictions compare with those based on current mapping approaches. ENMs have been applied successfully to other disease systems, allowing novel insights into ecology [10], current spatial distributions [11], climate-mediated shifts [12], and unknown interactions among species [13]. Their application to improving risk maps for STHs, however, has not been explored.

MATERIALS AND METHODS

Input Data

Georeferenced point prevalence data for *Ascaris lumbricoides* from 1975-2009 for East Africa were obtained from the *Global Atlas of Helminth Infections* [14]. The study area included four countries from which occurrence data were available (Burundi, Kenya, Rwanda, Tanzania), and ten surrounding countries to which we would "transfer" model predictions: Democratic Republic of Congo (DRC), Djibouti, Eritrea, Ethiopia, Malawi, Mozambique, Somalia, Sudan, Uganda, and Zambia (Figure 1). The original dataset for model calibration included 1,557 data points, with prevalence values ranging 0-96%. Duplicate records, studies based on <50 samples, sites with zero prevalence, and indeterminate geographic locations were removed from consideration, leaving an initial modeling dataset of 780 unique occurrence locations (Figure 2).

To determine optimal prevalence thresholds for inclusion as a "presence" record in the modeling dataset, we created subsets based on the 100^{th} (i.e., any detection of the parasite), 80^{th}, 60^{th}, 50^{th}, and 20^{th} (i.e., only the top fifth of measured prevalences) percentiles of the prevalence distribution. Corresponding prevalence thresholds and numbers of data points included in each subset are shown in Table 1. These five prevalence subsets were used for initial model testing to optimize choices in these model applications that require binary response variables. Monthly composites of the Normalized Difference Vegetation Index (NDVI) drawn from the Advanced Very High Resolution

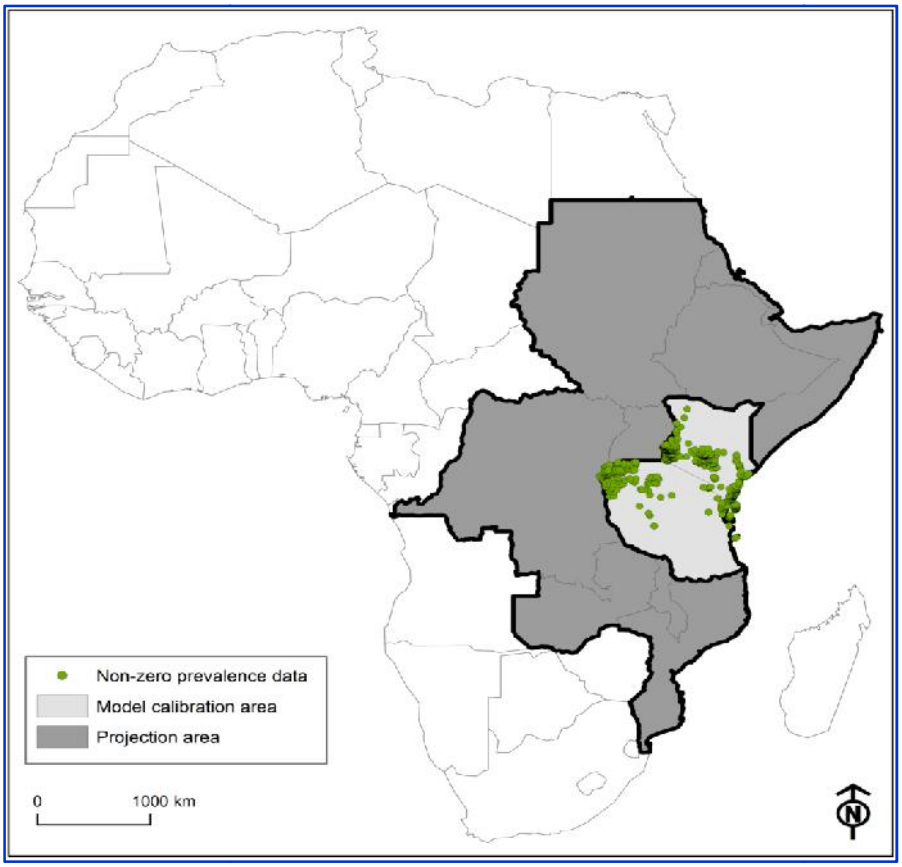

Figure 1. Study Area in East Africa
[Light gray shading represents the four countries for which occurrence data were available (Burundi, Kenya, Rwanda, Tanzania); Dark gray shading represents the 10 surrounding countries to which model predictions were "transferred" (Democratic Republic of Congo, Djibouti, Eritrea, Ethiopia, Malawi, Mozambique, Somalia, Sudan, Uganda, Zambia)]

Radiometer (AVHRR) sensor were used to characterize environments [15] for April 1992 through March 1993. This period corresponds roughly to the median age of records in our occurrence dataset. The 1-km resolution NDVI data layers reflect seasonal variation in photosynthetic mass ("greenness"), and thus serve as a proxy for measures of key dimensions of climate (temperature and precipitation), as well as vegetation type and aspects of land cover. Seasonal variation in 'greenness' summarized in NDVI values has been shown to be an important predictor of the geography of disease transmission in previous studies [16].

