THE BEST LOCAL-SCALE PREDICTION MAPS FOR DYNAMIC LANDSCAPE PATTERNS OF AQUATIC HABITATS OF ANOPHELINE LARVAE IN WESTERN LOWLAND KENYA

By

Nicole Jean Smith

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Geography—Master of Science

2016

ABSTRACT

IDENTIFYING THE BEST LOCAL-SCALE PREDICTION MAPS FOR DYNAMIC LANDSCAPE PATTERNS OF AQUATIC HABITATS OF ANOPHELINE LARVAE IN WESTERN LOWLAND KENYA

By

Nicole Jean Smith

The possibility of anopheline larval control and the need to understand the contribution of larval habitat distribution to the intensity of the malaria transmission cycle have generated inquiry into the relationships of anopheline larval habitats with environmental variables, including those variables that can be remotely-sensed across the landscape. These habitats are spatially predictable but their occurrence is unstable throughout time such that a map of their locations has a short lifespan of high accuracy. In this study, I create a dynamic environmental model of aquatic habitats of anopheline larvae for Asembo, a community in western Kenya, using topography, land-use/land-cover, and rainfall variables that have shown previous success in landscape models of *Anopheles* spp. habitats. I compare the success of the model's prediction maps when confronted with new data in another year at the same site to the accuracy of nearly-contemporaneous maps of the habitats as well as kriging-interpolated maps that exploit the habitat spatial clustering to increase the predictive power of the map. The dynamic environmental model shows the best predictive power of the three map types tested. The dominant input variable, the topographic position index, is further investigated, showing that the relationship is strongest at the 1710m scale and the predictions are moderately robust to elevation measurement errors. Though the prior knowledge of habitat locations does not accurately predict their future locations for long, I identify significant spatiotemporal autocorrelation in the distribution of the aquatic habitats that could be used in future prediction mapping to fine-tune generalized environmental models to site-specific patterns when some habitats have already been identified.

ACKNOWLEDGEMENTS

Grant #0723770 from the National Science Foundation funded work included in this thesis. I would like to thank my advisor, Dr. Joseph Messina, for his instruction and discussion contributing to my work, as well as for his generous and patient attitude over the time it took me to complete it. I would also like to acknowledge my committee member Dr. David Lusch for his feedback and my committee member Dr. Edward (Ned) Walker (co-PI with Dr. Messina for this project) for sharing the 2007 habitats dataset as well as his feedback. Dr. Robert McCann shared his 2011 habitats dataset as well as his hospitality and experience during my field work in Asembo. He also worked with me on the class project that piloted the creation of the custom digital elevation model used in this thesis. Dr. Ashton Shortridge shared useful feedback on that class project. Dr. Sarah Hession of MSU's CSTAT office provided consulting regarding the selection of the indicator kriging method. Philip Owera, CDC/KEMRI technician with the Kisumu office, was most valuable in the field for the ground reference data collection. The cooperation of the communities of Asembo and Seme has been indispensable for the larger project of which my study is a part and when one of my paths to some random location crossed someone else's there, I was warmly received.

Finally, I would like to thank my friends and family for their support and for their faith in me while I pursued the completion of this degree. I'm especially grateful to my beloved partner, Bob Murphy, for his unwavering belief in me, without which I most certainly could not have persisted through my setbacks on the way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

AIC	Akaike information criterion
An. arabiensis	*Anopheles arabiensis*
An. gambiae s.l.	*Anopheles gambiae sensu lato*
An. gambiae s.s.	*Anopheles gambiae sensu stricto*
AUC	Area under the (receiver operating characteristic) curve
BIC	Bayesian information criterion
DEM	Digital elevation model
DNA	Deoxyribonucleic acid
ENFA	Ecological Niche Factor Analysis
EROS	Earth Resources Observation and Science
GPS	Global Positioning System
GSOD	Global Summary of the Day
ILRI	International Livestock Research Institute
ISODATA	Iterative Self-Organizing Data Analysis Technique
LULC	Land-use/land-cover
P. falciparum	*Plasmodium falciparum*
RMSE	Root mean square error
ROC	Receiver operating characteristic
spp.	Species (plural)
SPOT	Satellite Pour l'Observation de la Terre
SRTM	Satellite Radar Topography Mission
TPI	Topographic position index
TWI	Topographic wetness index
TRMM	Tropical Rainfall Measuring Mission

USGS	United States Geological Survey

CHAPTER 1

BACKGROUND

The deadly disease malaria is transmitted by mosquitoes that depend on aquatic habitats for the juvenile phases of their life cycle. The ability to predict the spatial and temporal distribution of these larval habitats contributes to the ability to understand, forecast, and control malaria incidence. The objective of this thesis is to examine and compare the roles that landscape environmental correlates of aquatic habitat presence and autocorrelation of the habitats themselves play in predicting when and where habitats occur on a local scale at a malarial site in western Kenya.

1.1 Malaria overview

About half the world's population is at risk of being infected with malaria, with an estimated 214 million cases occurring in 2015 and most deaths occurring in children in sub-Saharan Africa (World Health Organization, 2016). Sub-Saharan Africa hosts the deadliest malaria parasite, *Plasmodium falciparum*, as well as the particularly effective vector species *Anopheles gambiae sensu lato*. In the last 15 years, malaria incidence fell 37%, but the decline has lagged in the sub-Saharan African countries most at risk. The decline is thanks in large part to effective use of vector control methods (predominantly those that prevent human contact with adult mosquitoes, indoor residual spraying and insecticide-treated bed nets), as well as antimalarial drugs. Unfortunately, both the disease and the vector are quick to evolve resistance to these means of control and there is still substantial progress to be made before malaria is eliminated. Any vulnerable point at which to interrupt the transmission cycle of malaria remains a candidate for interventions in a multifaceted control strategy.

1.1.1 Vector species in western Kenya

In western Kenya, three mosquito species in the *Anopheles* genus are responsible for the transmission of *P. falciparum*. The first two, *Anopheles gambiae sensu stricto* and *Anopheles arabiensis*, are species within the *Anopheles gambiae s.l.* complex of at least seven morphologically inseparable species. Although they can only be distinguished by their DNA sequence, *An. gambiae s.s.* and *An. arabiensis* interbreed very little and have different habitat preferences and behaviors-- *An. gambiae s.s.* preferentially feeds on humans (anthropophilic) and rests indoors (endophilic) while *An. arabiensis* exhibits a wider variety of behaviors, feeding on livestock in addition to humans and resting outdoors more often (Budiansky 2002). Both generally use small, seasonal, sunlit, clear, shallow aquatic habitats during the immature life stages, though *An. arabiensis* may be found using a wider variety of habitats including slow-moving water (Gimnig et al. 2001, Sinka et al. 2010). These habitats are rain-fed and usually directly related to topography and hydrology (the possible exception being certain man-made features). The third malaria vector species is *Anopheles funestus,* which uses larger habitats with emergent vegetation (Sinka et al. 2010).

1.2 Malaria vector habitats

1.2.1 Definition of terms

Anopheles spp. malaria vector habitats, the focus of this thesis, are associated with a variety of terms in the literature. Some are used ambiguously. Here, I will use these terms:

1) **Aquatic habitat**—any permanent or semi-permanent stagnant standing body of water presumed suitable for *Anopheles* spp. habitation, regardless of whether anopheline larvae are found within. In the context of landscape correlates, I will ignore artificial habitats as many other authors have. Aquatic habitats can be readily observed by a visitor to the site. Flowing water is categorically excluded but the distinction of flowing water and standing water may be judged subjectively *in situ,* as very slow flowing water can be considered stagnant enough. Lacustrine bodies of water, for example, Lake Victoria, typically host aquatic habitat at their margins, but are not conceived of here as a habitat themselves as a whole entity. These terms are also used equivalently: potential breeding site, water presence, larval habitat. The appearance of the term “larval habitat” requires close reading as some authors use it for this definition, while others use it for the following definition.

2) **Anopheline larval habitat**—an aquatic habitat (above) that hosts *Anopheline* larvae or pupae. In order to positively identify an anopheline larval habitat, one or more larvae must be observed by sampling using a specialized tool such as a dipper or pipette. Some investigators describe a subset of anopheline larval habitats occupied by a particular species, but many combine all habitats with any *Anopheles* species into a single population for study, as all species are malaria vectors and they can be difficult to distinguish. These terms are also used equivalently: anopheline-positive habitat, larval habitat, larval presence, larvae. The appearance of the term "larval habitat" requires close reading, as some authors use it for this definition, while others use it for the previous definition.

3) **Conditional anopheline larval presence**—whether an aquatic habitat (above) hosts *Anopheline* larvae. A variety of descriptions of this event may appear: the occurrence of anopheline larvae, *Anopheles* presence compared to water, site with *Anopheles* spp. larvae, proportion of habitats positive for *Anopheles* larvae. Most authors' descriptions have required a close reading to distinguish this definition from the previous one.

Bayes' theorem can be used to relate the probability of anopheline larval presence given that there is an aquatic habitat, $P(A|B)$, to the probability of anopheline larvae presence, $P(A)$, and the probability of aquatic habitat presence, $P(B)$.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(B|A)$, the probability of observing an aquatic habitat given that larvae are observed, can be assumed to be equal to 1[1], simplifying the equation:

$$P(A|B) = \frac{P(A)}{P(B)}$$

[1] While larval development requires standing water, *An. gambiae* eggs can survive desiccation for at least several days and may contribute to population build-up when water returns (Beier et al. 1990, Minakawa et al. 2001)

Further, that assumption leads to the re-definition of *P(A)* as the probability of anopheline larval habitat presence. Solving for *P(A)*

$$P(A) = P(A|B)P(B)$$

demonstrates that measuring the probability of anopheline larval habitat presence without measuring at least one of the two probabilities on the right-hand side of the equation provides no information about their relative contributions. For example, consider a study in which anopheline larval habitat presence is determined to be positively associated with a given land-use/land-cover (LULC) class and note that aquatic habitats without larvae were not also observed and analyzed in the study. Several possibilities remain:

1) The LULC class is positively associated with aquatic habitat presence, but has no relationship to the suitability of the same habitats for *Anopheles* larvae.
2) The LULC class has no relationship with aquatic habitat presence, but increases the suitability of the habitats for *Anopheles* larvae.
3) The LULC class is weakly negatively associated with aquatic habitat presence, but is strongly positively associated with the suitability of the same habitats for *Anopheles* larvae.
4) The LULC class is positively associated with both aquatic habitat presence and suitability of the same habitats for *Anopheles* larvae.
5) And the list of possibilities continues.

Because this thesis seeks to understand the distribution of aquatic habitats as a piece of a puzzle that contributes to conclusions about the distribution of anopheline larvae populations in the landscape and their resulting impact on malaria incidence, the relationships between the definitions identified here will be further considered in Chapter 4.

1.2.2 Significance of habitat locations

One approach that has been effective for malaria control in the past and may be incorporated more often into the latest control efforts is larval source management (Fillinger and Lindsay 2011). Additionally, isolating the change of malaria transmission rates due only to a particular intervention is necessary in order to monitor success, and can only be done if the variability in malaria transmission due to other factors such as the heterogeneous distribution of such habitats throughout space and time is well understood (Rejmankova et al. 2013). Therefore, detecting larval habitats at the local scale presents an opportunity to quantify and control vector abundance and distribution at the point where it is most confined, but they are not always easily found—they are small, ephemeral, and often in locations that are difficult to access.

Remote sensing with satellite imagery permits a synoptic view of conditions facilitating malaria transmission (Rogers et al. 2002). Direct detection of mosquito habitats using satellite imagery, even at the highest spatial resolutions currently available, is difficult owing to their small size (many less than 1 m^2) and ephemeral nature (Mutuku et al. 2009). Curran et al. (2000) described the generic framework that allows remote sensing to be integrated into landscape epidemiology for vector-borne diseases such as malaria. Remotely sensed data are used to identify information about variation in the landscape, the landscape variables are related to vector habitat, and the spatial distribution of the vector-borne disease is defined by its habitat. In order for this framework to be applied successfully in the overall context of the disease, the landscape elements critical to the survival of the disease vector must be understood and these elements must be detectable using remote sensors (Beck et al. 1994). While a large body of literature describes the ecological characteristics of larval habitats (see Rejmankova et al. (2013) for a recent review), I will focus here on only those landscape variables that are commonly extracted from satellite imagery.

1.3 Landscape environmental correlates of habitats

1.3.1 Relationship with land-use/land-cover

The earliest studies examining the relationship between remotely-sensed data and malaria focused on the identification of potential mosquito habitats in a very broad sense (Hay, Snow, and Rogers 1998). Starting in the 1970s, vegetative communities known to harbor mosquito species were mapped using color-infrared aerial photography for the purpose of directing larval control (Wagner et al. 1979). Barnes and Cibula (1979) were the first to take advantage of digital multispectral imagers for this task; a few years later, data from the earliest spaceborne sensor platforms (Landsat 1 and 2) were used (Hayes et al. 1985).

More recently, a number of studies including *An. gambiae s.l.* and *An. funestus* in an examination of LULC associations with aquatic and/or anopheline habitats have been conducted with *An. gambiae* at the focus in western Kenya. At a site in the Kakamega district in the western Kenya highlands, farmland is less likely to host aquatic habitats, but the habitats are most likely to be occupied—in a further experiment, these habitats were the only ones to produce *An. gambiae* adults under natural conditions (Minakawa et al. 2005, Munga et al. 2006, Mushinzimana et al. 2006, Munga et al. 2009). Conversely, the streams and swamp categories have many aquatic habitats, but few anopheline larvae. LULC classes at the Kakamega site include farmland, pasture, swamp, forest, streams, shrubs, and roads. At this site, farmland tends to be farther away from the streams. Mutuku et al. (2009) studied the relationship between LULC and both aquatic habitats and anopheline-occupied habitats in Asembo in 2005, finding that mature maize and newly-cultivated fields foster more aquatic habitats than would be expected randomly, and that aquatic habitats decrease with increasing distance from streams. In Asembo, agricultural lands tend to be located relatively close to streams, suggesting that these seemingly opposing conclusions might be reconciled if one were to evaluate the effect of LULC on aquatic habitat presence while controlling for correlated variables such as distance from stream or elevation. One attempt to include both variables in a multivariate model shows no significant relationship for LULC in the model for aquatic habitats, although there is a significant relationship for LULC in the model for anopheline larval habitats (Mushinzimana et al. 2006). When agricultural LULC classes enter a model controlling for topographic wetness index (TWI), soils, distance to stream, and rainfall, the odds of the location having an aquatic habitat increase by a factor of 1.33 (McCann et al. 2014), but the model's Akaike information criterion (AIC) shows only a modest improvement over a model with no LULC included.

1.3.2 Relationship with topography

Topographic variables have also shown considerable promise for predicting aquatic habitat locations and corresponding malaria risk. In the western Kenya highlands, topographic variables are associated with anopheline larval habitats (Mushinzimana et al. 2006). In Zambia, significant portions of the landscape are able to be excluded as not having water pools present using several topographic variables including topographic position index (TPI), but it is difficult to pinpoint where *An. arabiensis* actually occupies these pools (Clennon et al. 2010). Nmor et al. (2013) has found that a variety of topographic variables performed well in the rainy season. Drains, foot-prints, puddles, and swamps are especially predictable and the high-risk area comprises half the study site but detects 80% of the habitats. McCann et al. (2014) have found a moderate negative relationship between TWI and aquatic habitat presence.

1.3.3 Relationship with rainfall

Finally, while LULC and topographic conditions can be used to describe where habitats and anopheline larvae are located, an understanding of rainfall patterns is crucial to describing when they might be found there. Water must be present at a site for a minimum of 10 days under ideal conditions for *An. gambiae s.s.* to develop from egg to adult (Budiansky 2002), though larval mortality is higher at the shortest development time than when the temperature is just a few degrees cooler (Bayoh and Lindsay 2003, 2004).

Gridded satellite precipitation data such as the TRMM (Tropical Rainfall Measuring Mission) products provide information about rainfall at a high temporal and spatial resolution. They are increasingly used successfully to discriminate between conditions favoring aquatic vector-borne diseases and those that do not (Schuster et al. 2011, Xue et al. 2011, Debien et al. 2010). Two studies have examined TRMM datasets for use in models predicting malaria risk with results mixed between successful inclusion as an input variable (Kiang et al. 2006) and rejection as an input variable where precipitation may be too low to affect malaria risk (Adimi et al. 2010). Dambach et al. (2012) tried to predict larvae density (high/low) and adult density (high/low) using a combined remote sensing technique and found that the TRMM rainfall amount in the last 15 days is predictive of each in a univariate test. McCann et al. (2014) studied the relationship of rainfall with aquatic habitat presence and showed that the rainfall total in the previous 30 days is positively related to the aquatic habitat presence in a logistic regression. While other accumulation intervals were better related in other models, all of the rainfall accumulation intervals tested were correlated with each other. This study will re-examine the very same dataset, training a model that includes satellite precipitation estimate data before testing the result on data from another year.

1.4 Predicting habitat locations

1.4.1 Generality of landscape correlative models

Studies discussed here so far have performed reasonably well at prediction, as measured by the area under the receiver operating characteristic (ROC) curve. Comparing only logistic regressions, aquatic habitats can be predicted with the area under the curve (AUC) statistic ranging from 0.73 (McCann et al. 2014, lowest of those tested) to 0.81 (Clennon et al. 2010), while anopheline larval habitats can be predicted with AUC scores ranging from 0.76 (Nmor et al. 2013, lowest of those tested) to 0.85 (Clennon et al. 2010). Li et al. (2011) tested four types of models for generality at a site in western Kenya—logistic regression, spatial logistic regression, artificial neural network, and environment niche factor analysis (ENFA). Although spatial logistic regression was the only model type to incorporate spatial information, all model types except ENFA perform well overall and were able to predict at least 75% of the habitats (sensitivity) at the same site in a different year or season. Logistic regression had the greatest temporal generality, predicting 85% of the habitats on average. The authors write that since the spatial logistic regression explained the spatial structure of the training data well, the prior results of Li et al. (2009) give reason to believe that changing levels of spatial dependency over time limit the generality of this model type, as it may be overly dependent on the training data. The models tested here include only topographic variables, and the generality of a model including a dynamic variable has not been tested explicitly so far, although McCann et al. (2014) compared two models of similar construction and found agreement.

Li et al. (2009) have also explored the stability and spatial structure of the same habitats as the 2011 publication. They found approximately 25% direct overlap using a 20 m grid to measure overlap in aquatic habitats at 6 time points during the dry and rainy seasons of 3 consecutive years. Anopheline larval habitats showed slightly less overlap (15%). A nearest neighbor analysis showed that by starting with a rainy-season, anopheline larval habitat at one time point, one must search 190 m to find the first nearest anopheline larval habitat, on average, if visiting at another time point and that this distance varies with the reference season and year. They concluded that any known map of habitats should be updated frequently due to the low predictive power of the map in another season and year, though there is clustering to the habitats.

1.5 Study objectives

Given the comparatively strong predictive performance of even those environmental models with no dynamic predictor variables in the previously surveyed literature, it is worthwhile to investigate whether dynamic model predictions have achieved a higher utility than prior knowledge of the site in predicting habitat locations at other times. Another question that emerges is whether predictions based on the spatial autocorrelation of the habitats show increased detection power over predictions made with the habitats alone. Though habitats have thus far been too small to be detected directly with remotely-sensed imagery, future spatial and temporal resolution improvements may bring the opportunity to systematically identify a population of remotely-sensed habitats in some proportion to the total population. In this event, interpolation techniques will have the potential to fine-tune environmental correlative models of *Anopheles* spp. habitats to the spatial structure of a local region of study in order to produce better predictions and uncertainty estimates, especially in the vicinity of remotely-sensed habitats.

This study will identify the best prediction map for aquatic habitats in Asembo using the candidate predictors of LULC, topography, rainfall, soils, and prior knowledge of habitats at the site. The best-predicting model will be required to perform well at predicting new locations at new time points in order to be considered generalizable. This study will also quantify the differences in prediction power between the mapping possibilities tested.

1.6 Research questions

This study will answer the following questions:

Question 1: Which is more predictive of aquatic habitat locations—predictions resulting from an environmental model or predictions based on prior knowledge of habitat distribution at the site?

Question 2: What is the pattern of spatial and temporal autocorrelation of the aquatic habitats at this site?

Question 3: What error can be expected in predictions of aquatic habitat presence based on error measuring the elevation and the derived topographic index?

CHAPTER 2

METHODS

2.1 Study site

The study site encompasses the community of Asembo (0°11' S, 34°23' E) in the Nyanza Province of Kenya, located in a lowland region on the north shore of the Winam Gulf of Lake Victoria, approximately 50 km west of Kisumu. The study area is approximately 170 km^2 in size and ranges from approximately 1100 – 1325 m in elevation (Figure 1). The region receives rain throughout the year with two peaks, the "long rains" from March to May, and the "short rains" from November to December. Total annual rainfall averages about 1400 mm per year and daily maximum air temperatures range from 25.5°C – 33.0°C (Phillips-Howard et al. 2003). The study area is rural but densely populated, and features intensive agricultural land use including growing cereal grains (primarily maize) and forage for livestock including cattle and goats. Typical residences are mud-and-stick buildings with thatched or tin roofs arranged in compounds. Larval habitats in the area are characterized by 6 main types: burrow pits, drainage channels, livestock hoof prints, rain pools, tire tracks, and pools in streambeds (Mutuku et al. 2006). Malaria is holoendemic at this site with perennial transmission. Deployment of insecticide-treated bed nets and other interventions reduced the occurrence of malaria and the number of adult mosquitoes inside homes at the site (Lindblade et al. 2004), but the prevalence in children remained at about 40% in 2011 (Hamel et al. 2011).