To reduce dimensionality of the NDVI data layers, we used principal components analysis (PCA; principal components tool in ArcGIS 10, ESRI 2011) on the covariance matrix of the 12 monthly composite data layers. We retained all components with eigenvalues greater than unity, which turned out to be all 12. Hence, although we did not reduce dimensionality, at least we managed to create new variables with lower variable-to-variable correlations. This PCA-transformed dataset was used as input for all niche models in this study.

Niche Models

To identify and characterize associations between STH occurrence and environmental characteristics, we used the Genetic Algorithm for Rule-Set Prediction (GARP) [17]. GARP is a correlational ENM approach that uses non-random associations between known occurrence points and environmental data layers to evolve a set of conditional rules for describing ecological niches of species. Although discussion in the literature has pondered the predictive abilities of GARP [18], much of the criticism was based on inappropriate interpretations of summary

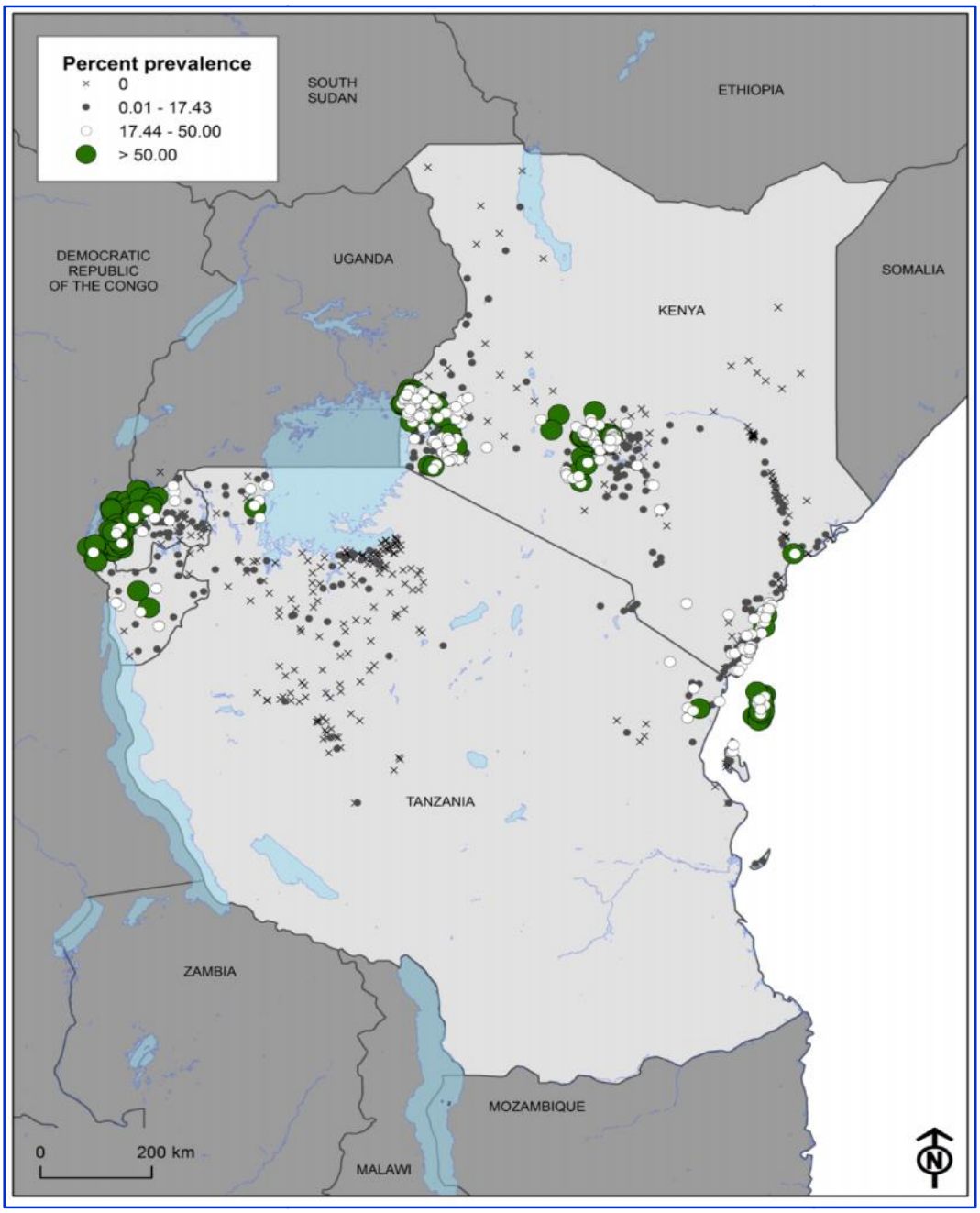

Figure 2. Initial Dataset of 780 Unique Locations
[Prevalence values ranging 0-97.7%. Locations with zero prevalence are shown for reference, but were not used in modeling]

statistics [19], and its utility as regards disease transmission is well documented [17-19].

GARP creates ENMs by dividing known occurrence data randomly into training and testing subsets: 25% for developing rules (training data), 25% for internal model refinement (intrinsic testing data), and 50% for external testing of model quality (extrinsic testing data). Using an iterative process of rule selection, GARP selects randomly from a set of inferential tools (logistic regression, bioclimatic rules, atomic rules, negated range rules), and generates initial rules; it then modifies those rules based on specific operators of the genetic algorithm designed to mimic chromosomal evolution (e.g., crossover, mutation), and either rejects or incorporates the result into the model. After each modification, rule quality is tested using the intrinsic testing data (to maximize both significance and predictive accuracy), and a size-limited set of rules is retained. Because

Table 1. Subsets of Data with Non-zero Prevalence Used for Initial Testing and Determination of Optimal Threshold Values for Model Calibration

Percentile in Data Set	Data Points	Prevalence Threshold (%)	Coarse Cells Occupied	Subsets with AUC Ratios <1.0
100	780	0.4	112	8/10
80	628	3.3	90	5/10
60	469	10.3	62	4/10
50	390	17.4	52	3/10
20	156	47.5	30	9/10

Percentile indicates the portion of the initial population of 780 points included in the analysis; Coarse Cells Occupied indicates numbers of coarse-resolution (60 km) grid cells available for final model tests; AUC, Area Under the Curve, shows results of model evaluations

rules are tested with independent data, performance values reflect the expected performance of the rule, and provide an independent verification that gives a more reliable estimate of true rule performance. The final result is a set of rules that can be projected into geographic space to produce maps of potential distributions of species [17].

Spatial Autocorrelation

One of the challenges of using human population disease prevalence data in ENM analyses is that data are often based on convenience sampling, not necessarily following a carefully-designed sampling scheme [9,20,21]. In these cases, use of the entire dataset risks focusing models overmuch on conditions in heavily-sampled areas [9]. Hence, we calculated the spatial lag of the NDVI PCAs (the distance at which data points are no longer spatially autocorrelated), and removed locations from the modeling dataset those that violated the assumptions of spatial independence, as follows.

We generated 5,000 points distributed randomly across the study area, removing locations that intersected large bodies of water such as Lake Victoria. This dataset was populated with attribute values corresponding to the values of the first principal component, which summarized 75.7% of overall environmental variation in the region. We used ArcGIS 10 to create an empirical semivariogram to model spatial dissimilarity: the semivariogram measures the strength of statistical correlation as a function of distance by calculating squared differences for all pairs of locations, and plotting half the squared difference on the *y*-axis versus the distance separating each pair on the *x*-axis. We considered the lag as the distance associated with 80% of the sill (the height that the semivariogram reaches when it levels off).

To rarefy the occurrence data spatially to reflect this autocorrelation structure, we created a vector grid spanning the entire study region, with grid cell sizes corresponding to the calculated spatial lag. For each of the five prevalence threshold subsets, we created 10 replicate datasets by selecting single representative points at random from each grid cell for which prevalence data existed. The result was 50 replicate datasets (10 for each prevalence threshold), each having no more than one location in each grid cell (Table 1).

For initial model tests, half of the data points from the replicate datasets were chosen at random for calibrating models and half were set aside for extrinsic testing. For each of the 50 replicate datasets, we used DesktopGARP (http://www.nhm.ku.edu/desktopgarp) to develop 100 random-walk models, and filtered out 80%

based on error statistics in accordance with best-practice recommendations [22]. A 'best subset' of 20 models was retained on the basis of omission (leaving out true potential distributional areas) statistics calculated from the extrinsic testing data. Specifically, we used a 'soft' omission threshold of 20%; the result was 20 low-omission models that were summed to produce a geographic prediction for each of the 50 datasets [22].

The performance of models based on each prevalence threshold was determined by calculating area under the curve (AUC) ratios (ratios of model AUC to the null expectation) using partial receiver operating characteristic (ROC) analysis [19]. Following this approach, only omission errors of <5% were considered (i.e., sensitivity >0.95). We selected as the best predictive niche model the prevalence threshold for which models showed the fewest partial AUC ratios of <1.

Projection to Other Regions

To test model transferability across broader regions, we used our best ENM based on the occurrence data from the four countries for which data were available initially, and projected it across 10 adjacent countries (Figure 1) following procedures outlined above. The 10 replicate models summed from the 20 'best subsets' were averaged pixel by pixel to produce a final geographic prediction across the broader region. Model performance in these regions was determined by calculating partial AUC ratios, as described above, on an independent set of data also obtained from the *Global Atlas of Helminth Infections* [14]. This dataset consisted of 143 point locations collected from Eritrea (in 2002) and Ethiopia (in 2010). We removed points with prevalence values falling below our empirically selected threshold, indeterminate locations, duplicates, and studies based on <50 samples, producing a validation dataset of 22 unique locations.

RESULTS

Using principal components analysis of the NDVI data for 1992-93, we determined that 75.7% of the variation of the data was explained by the first component, and retained components with eigenvalues of >1. Inspecting the semi-variogram of the first component, we calculated the spatial lag of this dataset at ~54 km (Figure 3). Figure 4 shows a vector grid network reflecting a 60-km grid over the region, with the training and testing cells used for one of the random subsets with which we tested the predictive capabilities of the models.