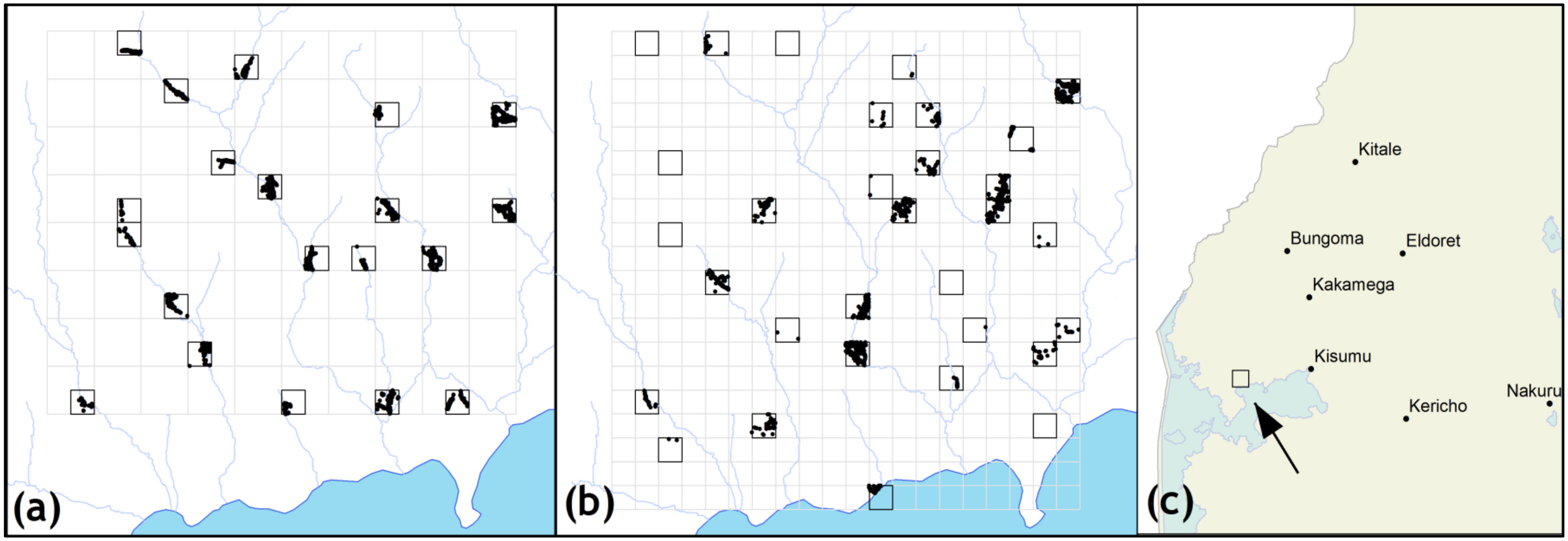

Figure 1 Site overview maps. (a) 2007 habitat survey locations. The lower-left survey grid is the only survey grid shared between years. (b) 2011 habitat survey locations. (c) Overview map showing location of Asembo on the north shore of Lake Victoria.

2.3 Aquatic habitat data

2.3.1 2011 habitat census

These data are the same described by McCann et al. (2014). A spatially-stratified, random sample was used to select 31 rectangular survey grids (500 m by 500 m in size) from a fishnet overlay of the study site. Between May 17 and July 4, each sampling grid was completely surveyed within one day for aquatic habitats and the locations were recorded as georeferenced points with GPS (Global Positioning System). An exhaustive census of aquatic habitats was recorded—each absence observation is verified by direct observation. These data were used for model training.

2.3.2 2007 habitat survey

A survey of larval habitats was performed with repeated visits from the start through the end of the rainy season. Similar to the 2011 survey, 20 rectangular survey grids (500 m by 500 m in size) were selected for the 2007 habitat survey from a fishnet overlay of the study site using a spatially-stratified sample that selected a survey grid in the vicinity of the streams from each stratum. The overall extent was overlapping with, but slightly smaller than that of the 2011 survey. Only one survey grid was in the same location as a survey grid for the 2011 site—all of the other survey grids were exclusive to either the 2007 or 2011 survey. At approximately 2-week intervals, all 20 grids were surveyed for aquatic habitat presence and the habitats were recorded as georeferenced points with GPS. Each location was visited nine times between May 9 and August 31, although different survey grids were observed on different days grouped together for each visit and locations within the same survey grids ere sometimes visited across several days as well. Two survey grids were moved before the third visit—the original locations were disregarded for this analysis and the new locations have missing observations for the first 2 visits. These observations were used for model testing.

For the indicator kriging analysis only, pseudo-absence data were defined. This choice allowed predictions near observed habitats to be non-zero—that is, the predictions are intended to reflect an underlying habitat probability that was not realized as positive observation on that particular day. Each habitat observation included dimensional measurements (length and width)—the maximum dimension plus 5 m was calculated and a buffer around each habitat was generated using this calculated value as the distance. Approximately 360 points, stratified by survey grid, were randomly generated for each visit in the region so that an approximately equal number of presence and pseudo-absence observations were included.

2.3.3 Gridded habitat observations

The point observations were overlaid with the 20 m grid originating from the satellite image used for the LULC classification (described below) in order to spatially align the input datasets in preparation for the logistic regression. Each cell (pixel) that contained at least one point was encoded with a “1” to indicate habitat presence and a “0” to indicate habitat absence. In order to assign observation dates to cells with habitat absence (which were not originally represented in the point data), the modal date of the habitat presence observations within the same survey grid was used—that is, if 90 habitats were observed on July 1, and data collection for that survey grid continued on July 3 with 10 additional habitats discovered, then July 1 was imputed for the absence cells. The imputation introduces the most potential error relative to the total rainfall amount when the rainfall accumulation variable has a short interval. Five survey grids in the 2011 data that contained no habitats at all were eliminated from the training dataset due to an unknown date of observation.[2]

2.2 Land-use/land-cover classification

A radiometrically-corrected SPOT-4 (Satellite Pour l’Observation de la Terre) multispectral image recorded May 13, 2011, with an angle of incidence of 5.61° was obtained and orthorectified with ground control points (RMSE = 1.4 m). SPOT-4 multispectral imagery is

[2] Since no habitats were found, the dates these grids were surveyed were not indicated in the source point data and the dates have not been requested from the original data collector. There are many remaining absence values in the dataset with a wide range of variation of the independent variables to be used in the model training.

collected with 20 m resolution at nadir and is re-sampled to 20 m when collected off-nadir. ISODATA unsupervised classification was used to produce 255 spectral classes in the ENVI software package that were then subjected to spectral separation testing. Class pairs with a transformed divergence value lower than 1990 were deleted and the 32 spectral classes that were distinct were used in a maximum-likelihood supervised classification, encompassing all of the spectral variance in the image using a maximum number of distinguishable classes (Messina and Walsh 2001). In August 2011, 254 ground reference points were collected using GPS. Visual interpretation of the spectral signatures, as well as evaluation of 162 of the ground reference locations were used to assign LULC classes, while 92 ground reference points were held aside for classification accuracy assessment.

The chi-square test statistic was calculated to measure the univariate correlation between LULC classes and habitat presence. For inclusion in the multivariate logistic regression, Class 29 was used as the reference category for odds ratios after the conversion to dummy variables due to its moderately large areal extent and approximately average frequency of habitat presence in the univariate analysis.

2.4 Topographic position index

2.4.1 Custom DEM creation

A 90 m resolution DEM (digital elevation model) was created using a regression kriging model based on two datasets. The first, an SRTM (Shuttle Radar Topography Mission) 3-arcsecond DEM, was downloaded from the United States Geological Survey Earth Resources Observation and Science (USGS EROS) Center. The SRTM 3-arcsecond DEM is a high-resolution, high-precision dataset derived from interferometric radar flown in 2000. Over Africa, 90% of the errors in SRTM heights fall within 11.9 m geolocation error, 5.6 m absolute height error, and 9.8 m relative height error (Rodriguez, Morris, and Belz 2006). A limitation of this DEM is that the radar does not significantly penetrate heavy vegetation canopies, and will provide an elevation measurement within the canopy rather than at the bare earth (Farr et al. 2007).

The second, non-gridded dataset is obtained from a demographic survey in which every household compound in the study community was georeferenced using differentially-corrected GPS (Ombok et al. 2010). Elevation was measured at 10,427 points in this dataset, with 90% of vertical errors below 9.3 m and 90% of horizontal errors below 5.9 m. Because households tended not to be located in low-lying areas, an additional 702 points were recorded in 2011 in the sparsely populated areas, particularly along streams, in order to improve the elevation estimates in these regions.

Universal kriging with external drift was used to predict a new DEM using the relationship between the SRTM and GPS observations. The predictions were derived from the model

$$\hat{\boldsymbol{Z}}(\boldsymbol{s}) = \beta_0 + \beta_1 \boldsymbol{SRTM} + \boldsymbol{\lambda_0}^{\mathsf{T}}(\boldsymbol{GPS} - (\beta_0 + \beta_1 \boldsymbol{SRTM}))$$

where ***SRTM*** is the vector of the SRTM elevations at the locations also measured with GPS, $\widehat{\boldsymbol{Z}}(\boldsymbol{s})$ is the vector of predicted elevations at unmeasured locations, β_0 and β_1 are the coefficients from the generalized linear regression, $\boldsymbol{\lambda_o}$ is a vector of kriging weights derived from a variogram model of the regression residuals, and ***GPS*** is a vector of the target variable (elevation) values at the measured locations. A map of best elevation estimates was developed using interpolation kriging and subsequently used in the logistic regression model for both training and testing (Appendix A.1). Additionally, sequential Gaussian simulation kriging was used to create 100 conditional simulations of the elevation surface by simulating the error surface (the difference between the GPS elevation and the SRTM elevation) and adding it to the SRTM elevation surface. All predictions were performed using the R statistical language version 2.14.0 and the gstat library for R. The interpolated estimates were compared with the predictions from 10-fold cross-validation to evaluate the performance of the regression kriging model.

2.4.2 Topographic position index calculation and scale selection

The resulting DEM rasters were processed using a simple high-pass convolution filter with a kernel size *n* using the ENVI/IDL software packages (Appendix A.2). The filter evaluates the difference between the center cell value and the mean of all cells in the neighborhood. For example, for $n = 5$ (i.e., a 5 x 5 neighborhood), the kernel took the form:

$$\frac{1}{25}\begin{bmatrix} -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 24 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \end{bmatrix}$$

The topographic position index (TPI) value resulted at each cell, with a positive value if a cell was higher than its neighborhood mean, a negative value if it was lower, and a zero value if the cell was in the center of a flat area or on a constant-slope hillside. In order to choose a neighborhood size (scale) for the multivariate model, a range of neighborhood sizes from $n = 1$ to $n = 33$, as well as the original absolute elevation values were tested in univariate logistic regressions with the habitat location data. The best scale was chosen by examining the AUC of the regression predictions with statistically significant relationships. The resulting choice was used as the only topographic input in the logistic regression.

Transformations of the TPI variable were investigated to explore the linearity of the relationship between this variable and the habitat presence logit. The TPI was scaled after calculation into a z-score in order to express relationships in terms of standard deviations above and below the mean TPI.

2.5 Soils

The soils dataset was provided as a polygon shapefile by the International Livestock Research Institute (ILRI). It was digitized from a 1:1,000,000 map produced by the Kenya Soil Survey in 1982. The shapefile was converted to the raster data structure with the same grid as the SPOT-4 dataset. Three soil types are found in Asembo and are described here by their drainage: well-drained, moderately drained, and slow draining. The moderately-drained soil type was used as the reference class for the odds ratios when the soil variable was converted to dummy variables.

2.6 Rainfall time series creation

The TRMM data product 3B42 Version 7 (Daily TRMM and Others Rainfall Estimate) was downloaded for all the months in 2007 and 2011 from the Goddard Earth Services Data and Information Services Center. This gridded data product results from an algorithm that merges TRMM radar measurements of precipitation with other satellite estimates, applies a rain gauge correction for a monthly estimate, and then uses the gauge correction to scale the cells in the 3-hourly 3B42 product. According to Huffman et al. (2007), the resulting products have a spatial resolution of 0.25 degrees latitude by 0.25 degrees longitude and have a nearly neutral monthly bias compared to gauges over land, though the errors in the shorter time scale products (such as 3B42) are large. They recommend that applications that average the data to longer time intervals (whatever is appropriate to the intended application) will have more success with the 3B42 product.

A single pixel in the TRMM dataset extends well beyond the extent of the Asembo study area on all sides (at this latitude, 0.25 x 0.25 degrees covers 27.62 x 27.83 km). Custom accumulation intervals were created by extracting this pixel from each daily image in chronological order and summing the daily precipitation estimates to the appropriate interval for a given calendar date using ArcGIS 10.1 (Appendix A.3). The rainfall variables were joined to the habitat data on the observation date for each presence/absence cell (and visit).

The TRMM rainfall accumulation estimates were compared with the GSOD (Global Summary of the Day) rainfall measurements from the Kisumu airport, approximately 40 km to the east of the study site. The TRMM pixel including the Kisumu airport is the pixel immediately east of the Asembo pixel. This pixel was used to create the custom accumulation intervals in the same fashion as before. The TRMM-Asembo estimates, TRMM-Kisumu estimates, and the GSOD-Kisumu estimates were all evaluated for correlation during the study period.

2.7 Logistic regression model

2.7.1 Model training (2011)

The “glm” procedure in R was used to fit a multivariate logistic regression model. Though logistic regression does not take into account the spatial autocorrelation between observations in the model, it has performed well against other modelling options when tested for generality (Li et al. 2011). The candidate variables for the model included the 32 LULC classes, 3 soil classes, TRMM daily, 1-week, TRMM 4-week, GSOD 1-week, GSOD 4-week, and the 19 x 19 neighborhood (1710 m x 1710 m) TPI. The number of presence observations in each categorical combination was examined to be sure that small cells did not present a significant problem to using logistic regression. The collinearity between variables was also investigated.

The final selection of variables included in the model was determined by the calculation of all possible combinations. The AIC and BIC (Bayesian information criterion) were consulted to determine which variables should be included but were not the final decision on which was the best model.

The model was trained using the 2011 data. The probability cutoff producing a 90% sensitivity rate was selected and the corresponding specificity and accuracy were calculated. This cutoff was selected in order to prioritize habitat detection over the costs of false positives. The AUC was used as a measure of prediction power regardless of cutoff. A binary map using the probability cutoff to split the predictions into presence and absence indicators was created and the geographic area of each resulting binary prediction was calculated. The spatial autocorrelation of the residuals was examined and the nugget:sill ratios for both the residuals variogram and the habitat variogram were compared to learn whether the model accounted for the spatial autocorrelation of the habitats, or whether substantial spatial dependence was still present in the model residuals.

2.7.2 Model testing (2007)

The 2007 dataset was divided so that each of the 9 visits constituted a separate trial of the model predictions. The predicted probabilities were calculated for each visit and the cutoff identified in the model training was applied to create binary prediction maps. The same metrics used to assess the training dataset performance were assessed for each of the 2007 visits: sensitivity, specificity, accuracy, AUC, and the geographic areal extent of the presence/absence predictions. In addition to using the cutoff selected in training, a cutoff that gave a similar sensitivity as the prior habitat autocorrelation predictions was tested to make comparisons between these two results more understandable.

2.7.3 Error propagation mapping

Using the best neighborhood size, the 100 simulated terrain surfaces were used as inputs to a Monte Carlo analysis to evaluate the propagation of the elevation measurement error through the TPI calculation and ultimately into the binary habitat prediction maps resulting from the best logistic regression model. The number of realizations out of 100 that the location was predicted to have a habitat was calculated (interpreted as the probability of predicting habitat) and mapped. Visit 5, the visit with the highest number of habitats, was chosen from the 2007 data for the error propagation analysis after its habitat pattern was shown to be the most predictive of other visits in the habitat stability analysis (described below).

2.8 Autocorrelation analysis

2.8.1 Habitat stability analysis

The 2007 visits were compared with each other for agreement in order to assess how stable habitat locations are over time. The rationale was that if the locations are stable, then using a map from another time point is a candidate for the best prediction map of an arbitrary time point. Each 2007 visit's gridded habitat map was overlaid pair-wise with every other visit's map for a total of 72 permutations. The unit used to express the time lag (t) is the difference in visits—e.g., the pair with visit 1 and visit 3 has t = 2—and each lag represents approximately 2 weeks. The number of matching habitat presence (only) observations were calculated and divided by the number of habitats in the second visit of the pair to calculate the proportion of successful presence predictions (the sensitivity) produced by the first visit. The number of matching habitats does not change when the pair order is reversed, but the sensitivity does. For each visit, two summary metrics were calculated. First, the mean sensitivity when predicting each other visit (8 pairs total) was calculated. Second, the mean sensitivity shown by the 8 other visits when predicting this visit was calculated (the predictability). In addition, for each relative lag, the mean sensitivity was calculated. The predictive success was compared to the predictions made by the environmental model for each visit.

A sum of the 9 visits' binary maps was calculated in order to map the habitat stability. Habitat stability for each 20 m cell was defined as the number of times out of 9 that a habitat was detected in that cell. Since each cell could have contained multiple habitats in a single visit, more than 9 habitats total might have been observed. Habitats were not necessarily the same habitats for all 9 visits, and they were not necessarily different—a single habitat persisting uninterrupted for 9 visits will have the same stability value per this definition as 12 distinct habitats of variable duration as long as at least one was present in each visit for 9 visits.

2.8.2 Indicator kriging stability analysis

Binary "hotspot" maps were created in order to test whether locations near a habitat at one time point are predictive of habitats discovered at another time point. For each visit, the indicator variogram was calculated using the point-based presence observations and the pseudo-absences. The variogram model was identified, de-emphasizing the behavior at larger lags due to the discontinuous nature of the collective spatial extent of the survey grids. Indicator kriging using the variogram model was used to predict the probability of habitat presence at each of the 20 m grid cells. Locations with a predicted probability of habitat presence that was significantly elevated above zero were encoded as "1" and the remaining locations were encoded as "0" in order to create the binary hotspot maps. For the test, the hotspot maps were used in pair-wise, time-lagged comparisons with the actual observed habitats in a similar fashion as before. The resulting sensitivity and specificity values were compared to those of the actual habitat stability analysis and the predictions of the environmental models.

2.8.3 Spatiotemporal variogram

In order to investigate the autocorrelation of the aquatic habitats in both space and time, the spatiotemporal variogram of the 2007 habitat observations was calculated with the "variogramST" function in the gstat library in R. The spatiotemporal variogram evaluates a pseudo-cross-variogram for each spatial lag (h) and each time lag (t) and averages the resulting gamma values over all map pairs with the given time lag (t). The earliest survey date was used as the original input for each visit time value and the function forced the dates into regular intervals. As with the habitat stability analysis, each time lag t is approximately 2 weeks.

CHAPTER 3

RESULTS

3.1 Aquatic habitats

3.1.1 2011 habitats

1,673 point habitat observations were recorded. These were converted to 982 presence cells and 14,269 absence cells in the gridded data. 6.44% of the total surveyed area was therefore positive for aquatic habitats.

3.1.2 2007 habitats

4,997 point habitat observations were recorded over the course of the 9 site visits. These were converted to 3299 presence cells and 106,701 absence cells in the gridded data. On average, 3.00% of the total surveyed area was positive for aquatic habitats in each visit. The number of habitats increased from the survey start in May until early July, when the number of habitats peaked and then decreased until the end of the survey in late August (Figure 2).

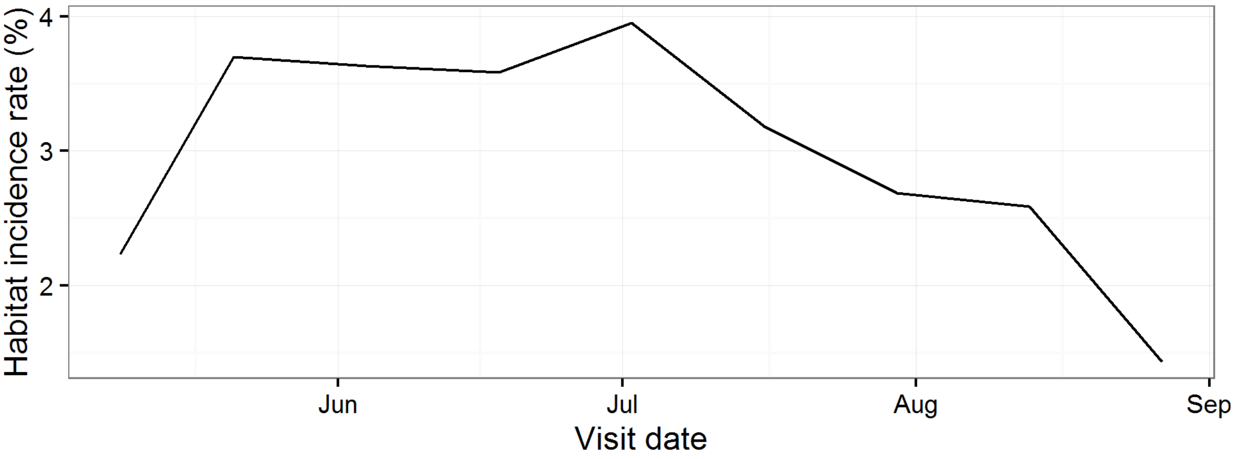

Figure 2 Seasonal rise and fall of habitat incidence rate throughout the 2007 study period.