Fifty preliminary replicate models (10 for each of the 5 prevalence subsets) were constructed to determine the best prevalence threshold for our dataset. The 50th percentile data subset produced AUC ratios with values >1 (i.e., statistically significant predictions) in all but three iterations (Table 1), which suggests that GARP has the best predictive abilities when the point prevalence data for *Ascaris lumbricoides* are above 17.4% (equivalent to the median prevalence value in the overall dataset, or the 50th percentile of the data). As a result, we used this subset in all subsequent analyses.

To test model transferability, we projected models based on all available presence data (but rarefied to 60 km resolution) across a broader region of East Africa (Figure 5). After thresholding, our best predictive model identified 51.4% of the overall region as suitable for *A. lumbricoides.* The least suitable locations were identified in the Sahara Desert in northern Africa and the Somali-Chalbi Desert east of the Ethiopian Highlands.

Transferring this model across Eritrea and Ethiopia (Figure 6) identified 59.9% of the area of the two countries as suitable for *Ascaris lumbricoides* transmission. This area anticipated all 22 of the test presence data in this independent region, such that omission error was zero. Partial AUC ratios were greater than one for all iterations, such that our model performs better than random at predicting the distribution of *A. lumbricoides* (P <0.001). Perhaps most intriguing is the low overall suitability predicted across Eritrea, where none of the 40 studies included in the initial validation dataset documented prevalence >3%, providing further corroboration of our model projections across northeastern Africa.

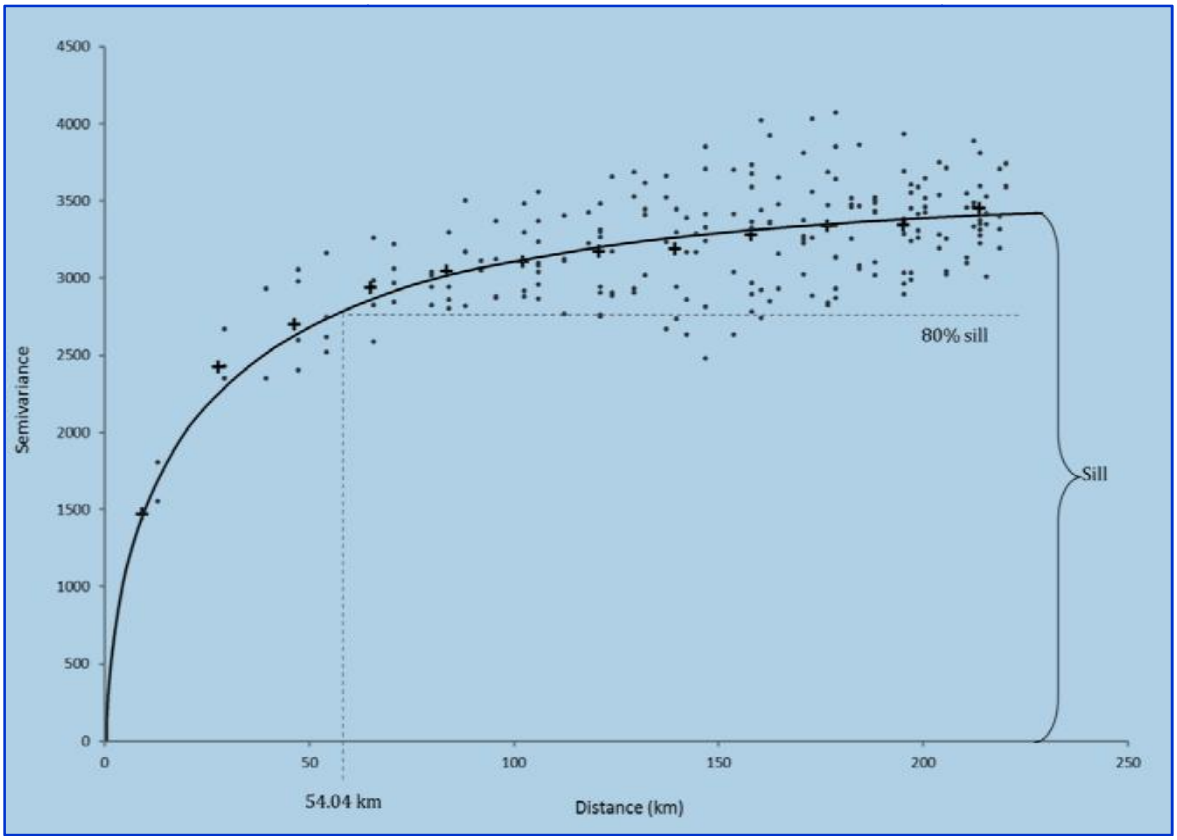

Figure 3. Calculating Spatial Lag
[Empirical semivariogram showing the spatial lag of the first principal component of normalized difference vegetation indices (NDVI) for April 1992 through March 1993. We defined the spatial lag as the distance associated with 80% of the sill (the height that the semivariogram reaches when it levels off), which was~54 km]

DISCUSSION

Historically, studies of STH transmission have consisted, in large part, of local epidemiological surveys and generalized global summaries. Epidemiological survey data are costly and time-consuming to produce [7], and can be intrusive and alienating to already-marginalized populations [4]. In such resource-poor environments, limited funding might better be spent on treatment and prevention [2,3,5]. In addition, local-scale analyses can miss elements of the "big picture," offering only limited insight into the complex spatio-temporal interactions that determine the geography of disease transmission.