3.2 Land-use/land-cover classification

There were 33 classes that emerged from the spectrally separable classification. The "unclassified" class contained no habitats and was mostly related to open water so it was combined with the other class for open water (Class 0) so that 32 classes remained for the statistical analysis. The LULC spectral classes were treated individually in the statistical analyses but were assigned nominal class names for interpretation. The interpretative classes described at the site were: open water, bare areas, croplands, shrubland, mixed, built-up, and grassland. Bare areas included highly-disturbed areas, overgrazed lands, and naturally sparsely vegetated areas. Grassland in this classification is grazing lands, natural grasslands, and areas vegetated with mixed grasses and low herbaceous plants that if ever cultivated, appeared to have been left fallow for some time (as opposed to lands that appeared to be seasonally fallow, which were classified as croplands). The 6 "Mixed" classes represent classes in which the pixels are an aggregate of more than one of the other class types, and the 6 classes do not represent the same mix. The classification accuracy was poor with many spectral classes showing considerable confusion (Table 1). The classification's percent correctly classified was 35.9% and the kappa coefficient of agreement was only 13.8%. Only 9 of the classes showed more agreement than disagreement with their ground truth assessment.

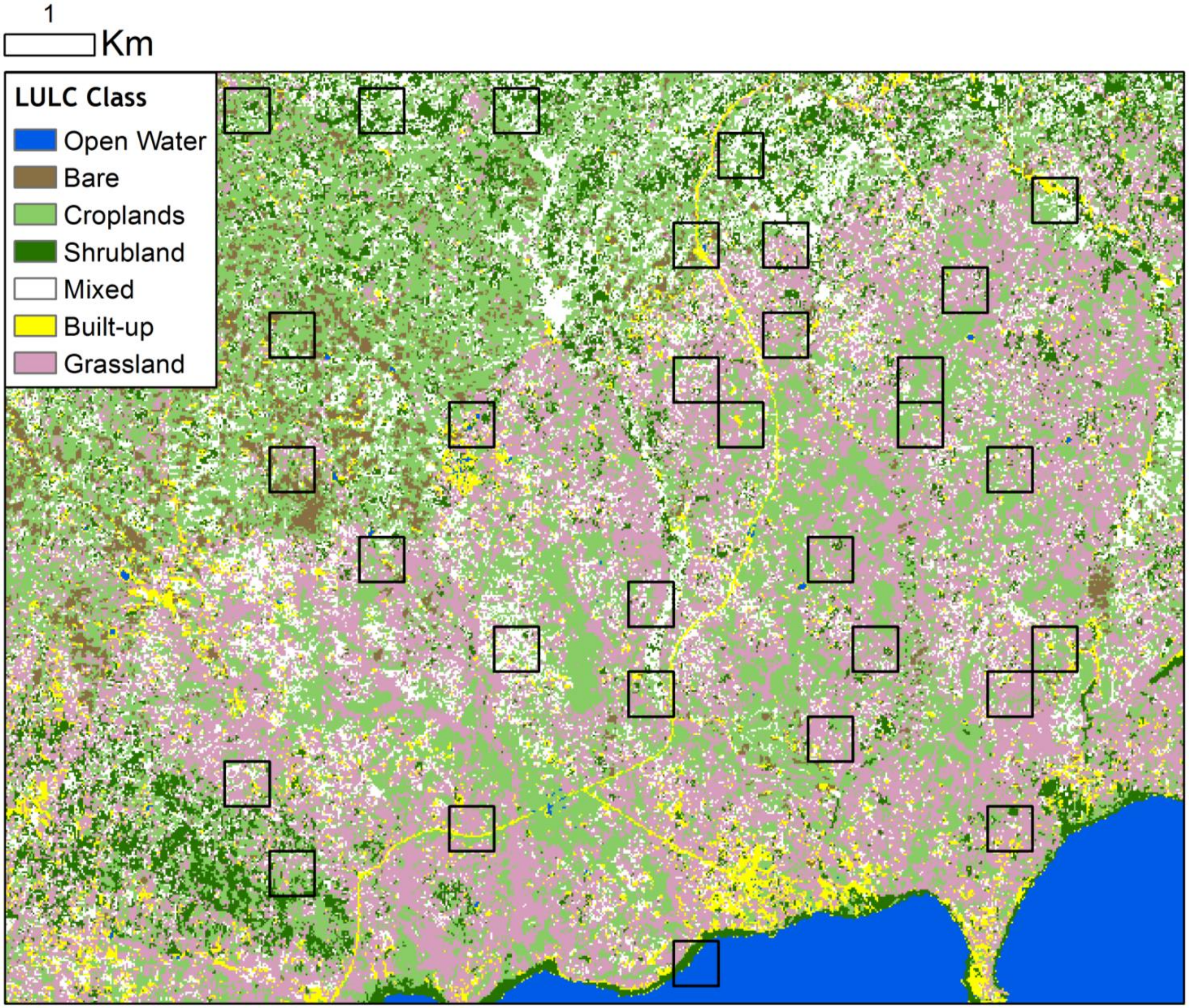

Figure 3 SPOT-4 LULC classification for Asembo, showing locations of 2011 rectangular survey grids.

Table 1 LULC spectral class confusion matrix.

		Ground Truth Class							
Assigned Class #	**Assigned Class**	**Bare**	**Built-up**	**Croplands**	**Grassland**	**Mixed**	**Shrubland**	**Grand Total**	**% Correct (Producer's accuracy)**
2	Bare			1	2			**3**	**0.0%**
13	Built-up			2	1			**3**	**0.0%**
3	Croplands			3				**3**	**100.0%**
7	Croplands			1	2			**3**	**33.3%**
10	Croplands			2		1		**3**	**66.7%**
12	Croplands			1	1	1		**3**	**33.3%**
18	Croplands			1	1		1	**3**	**33.3%**
23	Croplands			2	1			**3**	**66.7%**
28	Croplands				2	1		**3**	**0.0%**
30	Croplands			1	2			**3**	**33.3%**
31	Croplands	1		1	1			**3**	**33.3%**
16	Grassland			1	1	1		**3**	**33.3%**
17	Grassland				1	1	1	**3**	**33.3%**
19	Grassland			1	2			**3**	**66.7%**
20	Grassland			2	1			**3**	**33.3%**
22	Grassland				1	2		**3**	**33.3%**
25	Grassland			1	2			**3**	**66.7%**
26	Grassland			1	2			**3**	**66.7%**
27	Grassland			3				**3**	**0.0%**
29	Grassland				2		1	**3**	**66.7%**
1	Mixed	1		2				**3**	**0.0%**
5	Mixed			2		1		**3**	**33.3%**
6	Mixed				1	1	1	**3**	**33.3%**
14	Mixed				1	2		**3**	**66.7%**
21	Mixed				1	1		**2**	**50.0%**

Table 1 (cont’d)

24	Mixed		1	1	1			**3**	**0.0%**
4	Shrubland			2		1		**3**	**0.0%**
8	Shrubland			1	1	1		**3**	**0.0%**
9	Shrubland				1	1	1	**3**	**33.3%**
11	Shrubland			1			2	**3**	**66.7%**
15	Shrubland			2			1	**3**	**33.3%**
Grand Total		**2**	**1**	**35**	**31**	**15**	**8**	**92**	
% Correct (User's accuracy)		**0.0%**	**0.0%**	**34.3%**	**38.7%**	**33.3%**	**50.0%**		**35.9%**

The univariate chi-square test for association between the LULC spectral classes and the 2011 aquatic habitat presence was statistically significant ($X^2 = 351.9$, df = 31, $p < 0.001$). The soil and LULC spectral classes were also shown to be associated when tested for a correlation that could interfere with the interpretation of the multivariate model ($X^2 = 7715.7$, df = 62, $p < 0.001$).

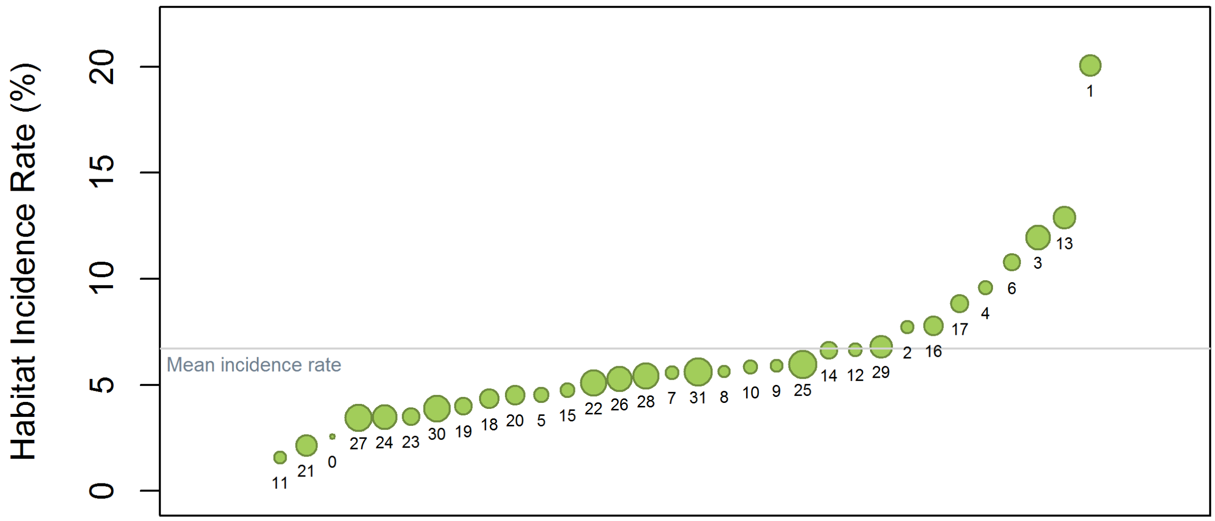

Figure 4 The habitat incidence rate for each LULC spectral class. Each bubble in the chart has an area proportional to the geographic area of its labeled class number.

Spectral Class 1 (Mixed) was by far the most likely to host a 2011 aquatic habitat (Figure 4). Class 1 had a strong negative relationship with TPI, with a mean scaled TPI of -0.92 compared to a mean TPI of -0.23 for cells that were not Class 1 (t = 19.4, df = 618, $p < 0.001$) This class is associated with low places and is clustered near the streams. In the ground reference data collected, it always contained shrubs in a mix with a variety of other LULC classes—3 out of 6 training sites contained shrubs mixed with cultivated land use, while another 2 were composed of shrubs mixed with bare land or overgrazed land approaching bareness. 2 of the 3 sites held out for accuracy assessment also contained cultivated land. The observations I classified as "croplands" frequently contained shrubs as a linear hedgerow that would be included in the pixel, but the “mixed” land cover contained higher proportions of shrubland and/or contained hedgerows along with other LULC classes in the pixel.

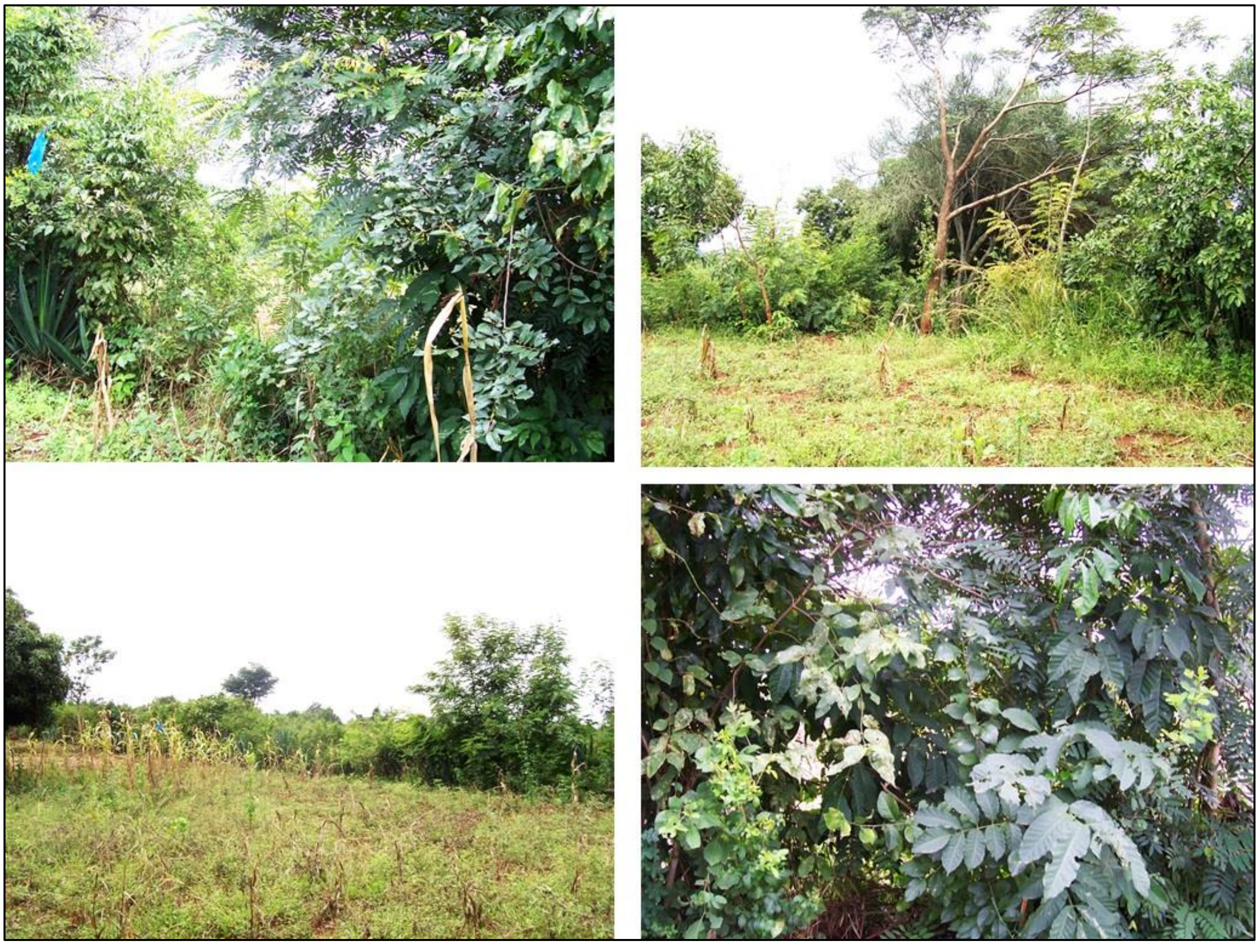

Figure 5 Photographs in 4 directions at a representative Class 1 location, demonstrating a mix of shrubs and agricultural land use.

In the multivariate model (reported further below), Class 13 (Built-up) also emerged as having a significantly strong positive relationship with aquatic habitats when controlling for the other variables in the model. Class 13 was positively related to TPI with a mean scaled TPI of -0.08, compared to a mean scaled TPI of -0.26 for cells that were not in Class 13 (t = -3.76, df = 663, p < 0.001). Class 27 and Class 30 were negatively related to aquatic habitats in the multivariate model.

3.3 Topographic position index

3.3.1 DEM creation

Elevation errors for the input to the simulation kriging ranged from -32.9 m to 25.5 m with a mean value of -2.52 m (positive values where GPS elevation is higher). The errors were not normally distributed and they were spatially auto-correlated in the landscape (Figure 6). Modeling the variogram of the residuals of a linear model predicting GPS from the SRTM elevation values ($R^2 = 0.98$) produced a Gaussian model with nugget = 12.6, sill = 36.6, and range = 247.9 m. Following the modeling of the variogram, the local neighborhood for the kriging was set to use the 60 nearest GPS points or a 1.5 km search radius maximum.

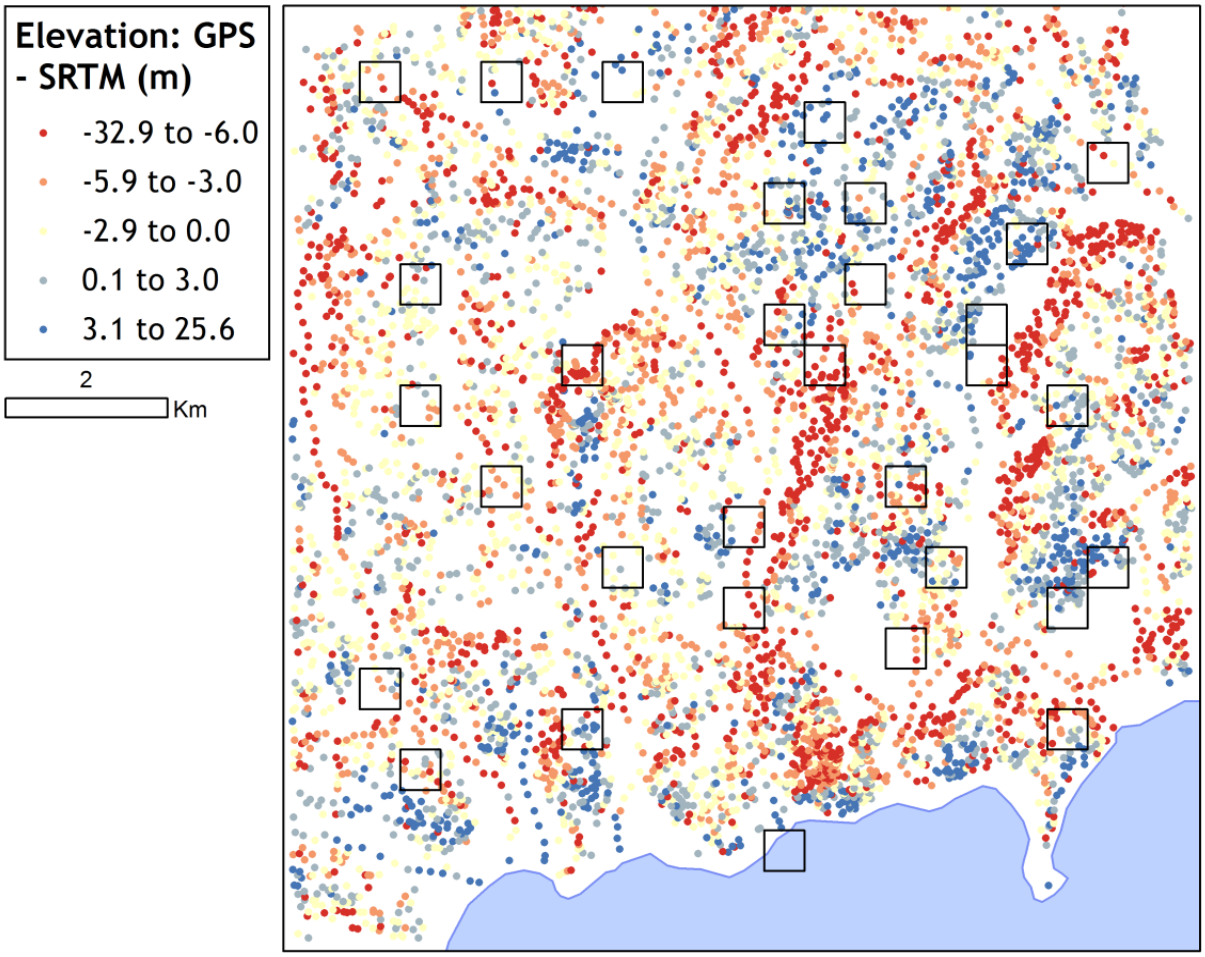

Figure 6 The difference between the GPS height observations and SRTM elevation estimates (in meters). Red values indicate SRTM elevation higher than GPS elevation.

The kriging model's residuals were examined with 10-fold cross-validation. The root mean square error (RMSE) of the cross-validation residuals was 5.13 m—much smaller than the standard deviation of the original GPS elevation measurements, 46.2 m, indicating that accounting for the spatial dependence in the data explains a significant fraction of the variation in the difference between the GPS and SRTM data. Kriging interpolation returns an estimate of the model's uncertainty at each location in addition to the actual prediction. This model was overly-optimistic in estimating the uncertainty on average (the variance returned with the interpolated estimate), with 8.1% of cross-validation predictions falling outside of an envelope 2 z-scores from the observed value (5% are expected at this distance). The DEM resulting from the interpolation kriging ranged in elevation from 1132 – 1365 m in the study area.

3.3.2 Index calculation and scale selection

The results of the scale test showed that the 19 x 19 (or 1.7 km square) neighborhood best matched the scale of the topographic process determining habitat presence and absence, predicting habitat locations in a univariate logistic regression with AUC of 0.867, the highest of all the scales tested (Figure 7).

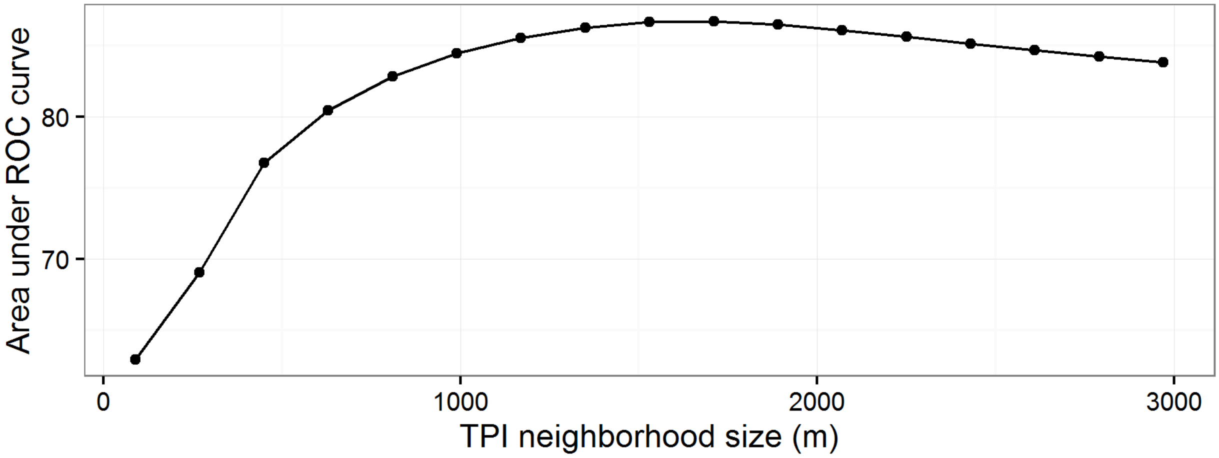

Figure 7 Area under the ROC curve for logistic regression models of aquatic habitat presence along a gradient of TPI neighborhood sizes.

The un-scaled TPI had a mean of 19,803 and a standard deviation of 1.08 x 10^6. The TPI z-score of 0.0182 corresponded to an un-scaled index of 0 (neutral position). From this point forward, “TPI” indicates TPI z-score. The 2011 survey grids captured most of the variation of TPI values at the site, with both high and low extremes falling inside at least one grid (Figure 8). There was a strong negative relationship in a univariate test between the TPI and aquatic habitat presence ($t = 37.1$, $df = 1314$, $p < 0.001$). Locations with aquatic habitats had a mean TPI of -1.07, while locations without aquatic habitats had a mean TPI of -0.19.

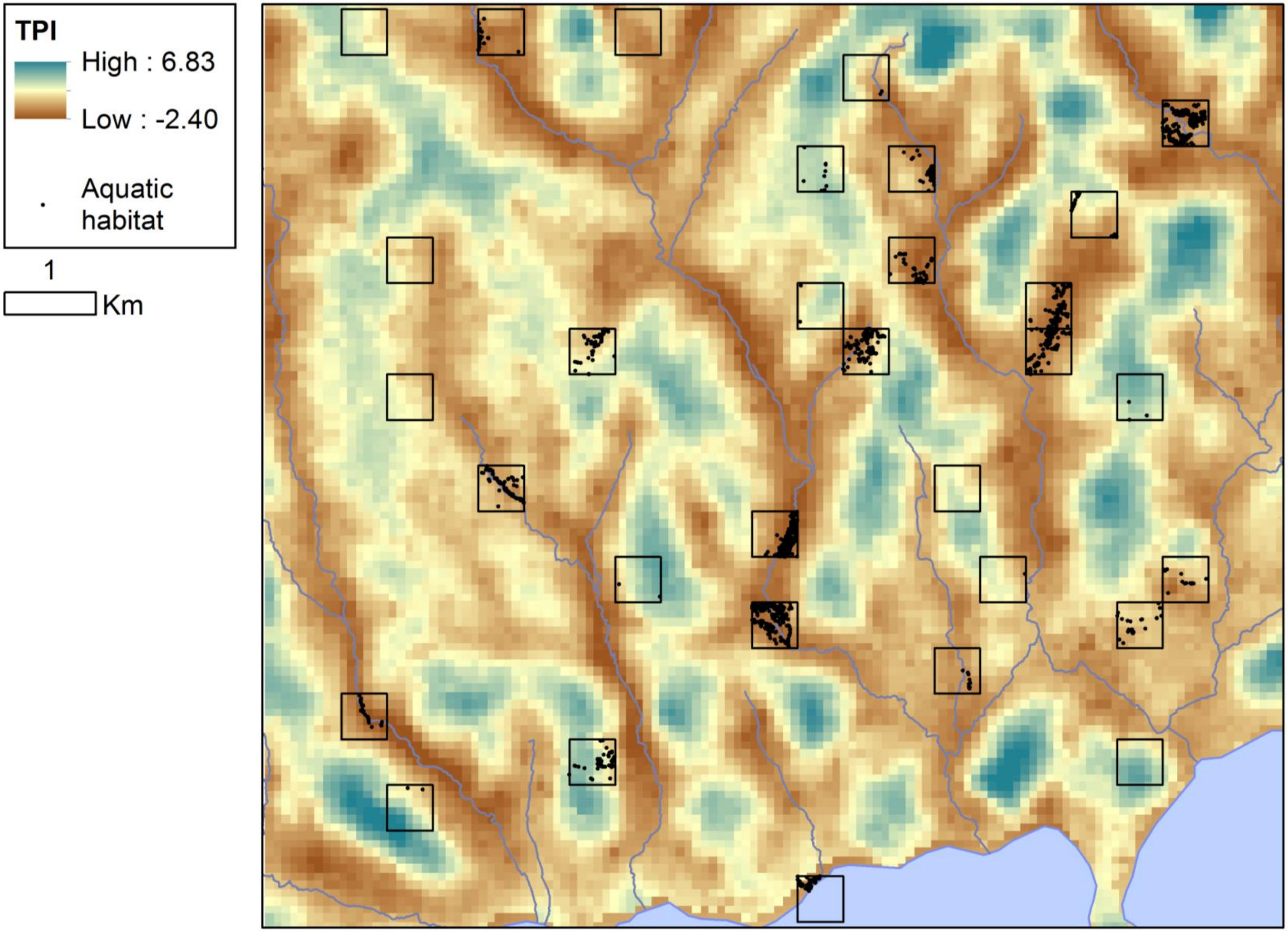

Figure 8 Map of TPI (1710 m scale) with 2011 aquatic habitats overlaid.

3.4 Soils

Among the locations included in the 2011 survey, the moderately-drained soil type dominated, covering 60.7% of the area. The slow-draining soil type covered 19.2% of the southern region of the survey area and the well-drained soil type covered the remaining 20.1%. In a chi-square test of association, there was a relationship between soil type and aquatic habitat presence ($X^2 = 21.7$, df = 2, $p < 0.001$), with the slow-drained soil type having fewer habitats than expected by its area, the well-drained soils having slightly more habitats than expected, and the moderately-drained soils having the highest proportion of additional habitats than expected by its area.

3.5 Rainfall

The TRMM daily rainfall, TRMM 1-week, TRMM 4-week, GSOD 1-week, and GSOD 4-week rainfall variables were the candidate variables representing rainfall for the testing of all possible model combinations. Only one rainfall variable was desired for the model, so out of the models including only a single rainfall variable, the best was selected. The model including TRMM previous 1-week rainfall performed well, resulting in the model with the lowest BIC and AIC out of all the options.

The TRMM-Asembo and TRMM-Kisumu rainfall variables were low-moderately to well-correlated during the course of the 2011 survey, increasing from $r = 0.47$ for the daily rainfall totals to $r = 0.94$ for the 4-week accumulations ($p <0.001$ for all). However, the strength of the correlations between the TRMM-Kisumu rainfall and the Kisumu airport GSOD rainfall did not rise as steadily and was weaker for all accumulation intervals. The 1-week correlation was low ($r = 0.30$, $p = 0.03$), but the 4-week rainfall was not even correlated at all ($r = 0.05$, $p = 0.73$). An examination of the longer-term correlation trend between TRMM-Kisumu rainfall and the GSOD rainfall shows a larger correlation between all rainfall variables: the 1-week correlation rises to $r = 0.49$ ($p < 0.001$) and the 4-week correlation rises to $r = 0.81$ ($p < 0.001$). In the end, the correlation between the TRMM-Asembo 1-week rainfall and the GSOD 1-week rainfall variable (analogous to the 6-day rainfall used in McCann et al. [2014]) was r = 0.37 (p < 0.001) during the longer time interval, which is low, but during the study period the correlation was actually negative, contrary to the expected relationship (r = -0.42, p = 0.002).

3.6 Logistic regression model

3.6.1 Model training (2011)

The model with the lowest AIC included the LULC classes while the model with the lowest BIC included all the other variables but not LULC. The model including the LULC classes was selected as the best model and reduced the deviance from the null model by 1,089 on 35 degrees of freedom (p < 0.001) with an AIC of 6,234. The most important variable in the model with the largest change in AIC was the TPI. Increasing the TPI by 1 standard deviation was associated with decrease in the odds of the location having an aquatic habitat by a factor of 0.30 (Table 2). The previous 1-week rainfall as recorded by TRMM was associated with a 1.02 factor increase in the odds of the location of having a habitat per additional millimeter of rainfall. Spectral LULC Class 1 (Mixed) and Class 13 (Built-Up) were significantly associated with an increase in odds of an aquatic habitat relative to Class 29 (Grassland - a relatively neutral class), while Class 27 (Grassland) and Class 30 (Cropland) were significantly associated with a decrease in the odds of an aquatic habitat relative to Class 29.

Table 2 Odds ratios of best logistic regression model.

	Odds Ratio	Lower CI 2.50%	Upper CI 97.50%
(Intercept)	0.015	0.010	0.023
TPI (scaled)	0.300	0.269	0.332
Soil mod.-drained: well-drained	1.191	0.955	1.488
Soil poorly-drained: well-drained	0.746	0.558	0.997
TRMM 1-week accumulation	1.024	1.017	1.031
LULC class 0: class 29	0.671	0.037	3.282
LULC class 1: class 29	2.385	1.590	3.630
LULC class 2: class 29	1.308	0.680	2.413
LULC class 3: class 29	1.559	1.028	2.394
LULC class 4: class 29	1.380	0.765	2.441
LULC class 5: class 29	0.672	0.322	1.311
LULC class 6: class 29	1.423	0.860	2.349
LULC class 7: class 29	0.976	0.468	1.905
LULC class 8: class 29	0.919	0.422	1.833
LULC class 9: class 29	1.104	0.516	2.193
LULC class 10: class 29	0.905	0.458	1.699
LULC class 11: class 29	0.309	0.073	0.882
LULC class 12: class 29	0.999	0.513	1.856
LULC class 13: class 29	2.583	1.712	3.946
LULC class 14: class 29	1.282	0.735	2.192
LULC class 15: class 29	0.779	0.376	1.511
LULC class 16: class 29	1.252	0.767	2.036
LULC class 17: class 29	1.398	0.855	2.277
LULC class 18: class 29	0.817	0.455	1.424
LULC class 19: class 29	0.547	0.285	0.999
LULC class 20: class 29	0.858	0.482	1.489
LULC class 21: class 29	0.454	0.223	0.859
LULC class 22: class 29	0.877	0.550	1.398
LULC class 23: class 29	0.455	0.229	0.852
LULC class 24: class 29	0.607	0.355	1.020
LULC class 25: class 29	0.802	0.520	1.244
LULC class 26: class 29	0.770	0.482	1.228
LULC class 27: class 29	0.433	0.261	0.709
LULC class 28: class 29	0.884	0.560	1.398
LULC class 30: class 29	0.439	0.268	0.712
LULC class 31: class 29	0.662	0.428	1.029

The AUC was 0.787 with the original training data for the model (Figure 9). The cutoff 0.0285 corresponded to a sensitivity of 90.0% and a specificity of 42.5%. The overall accuracy of the prediction at this cutoff was 45.6%.

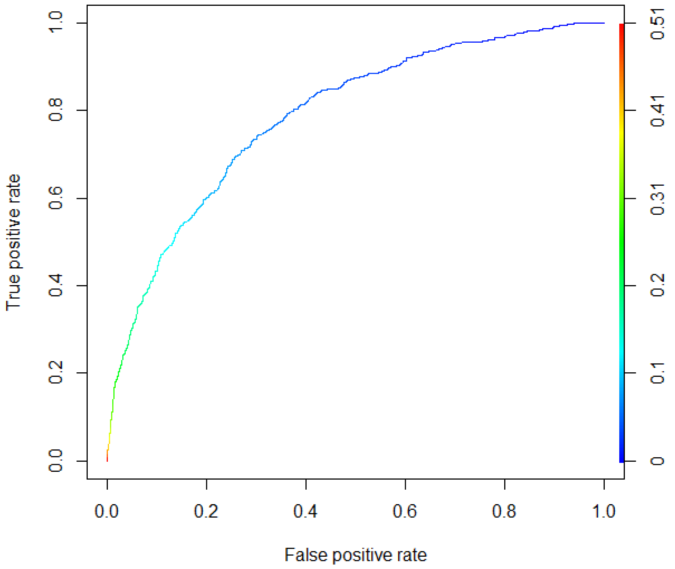

Figure 9 Receiver operating characteristic for the best logistic regression model with original 2011 training data.

A map of the binary predictions for the model shows that 59.6% of the survey area is predicted positive for aquatic habitats with corresponding 90% sensitivity (Figure 10). A map of the false positive and false negatives shows large areas of false positives but rare and somewhat spatially autocorrelated areas of false negatives at this cutoff (Figure 11). Comparing the nugget:sill ratios of an indicator variogram on the aquatic habitat observations and an indicator variogram on the residuals shows a very modest decrease in the spatial dependence in the data, decreasing from 0.47 to 0.48.

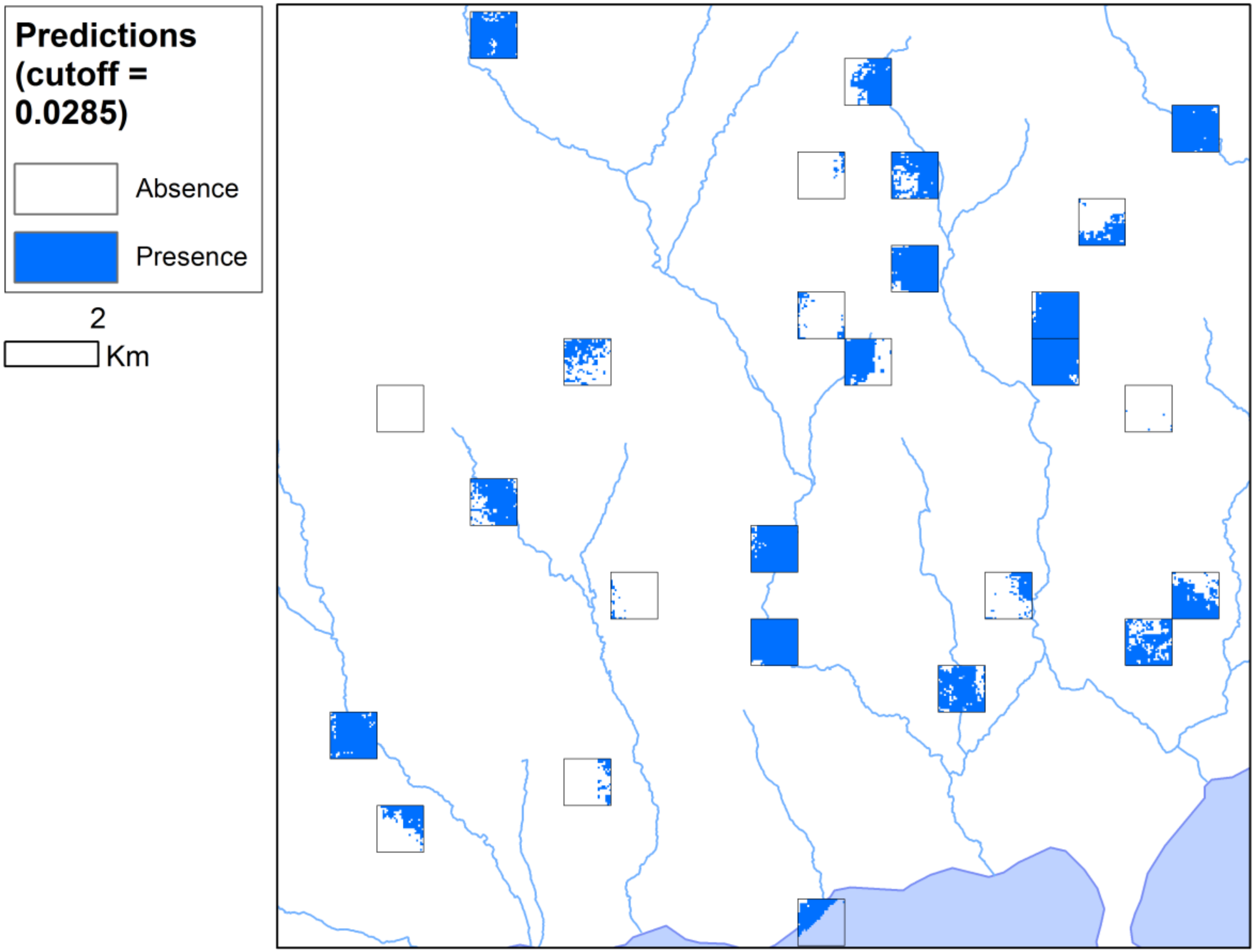

Figure 10 Map of aquatic habitat presence predicted by 2011 training data.

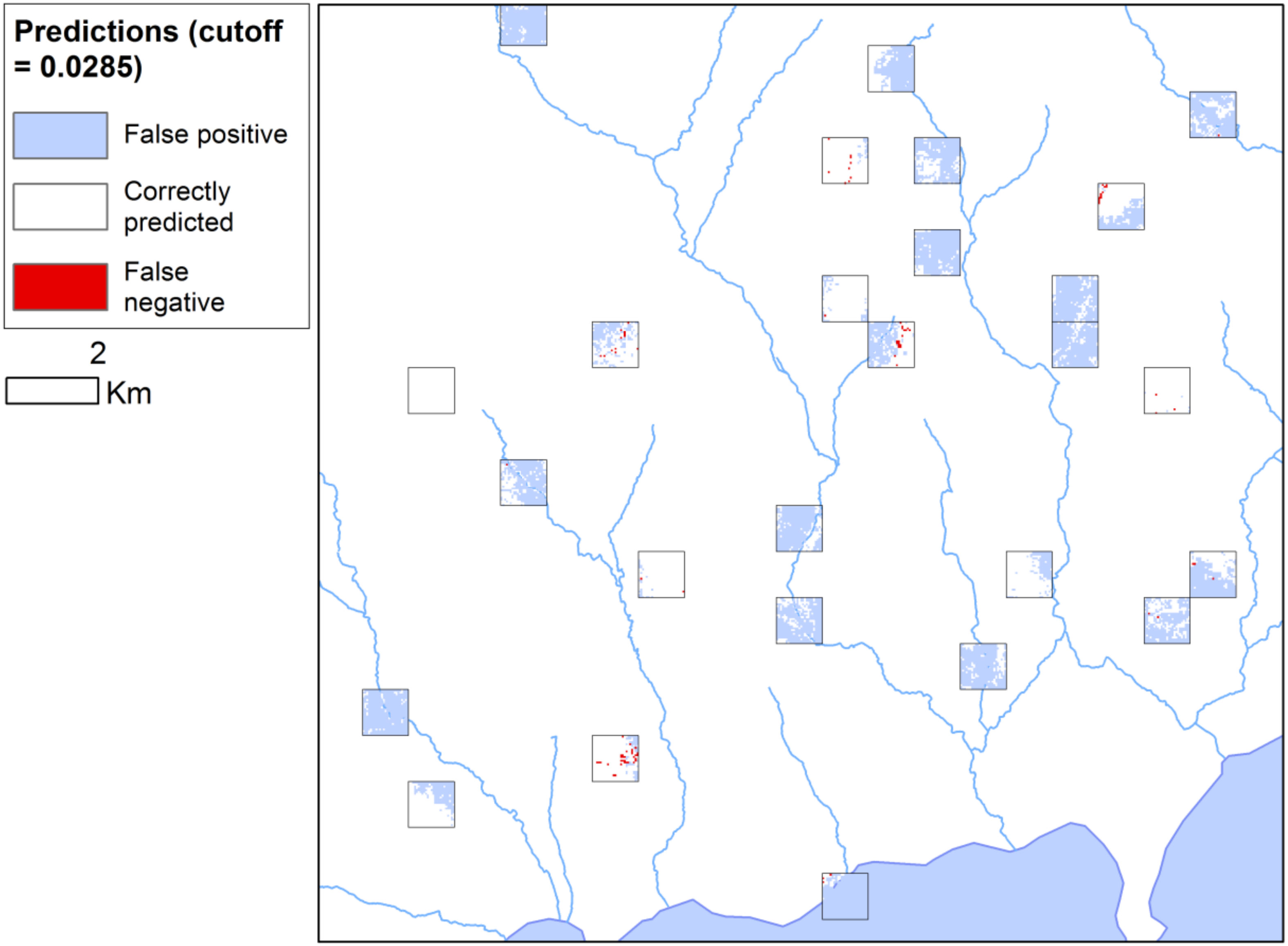

Figure 11 Map of false positive and false negative predictions for 2011 training data.

3.6.2 Model testing (2007)

The performance of the model was somewhat reduced with the 2007 habitats from its performance with the training data. The mean sensitivity for the predictions of all 9 visits increased to 96.8%; however, the mean specificity was 12.9%. The mean AUC was 0.65. The greatest sensitivity was seen with visit 7, the greatest AUC with visit 9, and the greatest specificity with visit 5. The most predictable visit in the autocorrelation analysis, visit 9, was tested again with a cutoff of 0.1277 to generate a sensitivity of 56.9% to match. The corresponding specificity is 75.2% (Table 6). In order to detect 95% of habitats in visit 5, 59.5% of the area was predicted positive for aquatic habitat presence (Figure 12).

Table 3 Comparison of performance of predictions of habitats in Visit 9.