Efforts to map generalized global distributions have interpolated between sampling locations to offer a more global picture of STH transmission geography. Although these efforts offer broad-scale overviews, the results are likely too general to meet needs of public health officials planning treatment and prevention interventions [23]. In addition, interpolations based on overly sparse prevalence surveys to target broad areas for mass treatment can be inefficient, and unnecessary remediation efforts may increase potential for developing drug resistance [5].

Recent studies have taken advantage of geographic information systems (GIS) and fine-resolution, remotely-sensed environmental data to produce more rigorous analyses of disease distributions based on associations between known occurrences and environmental variables [23]. One such approach has been explored in the *Global Atlas of Helminth Infections*, which uses Bayesian statistics in a GIS framework to map STH prevalence and predictive risk [14]. Pullan et al. [24] provided a detailed analysis of some of the same data (i.e., across Kenya) as we have examined, but using Bayesian space-time approaches, considering both human and environmental covariates and incorporating

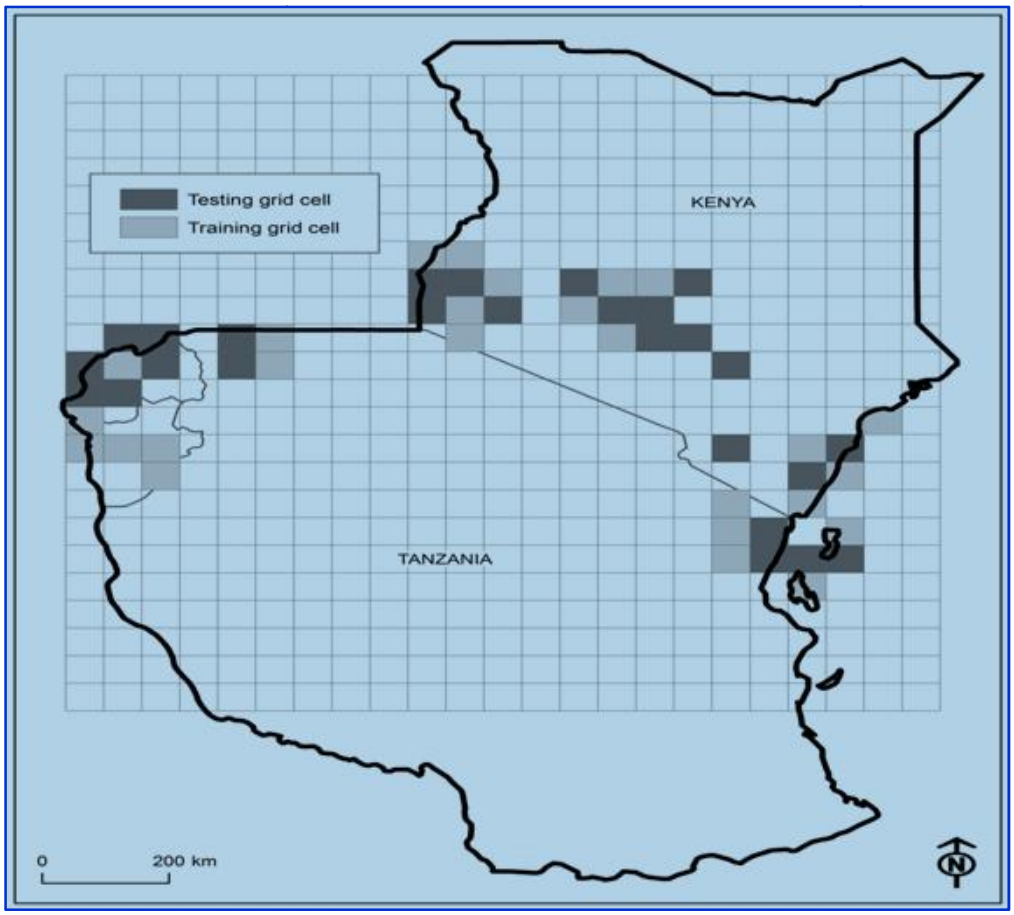

Figure 4. The Vector Grid Network
[The vector grid network is based on a 60-km grid over the calibration region. Dark shaded areas represent testing grid cells; light shaded areas represent training grid cells]

spatial and temporal autocorrelation. Their model evaluations, however, depended on random subsets of available data, and their use of traditional (i.e., non-partial) ROC approaches is subject to several complications [19,25]. Our analyses are simpler and data-driven, but have greater ability to project and transfer among regions, which opens greater opportunities for leveraging existing data into useful knowledge for less-well-known regions.