Predictor	Sensitivity (%)	Specificity (%)
Environmental model, cutoff = 0.0285	98.3	12.2
Aquatic habitat presence, another time point	57.2	26.6
Environmental model, cutoff = 0.1277	56.9	75.2
Indicator kriging of aquatic habitats, another time point	80.3	12.3
Environmental model, cutoff = 0.0859	79.9	45.3

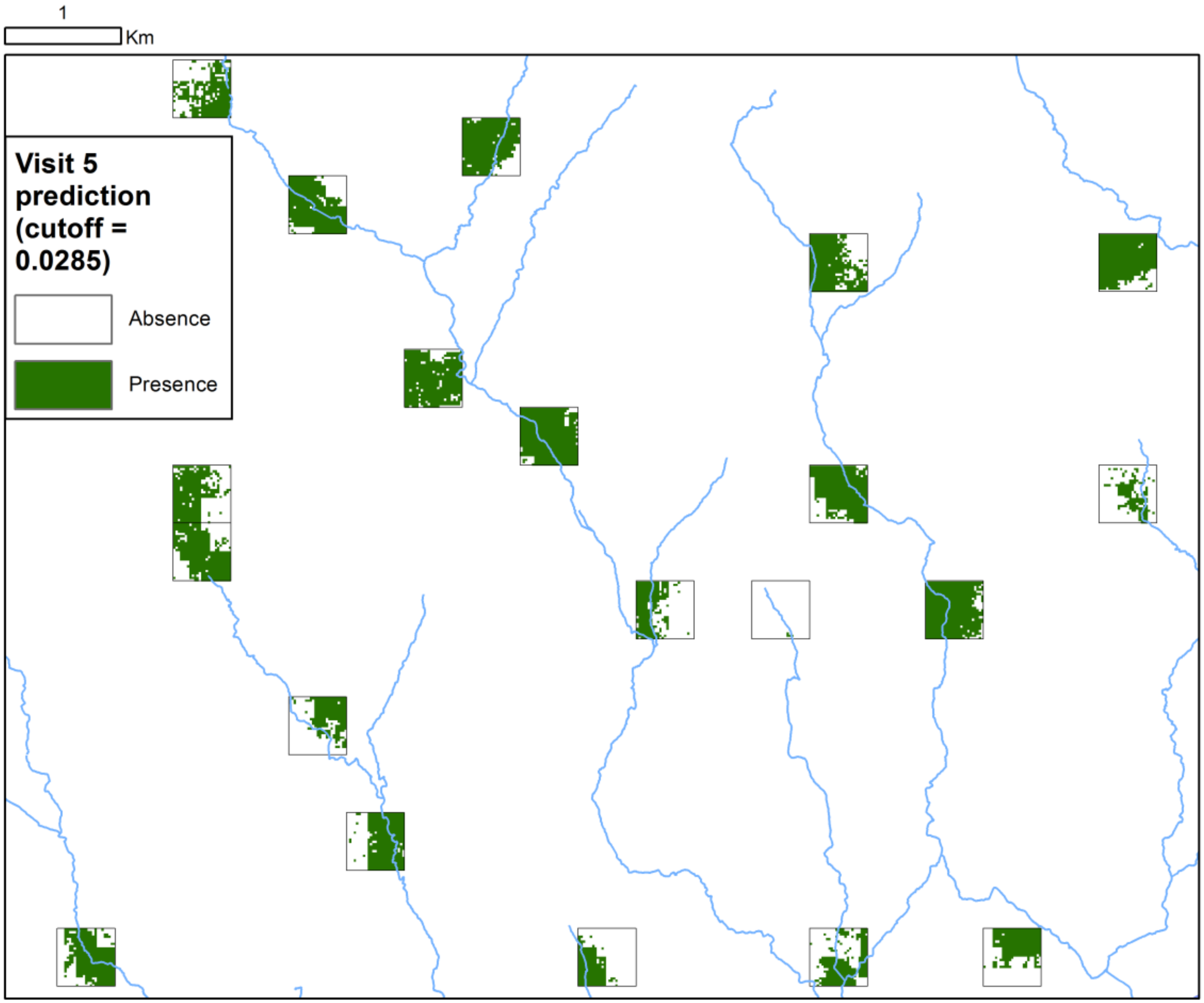

Figure 12 Best logistic regression model predictions in 2007 visit 5.

3.6.3 Error propagation mapping

Three-quarters (74.%) of the area that was predicted positive for habitat presence by the model was more than 95% likely to be predicted positive when the DEM is drawn randomly from a distribution reflecting the spatial structure of the vertical error at the site (Figure 13). Only very limited areas of pure negative prediction (white regions inside survey grids) were discovered at the cutoff used.

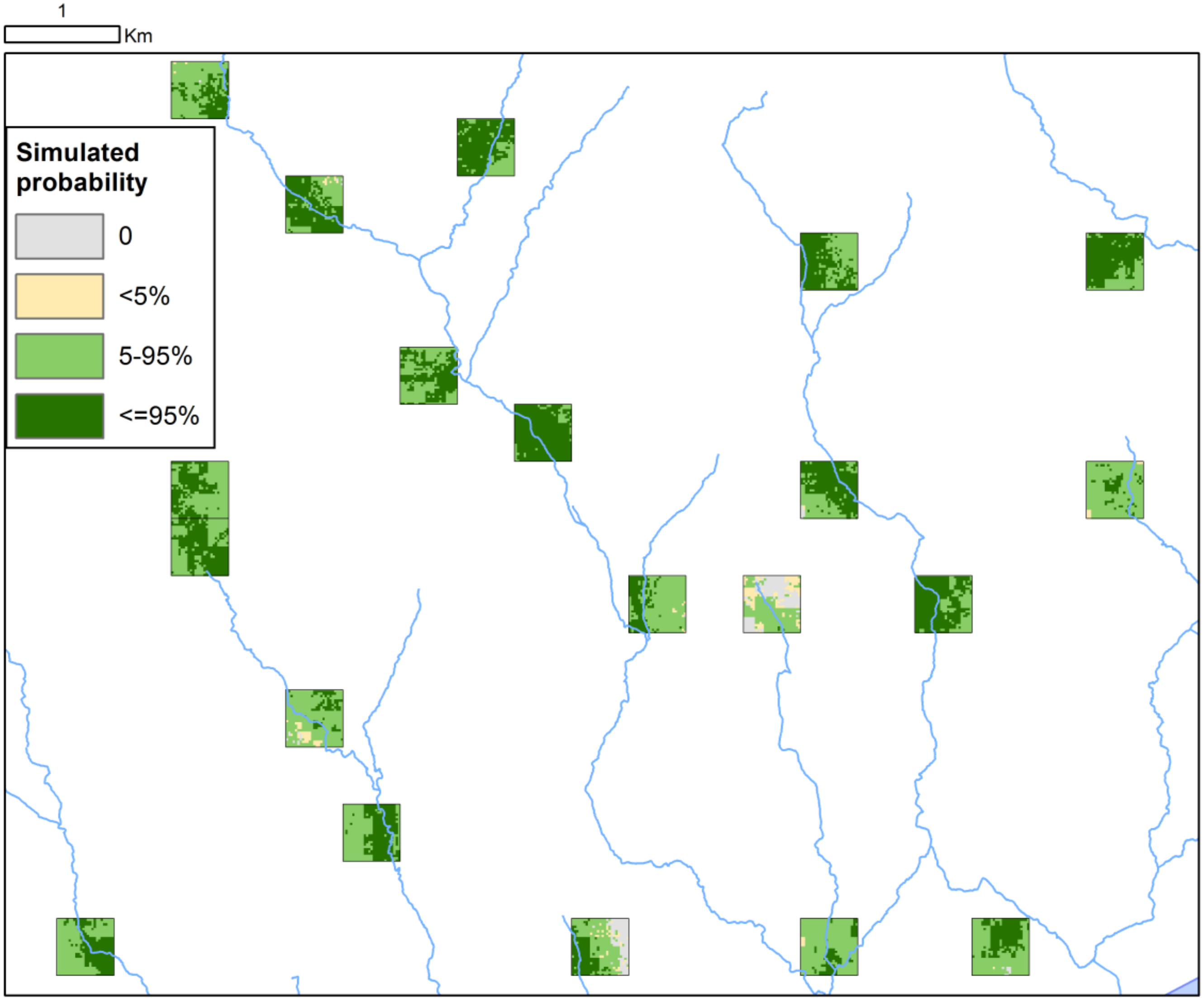

Figure 13 Probability of positive habitat prediction for 2007 visit 5 (with cutoff = 0.0285) when the GPS-SRTM difference is drawn from its random distribution during simulated trials.

3.7 Autocorrelation analysis

3.7.1 Habitat stability analysis

A small proportion of aquatic habitats, 1.7%, persisted during all 9 visits of the 2007 study period (Figure 14). Stable regions appeared to increase in the vicinity of streams. 8.5% of the site area was identified as habitat presence overall. 63.2% of the habitats persisted for at least 2 visits.

The average number of habitat presence grid cells was 367 per visit and the average number of habitats matched in the pair-wise comparison was 171 (Table 4). The habitats observed in visit 5 were the most predictive of habitats in other time points, detecting 61.5% of habitats on average (Table 5) with 44% specificity. The habitats observed in visit 9 were the most predictable habitats—57.2% of the visit 9 habitats were detected again in another visit. The proportion of true positives between pairs declined linearly with each increasing time lag, with a mean of 54.4% at a time lag of a single visit (or about 2 weeks), decreasing to a mean of 25.8% at the maximum time lag, 8 visits or about 16 weeks (Figure 15).

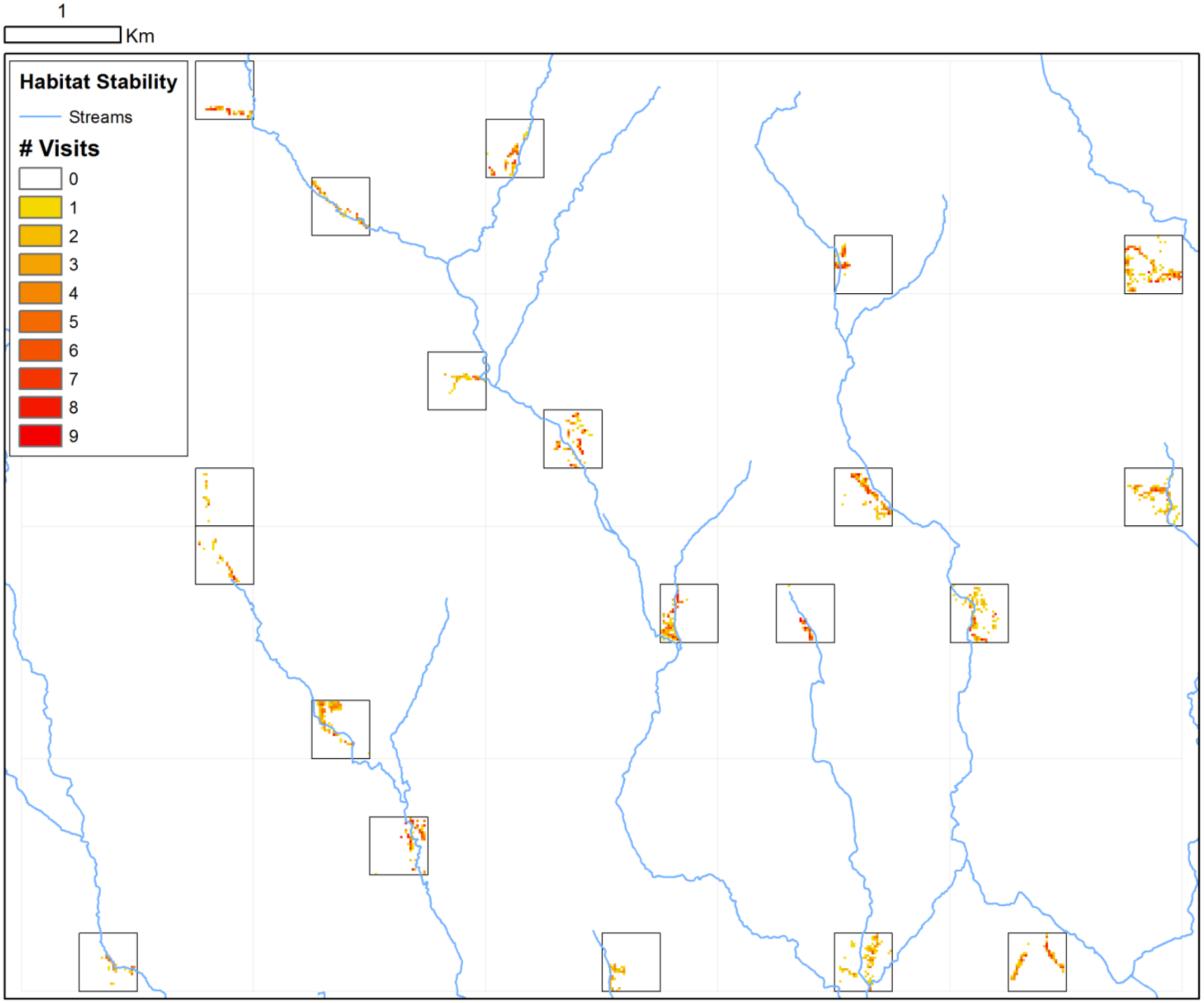

Figure 14 Habitat stability through 2007 study period.

Table 4 Number of overlapping aquatic habitats with 20 m grid cells by visit pair.

Visit	1	2	3	4	5	6	7	8	9	Mean
1	**251**	116	112	104	116	97	97	93	54	99
2	116	**416**	229	224	247	185	168	159	90	177
3	112	229	**454**	269	274	243	204	182	106	202
4	104	224	269	**448**	294	228	197	183	97	200
5	116	247	274	294	**494**	261	215	216	115	217
6	97	185	243	228	261	**398**	230	190	123	195
7	97	168	204	197	215	230	**336**	188	134	179
8	93	159	182	183	216	190	188	**323**	100	164
9	54	90	106	97	115	123	134	100	**179**	102
Mean	99	177	202	200	217	195	179	164	102	**171**

Table 5 Percent agreement for habitat presence in pair-wise comparison between predicting visit and predicted visit.

	Predicting visit									
Predicted visit	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Mean**
1	100.0	46.2	44.6	41.4	46.2	38.6	38.6	37.1	21.5	39.3
2	27.9	100.0	55.0	53.8	59.4	44.5	40.4	38.2	21.6	42.6
3	24.7	50.4	100.0	59.3	60.4	53.5	44.9	40.1	23.3	44.6
4	23.2	50.0	60.0	100.0	65.6	50.9	44.0	40.8	21.7	44.5
5	23.5	50.0	55.5	59.5	100.0	52.8	43.5	43.7	23.3	44.0
6	24.4	46.5	61.1	57.3	65.6	100.0	57.8	47.7	30.9	48.9
7	28.9	50.0	60.7	58.6	64.0	68.5	100.0	56.0	39.9	53.3
8	28.8	49.2	56.3	56.7	66.9	58.8	58.2	100.0	31.0	50.7
9	30.2	50.3	59.2	54.2	64.2	68.7	74.9	55.9	100.0	57.2
Mean	26.4	49.1	56.6	55.1	61.5	54.5	50.3	44.9	26.6	**47.2**

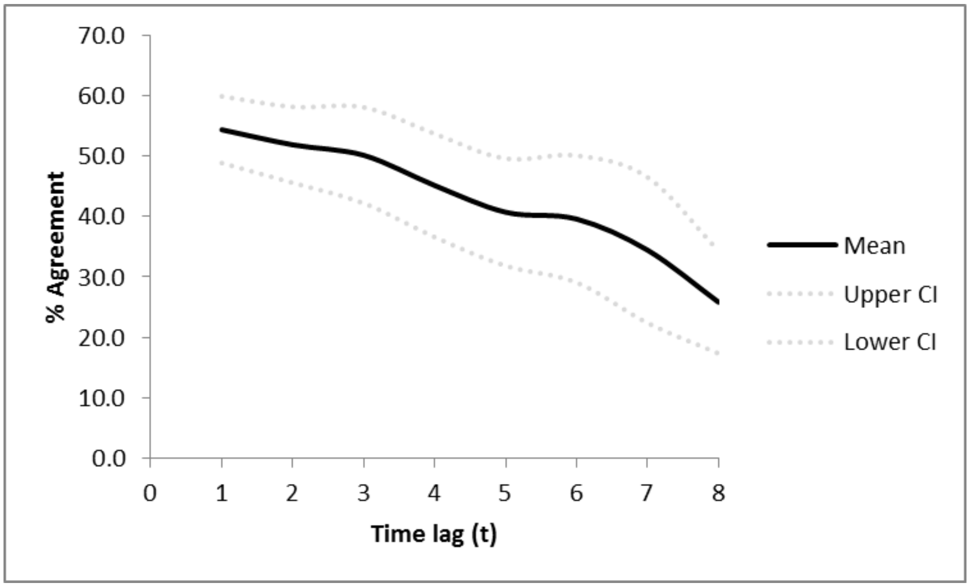

Figure 15 Habitat observation overlap/agreement as a function of relative time lag.

3.7.2 Indicator kriging stability analysis

Indicator kriging identified 21.6% of the site as habitat hotspots. 76.0% of the hotspot locations persisted for more than one visit. 2.1% of the hotspots were completely stable for the whole season, but it was not uncommon for a completely stable hotspot to contain no actual habitat observations throughout the 9 visits. No hotspot map predicted 100% of the observed habitat positives correctly, with each map missing just a few observed habitats (Table 6). This was due to pseudo-absence locations being permitted close enough to the aquatic habitats that both a pseudo-absence and a presence point could fall in the same cell. The growth in the area of the hotspots toward the middle of the study period is shown in Figure 16. The indicator kriging increased the sensitivity of the predictions to a mean value of 67.7%. The specificity was low, however, with a mean of 24.6%, indicating that a large proportion of predicted positives using the "hotspots" were false positives. Visit 5 was again the most predictive, and visit 9 was the most predictable. In visit 5, 12.5% of the site area is predicted to be habitat by the hotspots in order to detect 85% of the actual habitats.

Table 6 Performance of indicator kriging predictions vs. habitat observations.

Sensitivity

	Kriging prediction visit									
Observed visit	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Mean**
1	96.8	71.3	67.7	69.3	68.9	66.1	61.4	55.8	27.6	61.0
2	40.4	99.0	80.5	79.6	82.2	77.9	65.9	59.6	26.2	64.0
3	33.5	71.6	98.5	81.3	83.9	80.6	66.1	60.8	28.9	63.3
4	33.3	68.3	81.9	99.6	86.2	82.1	65.0	62.3	25.7	63.1
5	33.6	72.9	81.6	83.2	99.4	81.4	67.4	63.2	29.4	64.1
6	34.2	71.9	85.4	82.4	88.4	99.5	78.9	69.6	37.2	68.5
7	40.8	76.5	83.3	83.3	90.2	92.3	99.4	78.6	44.0	73.6
8	37.8	74.0	82.0	82.0	88.9	87.9	82.0	99.7	36.5	71.4
9	41.3	76.5	86.0	82.7	91.1	95.0	93.3	76.5	95.5	80.3
Mean	36.8	72.9	81.1	80.5	85.0	82.9	72.5	65.8	31.9	67.7

Specificity

	Kriging prediction visit									
Observed visit	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Mean**
1	55.4	13.4	13.1	10.6	11.0	10.7	14.8	13.9	23.6	13.9
2	38.3	30.9	25.7	20.1	21.8	20.8	26.2	24.6	38.4	27.0
3	34.6	24.4	34.3	22.4	24.2	23.5	28.7	27.3	46.1	28.9
4	33.9	23.0	28.2	27.1	24.6	23.6	27.9	27.6	40.5	28.7
5	37.8	27.0	31.0	24.9	31.2	25.8	31.9	30.9	51.1	32.5
6	31.0	21.5	26.1	19.9	22.4	25.4	30.1	27.4	52.1	28.8
7	31.2	19.3	21.5	17.0	19.3	19.9	32.0	26.1	52.1	25.8
8	27.8	17.9	20.4	16.1	18.3	18.2	25.4	31.9	41.5	23.2
9	16.9	10.3	11.8	9.0	10.4	10.9	16.0	13.6	60.2	12.3
Mean	31.4	19.6	22.2	17.5	19.0	19.2	25.1	23.9	43.2	24.6

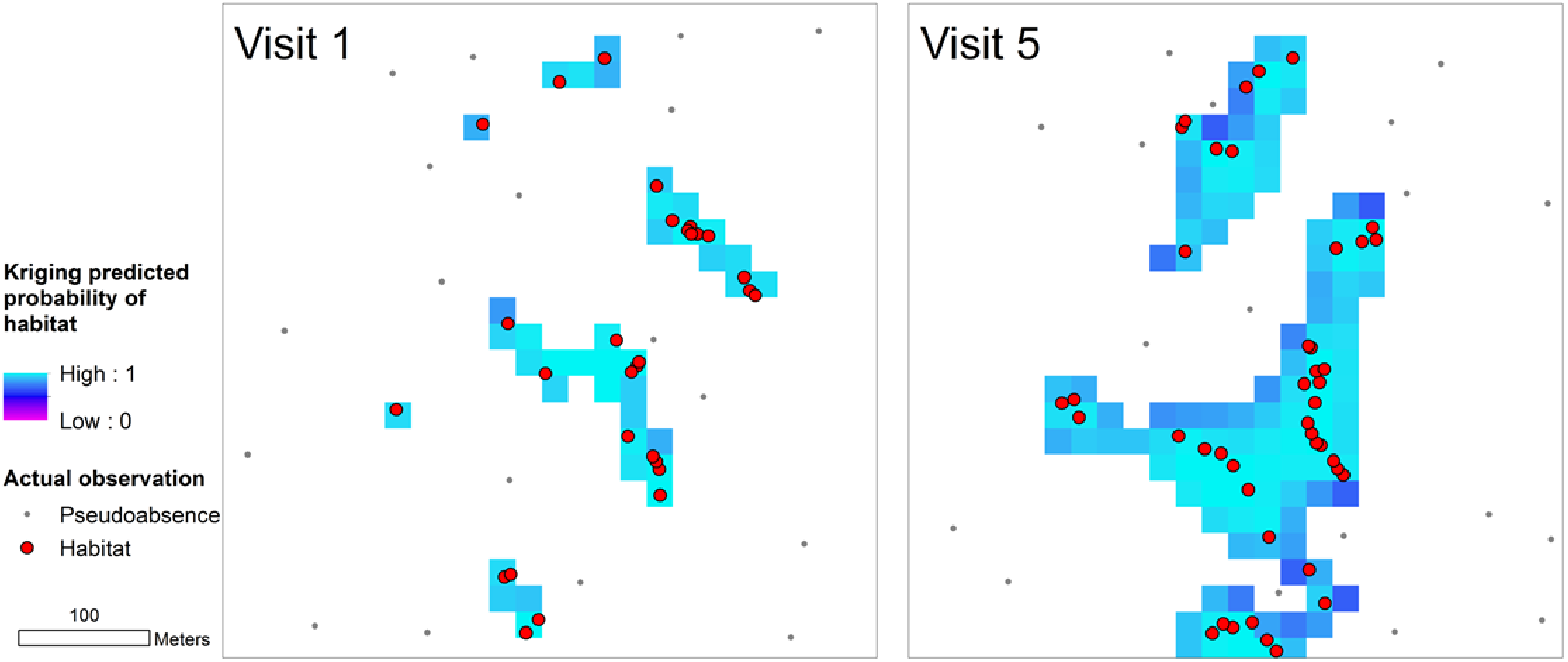

Figure 16 Maps of indicator kriging "hotspots" in a single 500 m x 500 m survey grid compared for two visits. The increase in the turquoise/blue area shows the growth in the area of the hotspot from the beginning of the study period (Visit 1) to the middle of the study period (Visit 5). The points show the distribution of the actual habitat observations as well as the generated pseudo-absences.

3.7.3 Spatiotemporal variogram

The spatiotemporal variogram shows considerable cross-correlation at each relative time lag. The nugget effect is slight smaller for time lags 1 – 3 compared to time lags 4 – 7 but at 30 m spatial lag the time lags 1 – 3 are not any more correlated than the lags 4 – 7. The variogram shape is similar for each pooled time lag. Time lag 8 (visit 1 and visit 9 pair only) is always more highly cross-correlated at all spatial lags than the other time lags. About half of the overall variation is spatially random for all time lags. All time lags show spatiotemporal dependence, with minimal cross-correlation at 180 m spatial lag. Habitat presence is more highly cross-correlated at all spatial lags in time lag 8 than in any other time lag.

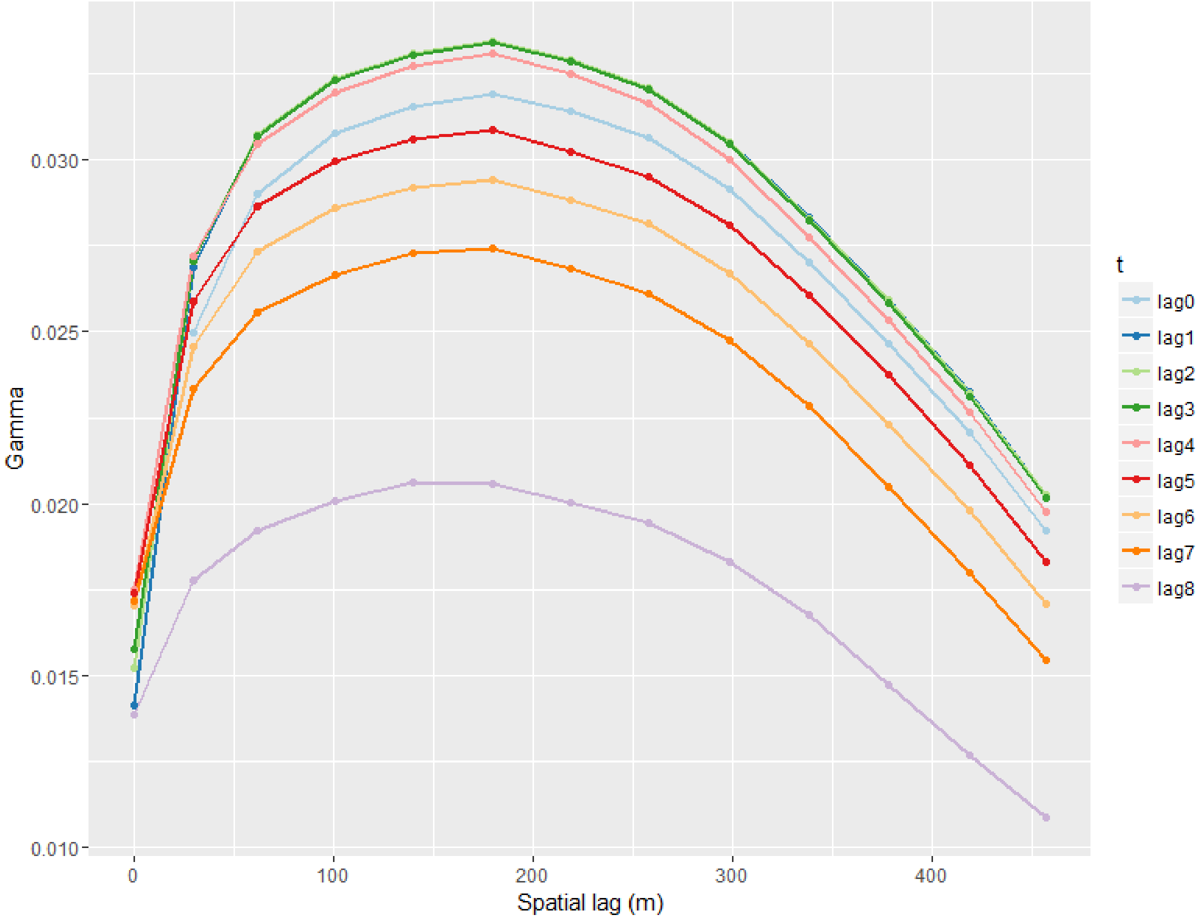

Figure 17 Spatiotemporal variogram for 2007 habitats.

CHAPTER 4

DISCUSSION

4.1 Aquatic habitats

Though the habitats were measured individually when the data were collected, ultimately this model does not measure the existence of a singular habitat with each observation. For example, the model does not delineate the bounds of a single hoof print, but identifies a location, about 20m in size, where one or many hoof prints are likely to be found, or any other habitat type is likely to be found. This is a specification of the model, not necessarily a limitation, but a related limitation is in the definition of the dependent variable. Not only does the model combine habitat definitions for three anopheline species, but even the habitats of just one of the species are variable in type and size. They originate from multiple processes which may not act at the same scale; even one process, such as topography, may not be acting at only one scale to produce to habitat pattern seen. Only the land-use/land-cover variable here attempts to capture the human activities in creating the habitats.

4.2 Land-use/land-cover classification

The 32 spectral classes created were difficult to translate to meaningful "on-the-ground" classes. A semantic LULC class, for instance, maize cultivation, almost certainly participates in multiple spectral classes. This situation could be handled by merging spectral classes after their semantic class is determined. However, another challenge is that a spectral class may also be associated with multiple semantic classes—for instance, the strong soil signature that is associated with overgrazed land, residential land, or land in a transitional cultivation status. This complexity challenges every LULC classification; however, for three reasons it is particularly troublesome for my classification.

First, the 20 m resolution of the SPOT-4 image is coarse compared to the size of the contiguous LULC patches seen at the site. In the residential and agricultural land use areas in Asembo, the pattern is highly heterogeneous. Agricultural fields are small and frequently delineated by hedgerows consisting generally of medium-height woody shrubs. Even “large” patches specially noted during ground reference data collection were as little as 40 m by 40 m—only 4 pixels in the SPOT image. With many mixed pixels, the shrubs of the hedgerows mixed with the coarse agricultural vegetation of maize and millet within a pixel was difficult to distinguish spectrally from the natural shrub and forbs cover during classification.

Second, the timing of the ground reference data collection was three months after the collection of the SPOT imagery. Typically, evidence of the recent land use was distinguishable in the field—evidence of maize that had been growing in May was provided by the stalks lying on the ground in August. However, for lands that appeared to be fallow, certain identification was more ambiguous for a transient visitor to the site. Cloud cover contributed to the selection of the timing of the SPOT image. Field work timing for the ground reference data was instead selected based on logistical concerns.

Third, the spectral separation technique used to classify the image attempts to maximize the amount of information extracted from the data by dividing the variation into the smallest, closest, but still statistically distinguishable clusters that can be detected. The results here show that the clusters are not always semantically distinguishable. The number of classes created with this method generates a high burden for the number of ground truth observations, which I was not able to fulfill within the available time for field work. This method may be better suited to regions where semantic distinctions among spectral classes are stronger, such as those with less heterogeneous LULC patterns or a better match between the patch size and the sensor spatial resolution. After difficulty producing a suitable standard of accuracy while maintaining the 32 individual spectral classes, two grouping schemes were attempted. In the first alternate scheme, the classes were grouped by the interpretative classes (cropland, grassland, etc.). In the second alternate scheme, the classes were sorted into a binary grouping in which one final class contained all of those spectral classes positively associated with aquatic habitats in a univariate test, and the second final classes contained all those negatively associated with aquatic habitats.

Neither of these alternate schemes was successful, and so the spectral classes were interpreted individually.

The class with the greatest odds increase of habitat presence (controlling for the effect of soil, rainfall, and most importantly TPI) could only be described as “Mixed” in the nominal class assignment. Its characteristics show that it usually contains shrubs and also another LULC class, typically cultivated land or overgrazed land. This class may be capturing a particular human disturbance pattern around on the fringe of agricultural land uses where the transition is due to landscape characteristics near streams (wetness, unevenness) that deter human use but promote ponding. That is, homogenous regions of cultivated land (primarily maize) do not lead to the creation of habitat, but the chaotic periphery of these areas where travel takes place or barriers and transitions near cultivation-suitable regions are found is by itself involved in a heterogeneous process generating cultivation-adjacent habitats.

The spectral class with the second-greatest odds increase is the built-up LULC class. Despite a poor performance in the accuracy assessment, this second spectral class has a distinctive pattern when the site image is viewed in entirety, showing a clear association with the tarmac highway and market centers. Even with this clear association, the spectral signature of this class shows clear signs of vegetation, indicating that most of the built-up class participates in mixed pixels with the surrounding vegetation and also includes patches of vegetated, but overgrazed land. It is unlikely that the tarmac includes many habitats, but the other roads, built-up areas, and disturbed or overgrazed lands contain many habitats that result from human, livestock, and vehicular activity. Class 27 (Grassland) and Class 30 (Cropland) are negatively associated with aquatic habitat presence, but neither performed well in the accuracy assessment or in the classification training. Both demonstrate the considerable confusion between the grassland and croplands LULC classes in this LULC classification.

Others have concluded that the performance of LULC may not merit the cost and effort, especially compared to topographic variables when one is only identifying aquatic habitats and not anopheline larval habitats (Clennon et al. 2010, McCann et al. 2014). The studies that have measured anopheline-positive habitats or conditional larval presence in addition to aquatic habitats commonly show LULC variables having no relationship with aquatic habitats, but coming in to the models once the larval presence is included (Clennon et al. 2010, Mushinzimana et al. 2006).

Minakawa et al. (2005) note that farmland is positively associated with conditional larval presence but there is some additional information that can be derived from their Table 1 regarding the aquatic habitats. This information provides a more complete interpretation when the relationships demonstrated with Bayes' theorem in chapter one are applied. Farmland is negatively associated with aquatic habitat presence, and in fact is negatively associated with anopheline larval habitat presence in their univariate test. That is, the negative effect of farmland on the total aquatic habitats is so strong that it overpowers the positive effect of farmland on conditional larval presence such that the net effect is a negative relationship with anopheline larval habitats as well. However, since farmland is the dominant LULC class at their site (60%), the end result is that the largest fraction of anopheline larval habitats (39%) was found in farmland. That is, if you seek to answer the question, "which LULC class explains the anopheline characteristics of this site?" you might answer "farmland" and cite the two facts that farmland is the highest LULC class at the site and the aquatic habitats within it are highly suitable for anopheline larval presence. However, if the goal is to convert land areas to land use less favorable for anopheline larvae, one should actually avoid farmland, at least at that site. Each of the habitat definitions related to malaria vector habitats measures a phenomenon that has something different to contribute to the applied use of habitat predictions, including cost-benefit analysis for control. LULC classes that contain highly suitable habitats but fewer aquatic habitats might be easy to manage with larval control measures despite their large areal extent, or an LULC class with many aquatic habitats might also have highly suitable habitats and so the anopheline population can be most effectively reduced by focusing on a small geographic area.

Some LULC classes may be associated with habitats simply by virtue of being found in the lower sites or other topographically favorable locations. At this site, (Mutuku et al. 2009) previously noted that the proportion of farmland is higher closer to the streams. Since there was no multivariate model in that study, it is difficult to conclude whether LULC classes or distance to stream better explained the aquatic habitat presence. Here, Class 1 (Mixed) is found in low places as well, but no collinearity problem was detected in the model. It seems that there is still a strong positive relationship with aquatic habitats once the TPI is controlled for.

4.3 Topographic position index

The results shown here are in agreement with the recent literature that shows topographic indices of various forms dominating the multivariate models in which they are tested. Where previous studies have tested the 500 m and 2 km TPI scales, here I have shown a systematic testing of the relationship between the scale of this topographic index and the pattern of habitat presence. The very local neighborhood kernel of 3 x3 (270 m square) showed a moderate improvement over using absolute elevation for predicting habitat presence. At the 5 x 5 (450 m square) neighborhood size, the largest increase in the AUC occurred, suggesting a strong rise in correspondence with the habitat occurrence as the neighborhood width grows linearly near this neighborhood size. In the vicinity of the maximum correspondence at the 19 x 19 m neighborhood selected for the rest of the analysis, the differences were not large among neighborhood sizes. A range of neighborhood sizes from 1350 m square to 2070 m square (or a similar size neighborhood of different shape) will likely be similarly predictive of aquatic habitat presence in a landscape similar to the one in Asembo (lowlands).

At this neighborhood size, the TPI spatial pattern seems to correspond well with the structure of the stream pattern at the site, in contrast with the smallest neighborhoods which barely hint at any watershed structure. While some of the other topographic variables used in aquatic habitat models may benefit from similar scale testing, it is easy to interpret the meaning of the neighborhood size as the “localness” of a “local lowness” interpretation of TPI in terms of how this variable may relate to the accumulation of standing water, while variables such as slope and aspect offer a less intuitive interpretation.

A limitation of this study is that I did not consider the use of more than one topographic variable here. The neighborhood size selected shows relatively coarse scale effect of the topography on the probability that a location hosts a habitat. One possibility that still uses only the TPI variable is that if the larger scale were controlled for in the multivariate model first, then the very local neighborhood could explain some of the remaining variation by detecting depressions in areas other than streams. Other authors have shown the use of multiple topographic variables in models—Nmor et al. (2013) tested for correlation between topographic variables (elevation, slope, aspect, curvature, convergence index, TWI, and TPI) in their model to avoid collinearity effects. They found that only TPI (500 m neighborhood) and convergence index were highly correlated at that site with a Pearson's correlation coefficient > 0.8; also, that of the two, TPI has a lower AIC in a model of anopheline larval habitats.

Given the reliably strong performance of topographic variables in models for all habitat definitions, more work should be done to determine the independence of topographic variables at a variety of sites and to build out the distinctions between topographic variables that promote aquatic habitat presence and those that are associated with conditional larval habitat presence. The logical next step with the current model according to this need is to try TWI, distance to stream (McCann et al. 2014) and the best TPI scale seen here (1710m) in a model together to clarify whether each variable is independent. Figure 8 suggests that distance to stream and the TPI at this scale may be highly correlated. So far, the two variables that have been demonstrated to be significantly associated with both (and thus, also anopheline larval habitats) are slope and TPI (Clennon et al. 2010). Their work also suggests that TWI works to increase the number of anopheline larval habitats via its effect on aquatic habitat presence only. These results are backed up by the related literature with consistent results, including the conclusions about TPI in this study.

These results and others suggest that other variables that are more easily altered or are critical to understanding habitat distributions should always be given more weight when they are also associated with local lowness and other topographically favorable features, or understood well with regard to topography. For instance, more models should test interaction effects with favored topographic variables. This study tested an interaction effect with TPI and LULC which did not show any statistical significance when all classes were included in the interaction. Individual classes were not tested.

The SRTM elevation values appear to be overestimating topographic features in the vicinity of the streams. The streams are often surrounded by shrubs due to the steeper slopes and less navigable terrain along their edges and the errors seen are consistent with the behavior of the radar wavelength (C-band; 5.6 cm) and return behavior in the canopy for SRTM data . These areas are truly low and the freely-available SRTM data do not measure them well. Since the topographic indices such as TPI are such reliable participants in correlative aquatic habitat models, it is important to understand how sensitive the habitat predictions are to error in the TPI due to error in the DEM.

The earlier mapping of a high density of GPS observations at this site afforded a serendipitous opportunity to test the error propagation with regards to aquatic habitat prediction also performed at the site. Three-quarters of the positive predictions are stable (95% confidence at the cutoff that is associated with 90% sensitivity) under the conditions of error in measuring the elevation. These regions are either so low that they can have any amount of error and will still be predicted as habitat presence reliably, or they are moderately low areas with small errors. The remaining quarter of the positive predictions are sensitive to typical error in the DEM used as an input to the model. An error propagation assessment with respect to the relationship between TPI and conditional larval presence would be a valuable next step to undertake.

The residuals of the model are larger where the TPI is lowest. Transformations of the TPI variable did not have much effect on the heteroscedasticity of the residuals and so I concluded that the errors are not independent due to the spatial autocorrelation of the TPI and the habitats, rather they are larger because the relationship of this variable was non-linear with the logit model of the habitat presence.

4.4 Rainfall

The TRMM previous 1-week rainfall is modestly positively associated with increased habitat presence. This variable is not necessarily accounting for long-term trends in wetness, predicting high or low habitat years due to wet or dry years seen in the interannual rainfall variation. Rather, it was hoped this variable could explain why within the 4-month sampling window, the site shows a different realization of the habitat pattern approximately every 2 weeks, with a mean agreement of only 54.4% between patterns after only 2 weeks has passed (Figure 15). As described in McCann et al. (2014), the training data were not collected to maximize the variation of this temporally dynamic variable, which most likely results in bias in the estimator. The negative correlation of the TRMM-Asembo rainfall values and the GSOD rainfall values during the study period seems to be an aberration given the long-term patterns. The TRMM-Asembo and TRMM-Kisumu comparisons show that the source of the disagreement is weighted more toward the TRMM error than toward the true difference in rainfall between the study site and the airport 40 km away. Given the sparsity of weather stations in this region, I expect the TRMM errors to be relatively high due to the dependency on weather station data for scaling adjustments. Though the study design tries to heed the suggestion by Huffman et al. (2007) to average the 3B42 data into longer time intervals to reduce the relative error, the 1-week interval selected for the model is not very long and probably still contains fairly large errors. It may be that the relationship seen here is spurious. This study does little to reverse a trend of mixed success predicting not only aquatic habitats but other aspects of the malaria transmission cycle at the local scale using TRMM. There is a clear seasonal component to the aquatic habitat occurrence that must be related to water availability somehow, with the number of habitats approximately doubling in the peak season compared to the May and August habitat observations. In order to double the habitat probability, 29.3 mm of additional weekly rainfall are needed according to this model. This is within the realm of possibility, but further modeling attempts should consider incorporating soil moisture models in addition to seeking out newer high-quality satellite precipitation estimates.

4.5 Prediction comparison

The environmental models perform better than the use of prior knowledge of the habitat locations. The most predictable visit, visit 9, was used to illustrate the performance comparison in Table 3. The logistic regression trained on 2011 data performs better in terms of prediction tradeoffs between sensitivity and specificity than both the direct habitat predictions and the indicator kriging hotspot predictions—with the sensitivity held the same, the specificity approximately triples when using the environmental model. The indicator kriging hotspot models do add detection power to the prediction, increasing the proportion of visit 9's habitats that are predicted by 23.1 percentage points while decreasing the sensitivity by 14.3 percentage points, which is a fruitful tradeoff if detection of as many habitats as possible is a priority. If the areas without any positively observed habitats near a habitat at one time point are unrelated to the probability of habitats at another time point, I would not expect to see this large increase in detection power.

In this study, the 2007 habitats were surveyed extensively but not exhaustively—these data suggest that even though no aquatic habitat event was observed at some point, there is an underlying predilection towards being a habitat that can be ascribed to the location based on its proximity to a confirmed habitat. The next step with this analysis would be to use regression kriging to see if the best model's predictions improve further and the residuals show no further spatial autocorrelation. Kriging is an interpolation method that is not useful when there are no existing data about habitat presence. If resolution improvements of remotely-sensed imagery in the future result in the ability to detect even a fraction of habitats (say, larger ones), then these data could be enough to seed a prediction that combines the general power of the environmental model with the site and season specific nature of the spatial autocorrelation.

The late season habitats are the most easily predicted by habitat data from other time points as well as by the environmental model. The stability of the habitats increases as the rainy season dwindles into the dry season. The earliest visit is especially poor at predicting habitats in any other time point, but a large improvement is seen with the predictive power of visit 2. The earliest visit is also the most poorly predicted by the environmental model by a large margin. The visit just at the end of the rainy season is the most predictive of other visits, although it also generates the greatest number of false positives. These results suggest that data from the late rainy season, when habitats peak, is useful for forecasting habitats for the upcoming months. There also is some critical turning point, seen here between the first and second visit, when the forecasting potential of the existing habitats greatly increases. Since it is desirable to be able to predict the habitats at their peak as well as in the late season, a next step is to identify whether this turning point has any general characteristics or whether it is a feature of these specific data.