ENMs are a relatively new set of tools in spatial epidemiology that emphasize development of distribution maps within explicit ecological and biogeographic frameworks [9]. Our model of transmission risk of *Ascaris lumbricoides* across East Africa showed greatest risk in savannah and forested tropical areas, and least risk in arid and semi-arid regions. This finding is consistent with studies that indicate that *A. lumbricoides* survival is maximal under warm, moist conditions, ceasing at temperatures exceeding 38°C [7], such that model predictions and expert experience coincide well. The best predictive ability of GARP for mapping occurrence of *A. lumbricoides* across East Africa occurred at the median prevalence values of 17.4%, which approximates the World Health Organization's minimum threshold for large-scale interventions of 20% [26].

In spite of the methodological differences between the two analyses (Bayesian and ENM approaches), both found two areas of highest transmission risk: in western Kenya in the areas close to Lake Victoria, and in southeastern Kenya close to the coast. The Bayesian analyses take human factors into account directly, while the niche-modeling analyses consider them only in that they affect prevalences. Our analyses are set more directly in the context of distributional ecology, and as such appear to be more easily extrapolated to other regions. Regardless, more detailed

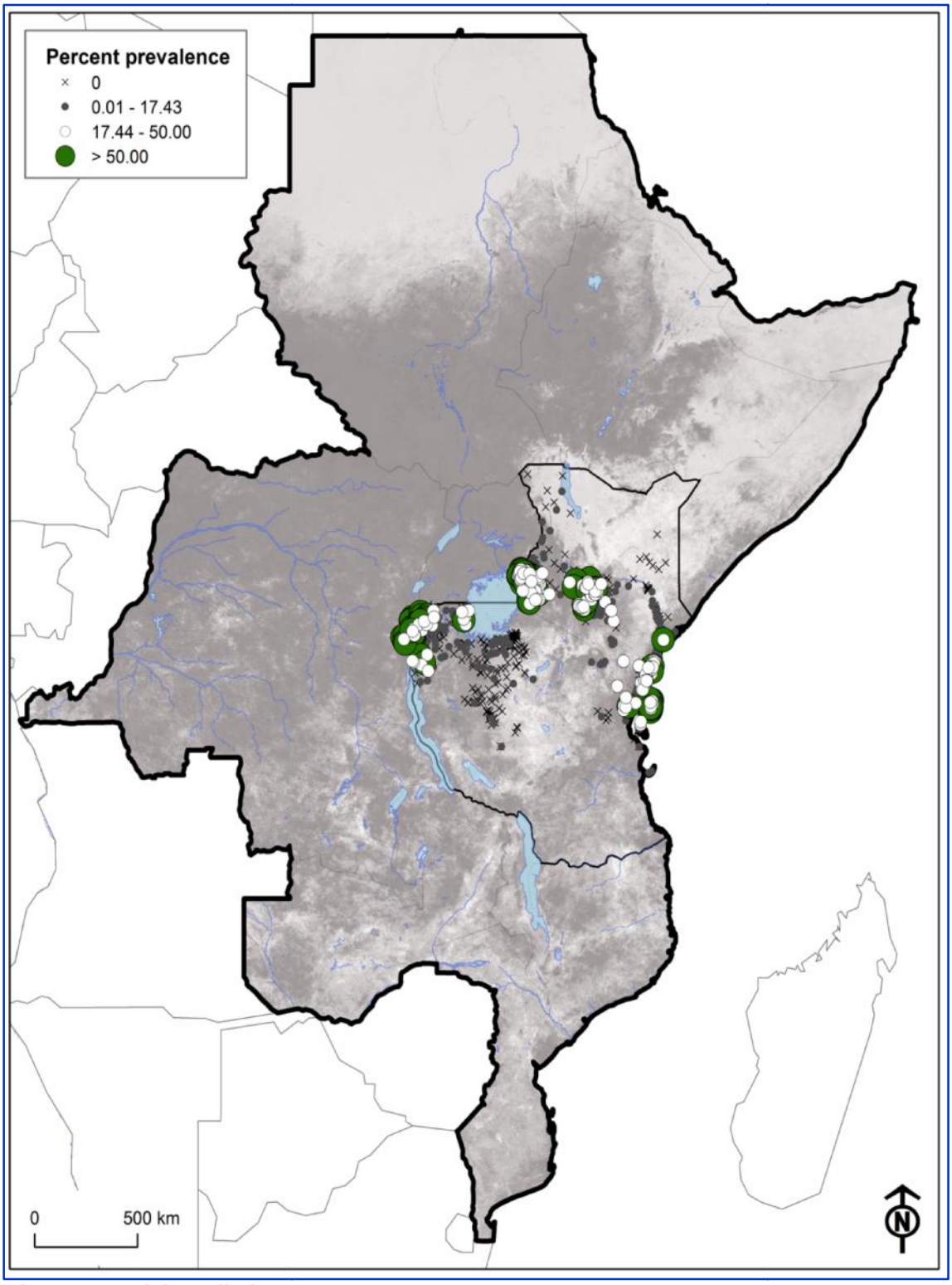

Figure 5. Model Predictions

[Model predictions for *A. lumbricoides* transferred to areas for which prevalence data were not used in model calibration. Gray scale indicates modeled suitability of *A. lumbricoides*, with lighter shades indicating low suitability, and darker shades indicating high suitability]

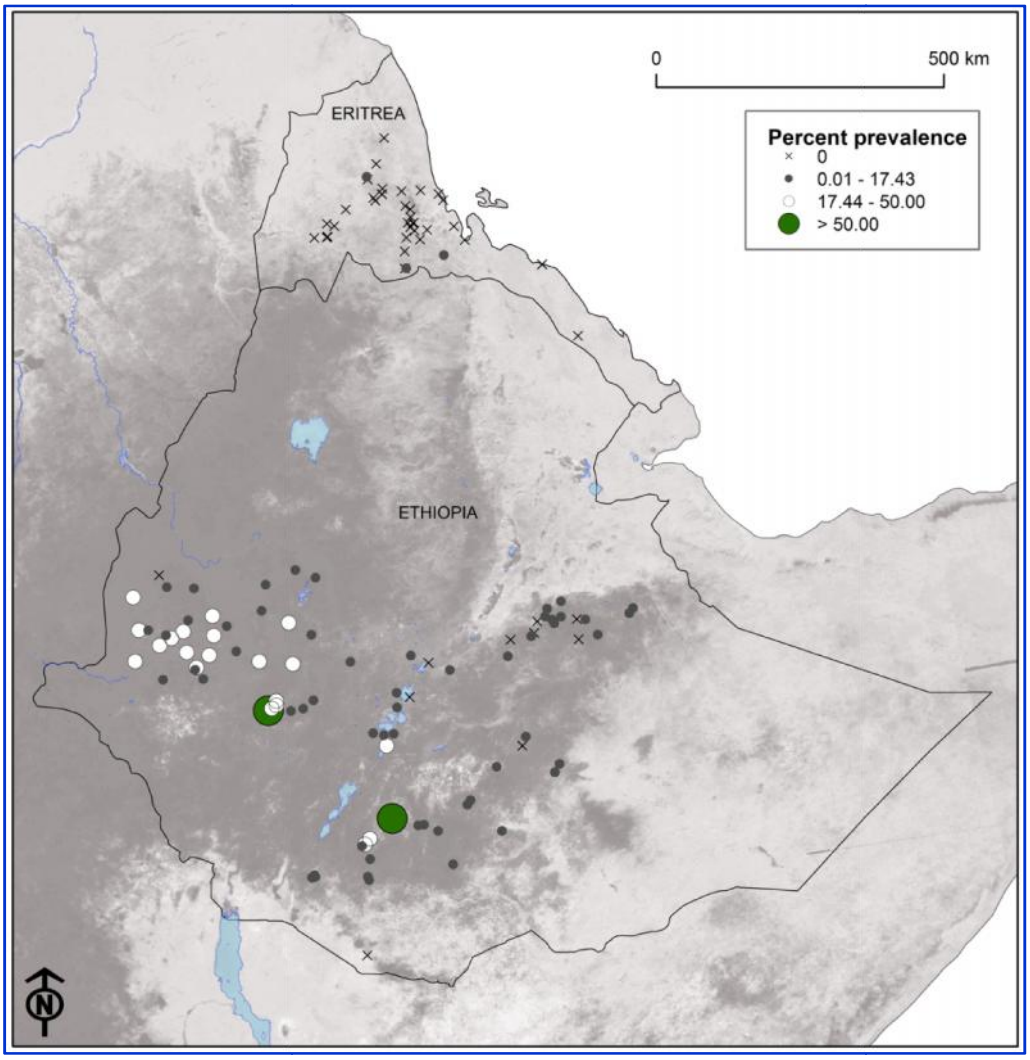

Figure 6. Model Validation
[Model validation across a region for which no data were available in model calibration, based on independent validation data from Eritrea and Ethiopia. Gray scale indicates modeled suitability of *A. lumbricoides*, with lighter shades indicating low suitability, and darker shades indicating high suitability. Partial AUC tests were significant for all iterations]

comparisons among these alternative methods may prove quite rewarding.

This study nonetheless has several significant limitations. In particular, the occurrence data available are highly clumped spatially, such that sample sizes after incorporating spatial autocorrelation were quite small (N = 52 at 17.4% prevalence), which limits the detail possible in calibrating ENMs, the statistical power of model evaluations, and potentially the regions to which these methods can be applied. More generally, historical, social, and cultural aspects of disease transmission (e.g., access to sanitation and hygiene) affect transmission of STHs as well, which can confuse model calibration easily. However, our models were highly predictive across broad, unsampled areas, so our risk maps may be useful in guiding public health agencies in planning treatment campaigns more effectively. Application of ENM to predict potential distributional shifts of STHs under future-climate conditions is the focus of coming research efforts.

CONCLUSION

Although preliminary, the occurrence-environment mapping approach discussed in this study provides predictions that can guide education and mitigation efforts for the control of

STH infections in the broader East African region.

AUTHORS' CONTRIBUTIONS

JG and ATP developed the analyses. All authors worked to develop the manuscript.

ACKNOWLEDGEMENTS

Simon Brooker kindly provided access to the occurrence data for this study. The C-CHANGE IGERT program supported JG during the development of the analyses.

CONFLICT OF INTEREST

Authors have declared that no competing interests exist.