These data contain a warning against using the accuracy metric as an evaluation of the success of the model when the modeled event is somewhat rare. Here, aquatic habitats comprise only 6.44% of the area in the 2011 survey grids, and even less in 2007. The accuracy is maximized by simply predicting no habitats everywhere (93.56% accuracy). Only an extraordinarily successful model will have higher accuracy at some other cutoff. I have used the sensitivity here and considered the models in terms of their detection power at varying cutoffs, but depending on the application it is equally valid to consider their specificity instead.

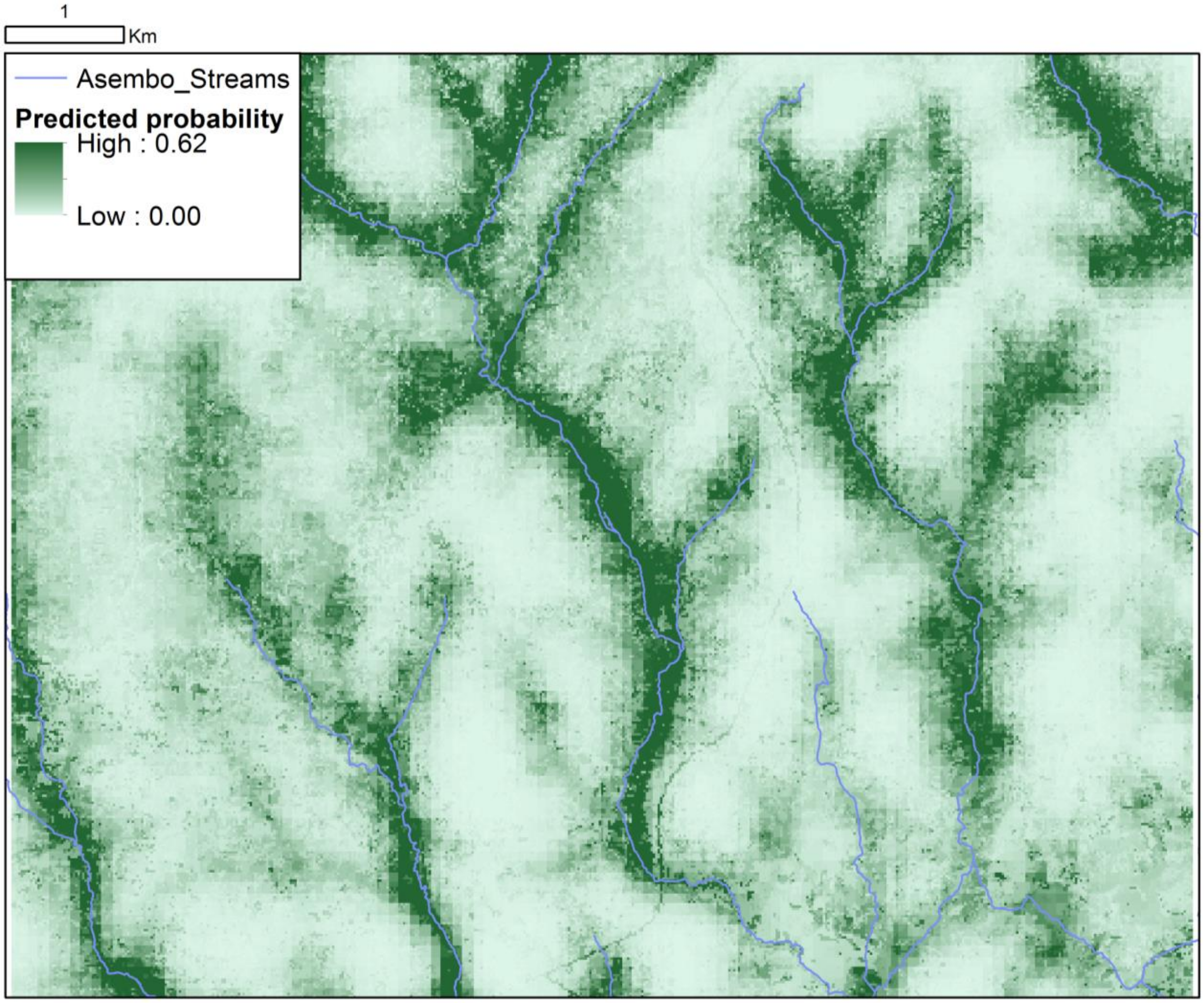

Figure 18 Predicted probability of habitat from the best model for Jul 2, 2007.

4.6 Spatiotemporal autocorrelation

There is a high degree of spatial and temporal autocorrelation in the repeated visits data. The crossover group of time lags 1 – 3 are a little more highly correlated when the spatial lag is zero, suggesting these lags are very effective at predicting each other directly. They are a little more temporally autocorrelated than the lags 4 – 7 but not any more spatially correlated. Time lag 0 is in the middle of the pack with respect to spatial autocorrelation. In general, habitats are at least as spatially autocorrelated with other habitats in other time points as they are with habitats in their own time point.

The variogram for time lag 8 shows a similar degree of cross-correlation between visits to the other time lags when the spatial lag is zero, suggesting that the habitats observed in visits 1 and 9 show as much aspatial correlation as any other pair combination. However, there is an immediate separation from the pack and habitat presence is more highly cross-correlated at all spatial lags. This is the time lag with the least combinations—only a single visit pair, visit 1 and visit 9 at opposite ends of the study period have this lag. While there are fewer pairs than at other time lags, there are still 11,250 paired observations and this time lag fits into a general pattern of higher cross-correlation at longer time lags. The variogram shows that when the habitat counts dwindle on either end of the rainy season, they settle into locations showing a high degree of spatial similarity.

The linear relationship seen in the sensitivity for each relative lag in Figure 15 must follow a large drop-off from the theoretical time lag 0. However, though the theoretical value there is 100%, I would not expect to see 100% agreement in reality if two observers visited the same habitats and recorded their locations as points with GPS independently. Both the GPS error and the random arrangement of the observer's position on the edge of the habitat are sufficient to lead to the habitat being observed in a neighboring grid cell rather than the same cell. I would expect steep drop-off regardless were these shorter time intervals to be measured, but determining the amount of disagreement due to these errors, perhaps via simulating them, would be necessary to understanding the adjustment that should be made to the values along the rest of the curve.

CHAPTER 5

CONCLUSION

The pattern of stability of aquatic habitats at the local scale over the course of a full seasonal change contains potentially valuable information for the forecasting of future aquatic habitat presence, but is not sufficiently powerful to predict future habitat distributions in its own right better than a moderately well-performing environmental model. Operational maps of a site must be updated frequently and supplemented by the predictions of an environmental model. Here, I have demonstrated the results of a generalizable, environmental, logistic regression model using TPI, soil data, an LULC classification, and satellite precipitation estimates.

TPI was a major contributor to the prediction power of the model, confirming results that low TPI is predictive of anopheline larval habitat presence because it is predictive of the presence of the aquatic habitats themselves. This variable is easily calculated from freely-available data and the predictions are moderately robust to errors inherent to satellite radar-based digital elevation models. Two possible investments are suggested by the importance of this variable and the current status of the available data. First, the cost of higher-resolution topographic data may be justified by the importance of the topographic variables. Such data could include, for example, airborne LIDAR data processed to extract the bare earth surface. In Kenya and other countries with confined regions of malaria transmission, the area that must be flown is relatively small and could be reduced further, if necessary, by using the freely-available topographic data in a "screening" program to identify local sites with both high malaria incidence and topography contributing to aquatic and/or anopheline larval habitats. The new data could be used to develop efficient larval control programs and larval habitat surveillance systems.

The new data could also be used as an input into a more substantial investment. Malaria is a serious problem degrading not only the physical health of those communities at risk of the disease, but also their economic health and growth. Well-planned landscape alteration in these communities has the potential to strategically eliminate larval habitats, so that further health and economic growth are returned by the initial investment. If the terrain could be altered strategically in these areas to ensure that water flows all the way into the streams or flows to designated reservoirs for human use that could be treated for larvae more efficiently, then adult mosquito production out of these regions could be reduced.

Many topographic variables have been associated with anopheline larval habitat presence, but with the exception of slope and TPI, previous study designs have been unable to describe whether the correlation is driven by an association with the number of aquatic habitats or the suitability of the habitats for anopheline larvae. A similar situation exists for other variables appearing consistently in landscape models of habitat presence, such as LULC. This distinction is important for connecting the results of studies focusing on landscape predictors of habitats with ecological studies of anopheline habitat characteristics, as well as for making larval vector control decisions.

Satellite precipitation estimates show an association strong enough to potentially account for seasonal swings in aquatic habitat presence, but doubt was cast on the reliability of the estimates when compared with local weather station data. Further testing of satellite rainfall estimates should consider incorporating soil moisture models to better capture the dynamic moisture dimension (saturation) that is likely associated with aquatic habitat presence. These data, though troublesome, are attractive because they present an opportunity to update a general model to any arbitrary date of interest. An error propagation analysis could be valuable to determine further whether satellite precipitation estimates as they currently exist are hopeless for this application, or useful in spite of their shortcomings.

Repeated observations in a spatially-stratified sampling scheme provided an opportunity to examine the spatiotemporal correlation of aquatic habitats. The result shows a strong degree of cross-correlation in both dimensions, indicating that any prior knowledge of an aquatic habitat can be used to elevate not only the estimate of future probability of another habitat at the same location, but also at nearby locations. The strength of association declines linearly with time over the 4-month time period tested, but results from the literature suggest that the decline may level off soon after the last observation here to persist at a low level across both years and seasons. In anticipation of the possibility that spaceborne sensor spatial resolutions will eventually yield systematic, direct observations of smaller aquatic habitats, future work could include collecting airborne imagery of habitats at fine resolutions and validating spatiotemporal kriging interpolated maps of aquatic and/or anopheline larval habitats at time intervals extending from several days out to at least a year.

The habitat stability analysis shown here demonstrated that after the peak habitat incidence rate is reached at some point after the long rains, the habitats dwindle into locations that are increasingly stable. The spatiotemporal variogram shows that these locations have a spatial structure similar to what is seen in the earliest weeks before the peak. The habitat stability map shows that habitats that are stable throughout the entire 4-month period were common and clustered. One larval control strategy should be to focus on identifying the habitats in late August and investigate habitat modifications or larviciding for these stable sites specifically, in order to target the mosquito populations when they are the most spatially confined throughout the year.

APPENDIX

A.1 Regression kriging and simulation kriging for DEM script

The extensive comments included from this code are retained from a personal blog post written to demonstrate the code and method. Though informal, they are informative and are retained here mostly unchanged.

~~~
Regression kriging and terrain simulations in R
========================================================

For my thesis project, we wanted to see if we could use a large collection of
differentially-corrected GPS elevation records gathered during previous
demographic work in the region to improve upon the available digital
elevation models (DEMs). (Though others are always involved, "we", for the
bulk of this portion, consisted of my fellow student within the larger
project group, Rob McCann, and our professor for our excellent spatial
analysis class, Dr. Ashton Shortridge). The approach I’ve ended up with so
far is to use the GPS records as secondary data in a regression kriging
(http://spatial-analyst.net/wiki/index.php?title=Regression-kriging_guide) to
attempt to improve upon the SRTM DEM estimates. I’ve also modeled the error
surface, showing that height errors are spatially autocorrelated-you can see
readily how the SRTM surface suffers from vegetation inversions, averaging
about 6m, or the height of a typical tall shrub canopy, in the vicinity of
the streams. I’m looking for mosquitoes and I’d prefer not to have my error
concentrated in low regions, since that’s exactly the place I need the data
the most.

This code is written up in R Markdown using knitr. I jumped in head first
wrestling with the knitr/LaTeX (well, Sweave, once upon a time) and while I'm
doing more complicated knitr-ing for my thesis writing, I must say that it is
shockingly easy to make a decent enough write-up or other "notebook"-style
presentation using this combo.

First, setup using a little GDAL grease: open the GPS points shapefile that I
already overlaid with the SRTM heights at each point and the difference/error
between the points, and open the SRTM DEM. Both were already in the same and
correct projection.

```{r message = FALSE}
#load necessary libraries
library(rgdal)
#Change directory to what you need
setwd("C:/Users/Nicole/Dropbox/Model")

#Read GPS points shapefile
#This file has the gps, srtm, and error values already, error: positive if
GPS was lower than SRTM
file.path <- "kriging_elevpts_NEWJUNE_FINAL"
elev.pts <- readOGR(".", file.path)
~~~

~~~
#copy the coordinates columns into the dataframe to use later to explicitly
check for trend surface
elev.pts$easting <- coordinates(elev.pts)[,1]
elev.pts$northing <- coordinates(elev.pts)[,2]

#Open up the SRTM data
srtm<-readGDAL("SRTM_5000m_from_2011_extent.tif") #comes in as
SpatialGridDataFrame
srtm <- as(srtm, "SpatialPixelsDataFrame") #but needs to be this instead for
kriging

#There's a lake by my site that I don't want to bother with and this is just
a quick and dirty way to get rid of it
lake <- 1134
#srtm <- srtm[srtm$band1 > lake,]
```

Quick glance at the datasets using the sp libary's plotting function--sp came
in when we loaded rgdal.
```{r fig.width = 7, fig.height = 6}
spplot(elev.pts, "srtm", main = "GPS Points Dataset Preview")
spplot(srtm, "band1", main = "SRTM Dataset Preview")
```

Looks fine. Before doing anything fancy, we'll just take a look at the
relationship between the GPS and SRTM data.


```{r fig.width=7, fig.height=6}
lmod <- lm(gps ~ srtm + northing , as.data.frame(elev.pts))
summary(lmod)
cor.gps.srtm <- cor.test(elev.pts$gps, elev.pts$srtm)
cor.gps.srtm

#save the residuals to the dataset
elev.pts$residuals <- residuals(lmod)
writeOGR(elev.pts, dsn = "C:/Users/Nicole/Dropbox/Model", layer = "elev.pts
with residuals new", driver = 'ESRI Shapefile')


#Plot the linear relationship, expecting high correlation
plot(gps ~ srtm, as.data.frame(elev.pts), main = "Elevation datasets are
highly correlated")
abline(lm(gps ~ srtm, as.data.frame(elev.pts)))
```

Unsurprisingly, there is a very high correlation. The reason we move on to
kriging, however, is that the residuals are highly spatially autocorrelated.
Time to bring in gstat.
~~~

~~~
```{r fig.width = 7, fig.height = 6, message = FALSE}
#Fit the variogram model of the residuals:
library(gstat)
elev.pts.vplot <- variogram(gps ~ srtm, elev.pts, cutoff=1100, width=25)

plot(elev.pts.vplot, main = "Variogram of residuals")
null.vgm <- vgm("Exp", psill = 18.0, nugget = 11.3, range=109)
vgm <- fit.variogram(elev.pts.vplot, null.vgm) #you can use this to help with
fitting
plot(elev.pts.vplot, vgm, main = "Variogram model")
vgm
```
~~~

As an aside, I first did all this ages ago when the gstat object was a mystery to me, and it worked all the same, but I think I actually get the whole thing now. Anyway, it's just a way to package up some things we'll use for actual kriging, you can bring it all together at that time as well if you prefer.

~~~
```{r}
#next part is copied faithfully from GEO 866 Spatial Analysis class with Dr.
Ashton Shortridge
g <- gstat(id="elev.g", formula = gps ~ srtm, data=elev.pts, model = vgm)
#create a gstat object with our data, formula, and variogram model
blue0 <- predict(g, newdata = elev.pts, BLUE=TRUE, debug.level=3)   # The GLS
trend estimates are returned
blue0$blue.res <- elev.pts$gps - blue0$elev.g.pred    # Calculate residuals
blue0$srtm <- elev.pts$srtm
```
~~~

After we get the BLUE residuals, we're supposed to check that the variogram model doesn't change. I've never seen it change even a little. Perhaps I'm doing something wrong. I've read somewhere that it usually doesn't have any impact and this step could be skipped, but it only takes a minute to check.

~~~
```{r fig.width = 7, fig.height = 6}
elev.g.vplot <- variogram(blue.res ~ srtm, blue0)
plot(elev.pts.vplot)
new.vgm <- vgm #try the old model first, change if needed
plot(elev.pts.vplot, new.vgm)
#NO CHANGES!
```
~~~

Now I just do a little bit of prep on the SRTM grid to make it work in the kriging. I get rid of any NA values, which will throw up an error, and though the SRTM surface and the GPS points do have the same projection, the proj4 strings are apparently saved a shade differently for vectors and rasters in ArcGIS, so I just copy one to the other and ignore the error that comes up.

~~~
```{r}
srtmgrid <- srtm

#have to remove NAs for the gstat stuff to work okay.
srtmgrid <- srtmgrid[!is.na(srtmgrid$band1),]

#cleanup
names(srtmgrid) <- "srtm"
proj4string(srtmgrid) <- proj4string(elev.pts) #ignore error, assuming you
know what you're doing
```


I'd like the estimates to tend toward the local means, so I set a few
parameters on the search distance and away we go with a regression kriging.

```{r}
num.max <- 60
dist.max <- 2500

rk <- krige(gps ~ srtm, elev.pts, srtmgrid, model=new.vgm, maxdist =
dist.max, nmax = num.max)

rk$stderr <- sqrt(rk$var1.var)

spplot(rk, "var1.pred", col.regions = terrain.colors(20), main = "Kriging
estimates")
spplot(rk, "stderr", col.regions=heat.colors(20), main = "Kriging error")
```

You are most likely going to want to save your results after you wait through
the kriging analysis, and you'll get a lot more capabilities for further
visualization and analysis out of a GIS, anyway.
```{r results='markup'}
out.name.interpolated <- "90m_FINAL_regkrige_nmax60_mdist1500_extended.img"
writeGDAL(rk, out.name.interpolated, driver="HFA")
```

My ultimate application for the terrain surfaces I'm making is to use them to
find "locally low" sites. I'll be using a simple topographic index to measure
a "textural" quality of the terrain, and therefore simulation kriging is
appropriate if I want to predict not only the best estimates at location, but
also come up with some idea what the error means for each location's position
with respect to its neighbors. Simulation kriging just requires adding a few
arguments to the krige function, the number of simulations, and a boolean
indicating I am not performing indicator kriging.

```{r fig.width =7, fig.height = 6}
~~~

```
rk.sim <- krige(gps ~ srtm, elev.pts, srtmgrid, model=vgm, maxdist =
dist.max, nmax = num.max, nsim =100, indicators = FALSE, debug.level = -1)

#plot just the first few simulated surfaces to get an idea how they all look
spplot(rk.sim, c("sim1", "sim2", "sim3", "sim4"), col.regions =
terrain.colors(20), main = "A few of the 100 simulations of the terrain in
Asembo")


out.name.simulated <- "90m_FINAL_SIMSkrige_d1500_nmax60_extended.img"
writeGDAL(rk.sim, out.name.simulated, driver="HFA")
```

And, finally, I won't go into any of the analytical details here, but there are a few diagnostics that you can look at after performing a cross validation of the regression kriging.

```{r results='hide'}
#cross validation kriging
cv.rk<- krige.cv(gps ~ srtm, elev.pts, srtmgrid[srtmgrid$srtm > lake,],
model=vgm, maxdist = 1500, nmax = 60)
```

```{r fig.width = 7, fig.height = 6}

#bubble plot of residual size
coordinates(cv.rk) <- c("coords.x1","coords.x2")
bubble(cv.rk, z="residual")

#some basic error stats
sd(cv.rk$residual)
rmse.cv.rk <- sqrt(sum((cv.rk$observed-
cv.rk$var1.pred)^2)/length(cv.rk$observed))

#can compare the rmse and original sd to demonstrate how much of the
variability the kriging "explained"
orig.sd <- sd(elev.pts$gps)
orig.sd

#check that variogram does not show significant autocorrelations, mine
usually shows a bit
cv.vg<-variogram(residual~1,cv.rk, width=12, cutoff=300)
plot(cv.vg)

## Compare errors to std errors ##

#check out where large errors are
cv.rk$big <- factor(ifelse(cv.rk$zscore <= -2, 'negative',
ifelse(cv.rk$zscore < 2, 'moderate', 'positive')))
spplot(cv.rk, "big")
```

~~~
#over.2z is the percent of errors that exceed two SD from the mean. Would
expect this to be 5% in a normal distribution. If higher, model optimistic.
If lower, model conservative.
over.2z <- 100*length(cv.rk$zscore[(abs(cv.rk$zscore) >
2)])/length(cv.rk$zscore)
over.2z
```
~~~

And that’s it. I have some confidence in my theoretical grasp of kriging and of course I certainly fretted over getting the details right for my project including which exact methods to use, but I’m no spatial analysis ninja, either, and if you are actually reading this trying to follow along, you might want to try (http://spatial-analyst.net/wiki/index.php?title=Main_Page) or find a copy of Applied Spatial Data Analysis with R (https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&ved=0CEMQFjAC&url=http%3A%2F%2Flink.springer.com%2F978-0-387-78170-9&ei=v95gUpHqJNKyygH_14HwDQ&usg=AFQjCNECcQtarQil4SmSaejbNGc54m98EA&sig2=jgCL5miZDxikf4D1PyCOdw) (try your academic library's e-book access, if you have it) for more authoritative advice.

A.2 Topographic position index calculation

```
PRO locally_low_final

COMPILE_OPT IDL2

; This program is to open the desired DEM file and process any relevant
elevation
; band with the kernel size 3x3 to 66x66 (1km at 30m) and save it as a file
CD, "/home/nicole/Dropbox/Model"

; Launch ENVI if you haven't already
ENVI

PRINT, "Waiting 10 seconds for ENVI to open."
WAIT, 5

filename = ENVI_PICKFILE(TITLE='Pick a DEM file', FILTER='*.img',
DEFAULT='/home/nicole/Dropbox/Model/')
ENVI_OPEN_FILE, filename, r_fid=fid
IF (fid EQ -1) THEN RETURN

; Necessary to get some variables
ENVI_FILE_QUERY, fid, dims=dims, nb=nb, data_ignore_value=d_ignore
inherited_details = ENVI_SET_INHERITANCE(fid, dims, /FILE_TYPE, /GEO_POINTS,
$
   /MAP_INFO, /PIXEL_SIZE, /SPATIAL)

; Find out if this is for the simulations, can avoid some looping that way

; This is all to process the right number of bands. Can only do quantity
```

```
; Cannot specify particular bands this way.
;PRINT, 'Number of bands:', nb
;READ, 'Type the number of bands to read (min:1, max:254)', input
;PRINT, input
;input = UINT(input)
;HELP, input
;input_test = input LT 254
;HELP, input_test
;IF input_test EQ 0 THEN BEGIN
;  PRINT, 'Unexpected input. Program will proceed filtering all bands of
file.'
;  selected_bands=pos
;ENDIF ELSE BEGIN
;  selected_bands = LINDGEN(input)
;  HELP, selected_bands
;ENDELSE

;;ENVI VERSION BUT ENVI CONVOLUTIONS ARE NOT WHAT I WANT
;; This makes a separate file for each kernal size
;CD, '../../tempnicole'
;lli_size = [3,5,7,9,11,13,15,17,19,21,23,25,27,29,31,33]
; FOREACH element, lli_size DO BEGIN
;  ENVI_DOIT, 'CONV_DOIT', DIMS=dims, FID=fid, POS=selected_bands,
/IN_MEMORY, KX=element, $
;  KY=element, OUT_NAME=STRING(element)+'LLIndex_'+filename, R_FID=con_fid
;  PRINT, "Filter size ", lli_size, " complete."
;  ENVI_FILE_MNG, con_fid, /REMOVE
; ENDFOREACH

;;IDL VERSION HAS MORE INTEGRITY AND TRANSPARENCY
data = ENVI_GET_DATA(fid=fid, dims=dims, pos=0)


; Change the Nan values in the lake region to a lake elevation
;IF input LT 3 THEN lake=1129 ELSE lake=1112
;1129 is the lowest of the 30mkriged, 1111 the lowest of the 30m simulation
nan_index = WHERE(finite(data) EQ 0, count)
IF count GT 0 THEN data[nan_index]=1125 ELSE data=data
PRINT, 'Waiting for kernel processing....'

;; This makes a separate file for each kernal size
READ, "What filename should be appended?", filename
lli_size = [3,5,7,9,11,13,15,17,19,21,23,25,27,29,31,33]
HELP, lli_size
 FOREACH element, lli_size DO BEGIN

  ;First make the kernel
  kern = INTARR(element, element)-1
  kern[(element/2), (element/2)]=(element^2)-1
  scale_fac = FLOAT(element)^(-2)

  filtered = FLTARR(dims[2]+1,dims[4]+1) ; initialize array to fill

  ;Then apply it to the image
  filtered=CONVOL(data, kern, scale_fac, /EDGE_MIRROR)
 ; and save to a new file
  out_element = STRCOMPRESS(STRING(element), /REMOVE_ALL)
  PRINT, out_element
  HELP, out_element
  out_name=out_element+'kLLI_'+ filename
  PRINT, out_name

  ENVI_WRITE_ENVI_FILE, filtered, DESCRIP="Results from the locally low 
filter kernel size " + STRING(element), $
    INHERIT=inherited_details, OUT_NAME=out_name,  /NO_OPEN

  ;remove the file before moving on
  PRINT, "Filter size ", element, " complete for ", OUT_NAME
;  ENVI_FILE_MNG, r_fid, /REMOVE


 ENDFOREACH
END
```

A.3 Rainfall time series extraction

```
#stack the trmm into

import datetime, os
import arcpy
import numpy as np

def get_3B42daily(workspace, extent_fc, output_csv):
    """For a folder containing 3B42-daily NetCDFS, makes a CSV table with two
    columns, date, and preciptation total (mm). Averages
    the values of the TRMM cells within the extent region to produce a single
    value per date.
    workspace: folder with the 3B42-daily NetCDFs inside, not nested. ALL of
                them will be included

    extent_fc: study site extent or region or interest polygon. 3B42 rasters
                will be clipped to this extent.

    output_csv: the output csv table path"""

    arcpy.env.workspace = workspace
    ncs = arcpy.ListFiles('*.nc')
    records = [','.join(['Date (MM-DD-YYYY)', 'Total Precipitation (mm)'])]

    for nc in ncs:
        print nc
        year, month, day = nc.split('.')[1:4]
        date_string = '-'.join([month, day, year])
        arcpy.MakeNetCDFRasterLayer_md(nc, 'r', 'longitude', 'latitude', 
'lyr')
        arcpy.Clip_management('lyr', '#', 'in_memory/pcp_clip', extent_fc)

        precip = arcpy.RasterToNumPyArray('in_memory/pcp_clip')
        records.append(','.join([date_string, str(np.mean(precip))]))

        for item in['lyr', 'in_memory/pcp_clip']:
            arcpy.Delete_management(item)

    with open(output_csv, 'a') as f:
        for line in records:
            f.write(line + '\n')

def run_2007():
    ws_2007 =
'C:/Users/Nicole/Dropbox/Feb_2014_WORKFILES/subset_wizard_2006_2007/daily_200
6_2007'
    asembo_extent =
'C:/Users/Nicole/Dropbox/Thesis.gdb/Study_Grid_2011_5KM_buffer'
    output_2007 =
'C:/Users/Nicole/Dropbox/Feb_2014_WORKFILES/2006_2007_rainfall_FIX.csv'
    get_3B42daily(ws_2007, asembo_extent, output_2007)

def run_2011():
    ws_2011 =
'C:/Users/Nicole/Dropbox/Feb_2014_WORKFILES/subset_wizard_2010_2011/3B42_dail
y'
    asembo_extent = 'C:/Malaria_Data/Thesis.gdb/Study_Grid_2011_5KM_buffer'
    output_2011 =
'C:/Users/Nicole/Dropbox/Feb_2014_WORKFILES/2010_2011_rainfall_FIX.csv'
    get_3B42daily(ws_2011, asembo_extent, output_2011)

run_2007()
```

A.4 Spatiotemporal variogram

```
#see this url for more information and help https://cran.r-
project.org/web/packages/gstat/vignettes/st.pdf

library(sp)
library(spacetime)
library(gstat)

#data frame with all 1-9 observations for pixid 1 then all for pixid 2 etc

d7 <- read.csv("C:/Users/Nicole/Dropbox/njs-
thesis/FINAL_2007_USEALL_VER2_includesNA.csv")
d7$Visit_Date <- as.Date(d7$Visit_Date, "%Y-%m-%d")
d7$lulc <- factor(d7$lulc, levels =
c(29,0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,
                                      22,23,24,25,26,27,28,30,31))
d7$Visit_Num <- ordered(d7$Visit_Num)
d7$soil_clay <- factor(d7$soil_clay, levels = c("KA", "IN", "MO"))
```

```
d7$presence.num <- as.numeric(d7$presence)

d7 <- d7[order(d7$Visit_Num, d7$pixid),]

d7ST <- subset(d7, select = presence.num)

#need a spatial points object of locations (each pixel as a point?)
d7.pts <- d7[d7$Visit_Num == 3,]
d7.pts.only <- subset(d7.pts, select = c(easting, northing))
row.names(d7.pts.only) <- d7.pts$pixid
coordinates(d7.pts.only) <- ~ easting + northing

#list of timepoints (visits 1-9)
d7.nona <- d7[!is.na(d7$Visit_Date),]
dates <- as.vector(by(d7.nona$Visit_Date, d7.nona$Visit_Num, min))
dates <- as.Date(dates, origin = "1970-01-01")

visits <- dates

habST <- STFDF(d7.pts.only, visits, d7ST)

vv <- variogram(presence.num ~ 1, habST, width = 40, tlags=0:8, cutoff = 500)
plot(vv)
plot(vv, map = FALSE, col = heat.colors(8))

library(ggplot2)

ggplot(data=vv, aes(x=dist, y=gamma, group=t, colour=t)) +
  geom_line(size = 1) +
  geom_point() + scale_colour_brewer(palette="Paired") + xlab("Spatial lag
(m)") + ylab("Gamma")

stplot(habST)

library(lattice)


wireplot <-plot(vv, all=T, wireframe=T, zlim=c(0,.03),
zlab=NULL,
xlab=list("distance (km)", rot=30),
ylab=list("time lag (days)", rot=-35),
scales=list(arrows=F, z = list(distance = 8)))
```

LITERATURE CITED

LITERATURE CITED

Adimi, F., R. P. Soebiyanto, N. Safi, and R. Kiang. 2010. "Towards malaria risk prediction in Afghanistan using remote sensing." Malaria Journal 9. doi: 12510.1186/1475-2875-9-125.

Barnes, C., & Cibula, W. 1979. “Some implications of remote sensing technology in insect control programs including mosquitoes”. Mosquito News 39:271-282.

Bayoh, M. N., and S. W. Lindsay. 2003. "Effect of temperature on the development of the aquatic stages of Anopheles gambiae sensu stricto (Diptera : Culicidae)." Bulletin of Entomological Research 93 (5):375-381. doi: 10.1079/ber2003259.

———. 2004. "Temperature-related duration of aquatic stages of the Afrotropical malaria vector mosquito Anopheles gambiae in the laboratory." Medical and Veterinary Entomology 18 (2):174-179. doi: 10.1111/j.0269-283X.2004.00495.x.

Beck, L. R., M. H. Rodriguez, S. W. Dister, A. D. Rodriguez, E. Rejmankova, A. Ulloa, R. A. Meza, D. R. Roberts, J. F. Paris, M. A. Spanner, R. K. Washino, C. Hacker, and L. J. Legters. 1994. " Remote-sensing as a landscape epidemiologic tool to identify villages at high-risk for malaria transmission." American Journal of Tropical Medicine and Hygiene 51 (3):271-280.

Beier, J. C., R. Copeland, C. Oyaro, A. Masinya, W. O. Odago, S. Oduor, D. K. Koech, and C. R. Roberts. 1990. " Anopheles-gambiae complex egg-stage survival in dry soil from larval development sites in estern Kenya." Journal of the American Mosquito Control Association 6 (1):105-109.

Budiansky, S. 2002. "Creatures of our own making." Science 298 (5591):80-86.

Clennon, J. A., A. Kamanga, M. Musapa, C. Shiff, and G. E. Glass. 2010. "Identifying malaria vector breeding habitats with remote sensing data and terrain-based landscape indices in Zambia." International Journal of Health Geographics 9. doi: 5810.1186/1476-072x-9-58.

Curran, P. J., P. M. Atkinson, G. M. Foody, and E. J. Milton. 2000. "Linking remote sensing, land cover and disease." In Advances in Parasitology, Vol 47, 37-80.

Dambach, P., V. Machault, J. P. Lacaux, C. Vignolles, A. Sie, and R. Sauerborn. 2012. "Utilization of combined remote sensing techniques to detect environmental variables influencing malaria vector densities in rural West Africa." International Journal of Health Geographics 11. doi: 10.1186/1476-072x-11-8.

Debien, A., S. Neerinckx, D. Kimaro, and H. Gulinck. 2010. "Influence of satellite-derived rainfall patterns on plague occurrence in northeast Tanzania." International Journal of Health Geographics 9. doi: 6010.1186/1476-072x-9-60.

Farr, T. G., P. A. Rosen, E. Caro, R. Crippen, R. Duren, S. Hensley, M. Kobrick, M. Paller, E. Rodriguez, L. Roth, D. Seal, S. Shaffer, J. Shimada, J. Umland, M. Werner, M. Oskin, D. Burbank, and D. Alsdorf. 2007. "The shuttle radar topography mission." Reviews of Geophysics 45 (2). doi: 10.1029/2005rg000183.

Fillinger, Ulrike, and Steven W. Lindsay. 2011. "Larval source management for malaria control in Africa: myths and reality." Malaria Journal 10. doi: 10.1186/1475-2875-10-353.

Gimnig, J.E., M. Ombok, L. Kamau, and W.A. Hawley. 2001. "Characteristics of larval anopheline (Diptera: Culicidae) habitats in Western Kenya." Journal of Medical Entomology 38 (2):282-288.

Hamel, M. J., K. Adazu, D. Obor, M. Sewe, J. Vulule, J. M. Williamson, L. Slutsker, D. R. Feikin, and K. F. Laserson. 2011. "A Reversal in Reductions of Child Mortality in Western Kenya, 2003-2009." American Journal of Tropical Medicine and Hygiene 85 (4):597-605. doi: 10.4269/ajtmh.2011.10-0678.

Hay, S. I., R. W. Snow, and D. J. Rogers. 1998. "From predicting mosquito habitat to malaria seasons using remotely sensed data: Practice, problems and perspectives." Parasitology Today 14 (8):306-313.

Hayes, R. O., E. L. Maxwell, C. J. Mitchell, and T. L. Woodzick. 1985. " Detection, identification, and classification of mosquito larval habitats using remote-sensing scanners in earth-orbiting satellites." Bulletin of the World Health Organization 63 (2):361-374.

Huffman, G. J., R. F. Adler, D. T. Bolvin, G. J. Gu, E. J. Nelkin, K. P. Bowman, Y. Hong, E. F. Stocker, and D. B. Wolff. 2007. "The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales." Journal of Hydrometeorology 8 (1):38-55. doi: 10.1175/jhm560.1.

Kiang, R., F. Adimi, V. Solka, J. Nigro, P. Singhasivanon, J. Sirichaisinthop, S. Leemingsawat, C. Apiwathnasorn, and S. Looareesuwan. 2006. "Meteorological, environmental remote sensing and neural network analysis of the epidemiology of malaria transmission in Thailand." Geospatial Health 1 (1):71-84.

Li, Li, Ling Bian, Laith Yakob, Guofa Zhou, and Guiyun Yan. 2009. "Temporal and spatial stability of Anopheles gambiae larval habitat distribution in Western Kenya highlands." International Journal of Health Geographics 8. doi: 10.1186/1476-072x-8-70.

———. 2011. "Analysing the generality of spatially predictive mosquito habitat models." Acta Tropica 119 (1):30-37. doi: 10.1016/j.actatropica.2011.04.003.

Lindblade, K. A., T. P. Eisele, J. E. Gimnig, J. A. Alaii, F. Odhiambo, F. O. ter Kuile, W. A. Hawley, K. A. Wannemuehler, P. A. Phillips-Howard, D. H. Rosen, B. L. Nahlen, D. J. Terlouw, K. Adazu, J. M. Vulule, and L. Slutsker. 2004. "Sustainability of reductions in malaria transmission and infant mortality in Western Kenya with use of insecticide-treated bednets - 4 to 6 years of follow-up." Jama-Journal of the American Medical Association 291 (21):2571-2580.

McCann, R. S., J. P. Messina, D. W. MacFarlane, M. N. Bayoh, J. M. Vulule, J. E. Gimnig, and E. D. Walker. 2014. "Modeling larval malaria vector habitat locations using landscape features and cumulative precipitation measures." International Journal of Health Geographics 13:12. doi: 10.1186/1476-072x-13-17.

Messina, J. P., and S. J. Walsh. 2001. "2.5D Morphogenesis: modeling landuse and landcover dynamics in the Ecuadorian Amazon." Plant Ecology 156 (1):75-88. doi: 10.1023/a:1011901023485.

Minakawa, N., J. I. Githure, J. C. Beier, and G. Y. Yan. 2001. "Anopheline mosquito survival strategies during the dry period in western Kenya." Journal of Medical Entomology 38 (3):388-392. doi: 10.1603/0022-2585-38.3.388.

Minakawa, N., S. Munga, F. Atieli, E. Mushinzimana, G. F. Zhou, A. K. Githeko, and G. Y. Yan. 2005. "Spatial distribution of anopheline larval habitats in Western Kenyan highlands: Effects of land cover types and topography." American Journal of Tropical Medicine and Hygiene 73 (1):157-165.

Munga, S., N. Minakawa, G. F. Zhou, E. Mushinzimana, O. O. J. Barrack, A. K. Githeko, and G. Y. Yan. 2006. "Association between land cover and habitat productivity of malaria vectors in western Kenyan highlands." American Journal of Tropical Medicine and Hygiene 74 (1):69-75.

Munga, S., L. Yakob, E. Mushinzimana, G. F. Zhou, T. Ouna, N. Minakawa, A. Githeko, and G. Y. Yan. 2009. "Land Use and Land Cover Changes and Spatiotemporal Dynamics of Anopheline Larval Habitats during a Four-Year Period in a Highland Community of Africa." American Journal of Tropical Medicine and Hygiene 81 (6):1079-1084. doi: 10.4269/ajtmh.2009.09-0156.

Mushinzimana, E., S. Munga, N. Minakawa, L. Li, C. C. Feng, L. Bian, U. Kitron, C. Schmidt, L. Beck, G. F. Zhou, A. K. Githeko, and G. Y. Yan. 2006a. "Landscape determinants and remote sensing of anopheline mosquito larval habitats in the western Kenya highlands." Malaria Journal 5. doi: 1310.1186/1475-2875-5-13.

Mutuku, F. M., J. A. Alaii, M. N. Bayoh, J. E. Gimnig, J. M. Vulule, E. D. Walker, E. Kabiru, and W. A. Hawley. 2006. "Distribution, description, and local knowledge of larval habitats of Anopheles gambiae s.l. in a village in western Kenya." American Journal of Tropical Medicine and Hygiene 74 (1):44-53.

Mutuku, F. M., M. N. Bayoh, A. W. Hightower, J. M. Vulule, J. E. Gimnig, J. M. Mueke, F. A. Amimo, and E. Walker. 2009. "A supervised land cover classification of a western Kenya lowland endemic for human malaria: associations of land cover with larval Anopheles habitats." International Journal of Health Geographics 8:13. doi: 1910.1186/1476-072x-8-19.

Nmor, J. C., T. Sunahara, K. Goto, K. Futami, G. Sonye, P. Akweywa, G. Dida, and N. Minakawa. 2013. "Topographic models for predicting malaria vector breeding habitats: potential tools for vector control managers." Parasites & Vectors 6:13. doi: 10.1186/1756-3305-6-14.

Ombok, Maurice, Kubaje Adazu, Frank Odhiambo, Nabie Bayoh, Rose Kiriinya, Laurence Slutsker, Mary J. Hamel, John Williamson, Allen Hightower, Kayla F. Laserson, and Daniel R. Feikin. 2010. "Geospatial distribution and determinants of child mortality in rural western Kenya 2002-2005." Tropical Medicine & International Health 15 (4):423-433. doi: 10.1111/j.1365-3156.2010.02467.x.

Phillips-Howard, P. A., B. L. Nahlen, J. A. Alaii, F. O. ter Kuile, J. E. Gimnig, D. J. Terlouw, S. P. Kachur, A. W. Hightower, A. A. Lal, E. Schoute, A. J. Oloo, and W. A. Hawley. 2003. "The efficacy of permethrin-treated bed nets on child mortality and morbidity in western Kenya I. Development of infrastructure and description of study site." American Journal of Tropical Medicine and Hygiene 68 (4):3-9.

Rejmankova, E., J. Grieco, N. Achee, and D. R. Roberts. 2013. "Ecology of Larval Habitats." In Anopheles Mosquitoes - New Insights into Malaria Vectors, edited by S. Manguin, 397-446. Rijeka: Intech Europe.

Rodriguez, E., C.S. Morris, and J.E. Belz. 2006. "A global assessment of the SRTM performance." Photogrammetric Engineering and Remote Sensing 72 (3):249-260.

Rogers, D. J., S. E. Randolph, R. W. Snow, and S. I. Hay. 2002. "Satellite imagery in the study and forecast of malaria." Nature 415 (6872):710-715.

Schuster, G., E. E. Ebert, M. A. Stevenson, R. J. Corner, and C. A. Johansen. 2011. "Application of satellite precipitation data to analyse and model arbovirus activity in the tropics." International Journal of Health Geographics 10. doi: 810.1186/1476-072x-10-8.

Sinka, M. E., M. J. Bangs, S. Manguin, M. Coetzee, C. M. Mbogo, J. Hemingway, A. P. Patil, W. H. Temperley, P. W. Gething, C. W. Kabaria, R. M. Okara, T. Van Boeckel, H. C. J. Godfray, R. E. Harbach, and S. I. Hay. 2010. "The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic precis." Parasites & Vectors 3:34. doi: 10.1186/1756-3305-3-117.

Wagner, V. E., R. Hillrowley, S. A. Narlock, and H. D. Newson. 1979. " Remote-sensing - rapid and accurate method of data acquisition for a newly formed mosquito-control district." Mosquito News 39 (2):283-287.

Weiss, Andrew. 2001. "Topographic position and landforms analysis." Accessed Jul 28, 2016. http://www.jennessent.com/downloads/tpi-poster-tnc_18x22.pdf.

World Health Organization. 2016. “Malaria” [Fact sheet]. Accessed Jul 18, 2016. http://www.who.int/mediacentre/factsheets/fs094/en/.

Xue, Z., M. Gebremichael, R. Ahmad, M. L. Weldu, and A. C. Bagtzoglou. 2011. "Impact of temperature and precipitation on propagation of intestinal schistosomiasis in an irrigated region in Ethiopia: suitability of satellite datasets." Tropical Medicine & International Health 16 (9):1104-1111. doi: 10.1111/j.1365-3156.2011.02820.x.