REFERENCES

1. WHO. Working to Overcome the Global Impact of Neglected Tropical Diseases: First WHO Report on Neglected Tropical Diseases. Geneva: *World Health Organization*; 2010.
2. de Silva NR, Brooker S, Hotez PJ, Montresor A, Engels D, Savioli L. Soil-transmitted helminth infections: Updating the global picture. *Trends Parasitol*. 2003; 19: 547-51.
3. Chan MS, Medley GF, Jamison D, Bundy DA. The evaluation of potential global morbidity attributable to intestinal nematode infections. *Parasitology*. 1994; 109: 373-87.
4. Crompton DW. How much human helminthiasis is there in the world? *J Parasitol*. 1999; 85: 397-403.
5. Bethony J, Brooker S, Albonico M, Geiger SM, Loukas A, Diemert D *et al*. Soil-transmitted helminth infections: ascariasis, trichuriasis, and hookworm. *Lancet*. 2006; 367: 1521-32.
6. Blackwell AD, Tamayo MA, Beheim B, Trumble BC, Stieglitz J, Hooper PL *et al*. Helminth infection, fecundity, and age of first pregnancy in women. *Science*. 2015; 350: 970-2.
7. Brooker S, Clements AC, Bundy DA. Global epidemiology, ecology and control of soil-transmitted helminth infections. *Adv Parasitol*. 2006; 62: 221-61.
8. Weaver HJ, Hawdon JM, Hoberg EP. Soil-transmitted helminthiases: Implications of climate change and human behavior. *Trends Parasitol*. 2010; 26: 574-81.
9. Peterson AT. *Mapping Disease Transmission Risk in Geographic and Ecological Contexts.* Baltimore: Johns Hopkins University Press; 2014.
10. Costa J, Peterson AT, Beard CB. Ecological niche modeling and differentiation of populations of *Triatoma brasiliensis* Neiva, 1911, the most important Chagas disease vector in northeastern Brazil (Hemiptera, Reduviidae, Triatominae). *Am J Trop Med Hyg*. 2002; 67: 516-20.
11. Nakazawa Y, Lash RR, Carroll DS, Damon IK, Karem KL, Reynolds MG *et al*. Mapping monkeypox transmission risk through time and space in the Congo Basin. *PLOS ONE*. 2013; 8: e74816.
12. Campbell LP, Luther C, Moo-Llanes D, Ramsey JM, Danis-Lozano R, Peterson AT. Climate change influences on global distributions of dengue and chikungunya virus vectors. *Philos Trans R Soc Lond B Biol Sci*. 2015; 370: 20140135.
13. Peterson AT, Sánchez-Cordero V, Beard CB, Ramsey JM. Ecologic niche modeling and potential reservoirs for Chagas disease, Mexico. *Emerg Infect Dis*. 2002; 8: 662-7.
14. GAHI. *The Global Atlas of Helminth Infection*. (Accessed 15 December 2015, at http://www.thiswormyworld.org).
15. James M, Kalluri S. The Pathfinder AVHRR land data set: An improved coarse resolution data set for terrestrial monitoring. *Int J Remote Sens*. 1994; 15: 3347-63.
16. Bodbyl-Roels S, Peterson AT, Xiao X. Comparative analysis of remotely-sensed data products via ecological niche modeling of avian influenza case occurrences in Middle Eastern poultry. *Int J Health Geogr*. 2011; 10: 21.
17. Stockwell DRB, Peters DP. The GARP modelling system: Problems and solutions to automated spatial prediction. Int J Geogr Inf Sci. *Int J Geogr Inf Sci*. 1999; 13: 143-58.
18. Elith J, Graham C, Anderson RP, Dudík M, Ferrier S, Guisan A *et al*. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*. 2006; 29: 129-51.
19. Peterson AT, Papeş M, Soberón J. Rethinking receiver operating characteristic analysis applications in ecological niche modelling. *Ecol Model*. 2008; 213: 63-72.
20. Diniz-Filho JAF, Bini LM, Hawkins BA. Spatial autocorrelation and red herrings in geographical ecology. *Glob Ecol Biogeogr*. 2003; 12: 53-64.

21. Dormann CF, McPherson JM, Araújo MB, Bivand R, Bolliger J, Carl G *et al*. Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*. 2007; 30: 609-28.
22. Anderson RP, Lew D, Peterson AT. Evaluating predictive models of species' distributions: Criteria for selecting optimal models. *Ecol Model*. 2003; 162: 211-32.
23. Brooker S, Michael E. The potential of geographical information systems and remote sensing in the epidemiology and control of human helminth infections. *Adv Parasitol*. 2000; 47: 245-88.
24. Pullan RL, Gething PW, Smith JL, Mwandawiro CS, Sturrock HJ, Gitonga CW *et al*. Spatial modelling of soil-transmitted helminth infections in Kenya: A disease control planning tool. *PLoS Negl Trop Dis*. 2011; 5: e958.
25. Lobo JM, Jiménez-Valverde A, Real R. AUC: A misleading measure of the performance of predictive distribution models. *Glob Ecol Biogeogr*. 2008; 17: 145-51.
26. WHO. Preventive Chemotherapy in Human Helminthiasis: Coordinated Use of Antihelminthic Drugs in Control Interventions: A Manual for Health Professionals and Programme Managers. Geneva: *World Health Organization*; 2006.