NATIONWIDE DISTRIBUTIONAL DETERMINANTS FOR

A WILDLIFE AND HUMAN PARASITE

by

IAN DAVID BULLER

B.A., University of Colorado, Boulder, 2012

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Master of Arts

Department of Ecology and Evolutionary Biology

2013
This thesis entitled:
Nationwide Distributional Determinants For A Wildlife And Human Parasite
written by Ian David Buller
has been approved for the Department of Ecology and Evolutionary Biology

Dr. Pieter Johnson, Dept. of Ecology and Evolutionary Biology

Dr. Valerie McKenzie, Dept. of Ecology and Evolutionary Biology

Dr. Carol Wessman, Dept. of Ecology and Evolutionary Biology

Date

The final copy of this thesis has been examined by the signatories, and we
Find that both the content and the form meet acceptable presentation standards
Of scholarly work in the above mentioned discipline.
Understanding the factors that constrain species distributions is a long-standing goal of ecology, although many studies involve only free-living species. Studies of disease occurrence and spread often require broader knowledge of distributional overlap for free-living and parasitic species, emphasizing the importance of determining the distributional constraints on parasites. Flatworm trematodes in the genus *Alaria* are a broadly distributed group of parasites with wildlife and human health implications. Using a 10-year survey of 624 ponds across the United States, we evaluated the relative roles of climate, geology, and land cover for *Alaria* occurrence using species distribution modeling (Maxent). We also conducted a step-wise parameterization of Maxent and a sampling bias control method, which may be useful for improving the functionality of Maxent. From among 26 considered models simulations, we identified the primary *Alaria* occurrence areas that included western and mid-western US with a low probability of predicted occurrence in the central and southern US. The best-fitting *Alaria* model (mean test AUC: 0.829 ± 0.070 SD; average of 10 ensemble models) is comprised of 9 variables including climate, geology, and land cover. Bootstrapping with 20 replicates was found to be the best Maxent method because it maximized mean test AUC and decreased mean standard deviation. Geology was the most important variable explaining 24% of the variation followed by precipitation in the wettest month (15%) and mean temperature of the driest quarter (14%). Land cover was not a substantial explanation of *Alaria* occurrence (7%). Geology is likely mediating through its effect on water quality and pH, while climatic variables may affect the composition of *Alaria* hosts. Our results may help inform predictions of infection risk in wildlife and humans.
ACKNOWLEDGEMENTS

This study would not have been possible without the generous assistance of many individuals. First and foremost, the guidance and support from my advisor, Dr. Pieter T.J. Johnson, is appreciated. I am grateful for the time and availability of the other members of the committee for their helpful advice. For data collection, I am indebted to members of the field crew, fellow amphibian necropsy teammates, and collaborators at the US National Fish and Wildlife Service (USFWS). In particular, I would like to thank Dr. Kim Medeley and Katie Richgels for their insights, revisions, and methodological support. This project was supported by funds from the USFWS, the National Science Foundation, the David and Lucile Packard Foundation, and the Barry M. Goldwater Scholarship.
CONTENTS

SECTION

I. INTRODUCTION .............................................................................................................. 1
   Questions, Hypotheses, Predictions .............................................................................. 4

II. Methods .......................................................................................................................... 6
   Pond Survey ....................................................................................................................... 6
   Laboratory Procedures .................................................................................................... 6
   Distribution Modeling ....................................................................................................... 7
   Selection of Occurrence Points ......................................................................................... 7
   Maxent Settings ............................................................................................................... 8
   Model Selection ............................................................................................................... 9
   Assessing the Most Informative Model .......................................................................... 10

III. Results .......................................................................................................................... 10

IV. Discussion ..................................................................................................................... 12

BIBLIOGRAPHY .................................................................................................................. 19

APPENDIX

A. Tables ............................................................................................................................ 26

B. Figures ........................................................................................................................... 33

C. Panels ............................................................................................................................ 27
TABLES

TABLE

1. Selected environmental variables in Maxent model for *Alaria* in the United States ........................................26

2. Explanation of the three types of replication in Maxent .................................27

3. List of surveyed amphibian hosts with *Alaria* ..............................................28

4. The components of each model .................................................................29

5. The predictive rock types of the best-fitting *Alaria* distribution model and their effect on water pH for each rock type ..............................................................30

6. Comparison of resampling types by mean test AUC score .........................31

7. Comparison of number of bootstrap replications ........................................32
FIGURES

FIGURE

1. All survey sites and which sites were positive for *Alaria*........................33

2. The *Alaria* occurrence sites used in the Maxent models..........................34

3. Importance of environmental variables in
   Slim Global Model..........................................................................................35

4. Effect on water pH relative to predictive rock types .........................36

5. Response curve of BIO 13
   (Precipitation of the Wettest Month) .........................................................37

6. Response curve of BIO 9
   (Mean Temperature of the Driest Month).................................................38

7. Omission/commission curve for Slim Global Model ..........................39

8. Receiver Operating Characteristic (ROC) curve
   for the best-fitting model .............................................................................40

9. Map of the average likelihood of *Alaria* presence
   in the contiguous United States.................................................................41
PANELS

PANEL

1. Photo of *Alaria* mesocercaria.................................................................42
2. The general life cycle of the genus *Alaria*...............................................43
3. Flowchart of the model selection approach.............................................44
INTRODUCTION

The study of species ranges is an active area of research asking the ecological question: Why do species occur in some locations and not others? A species’ range is governed by myriad biotic interactions involving competitors, predators and pathogens as well as abiotic gradients (Brown et al., 1996; Gaston, 2003). Recent advances in technology and sharing of large databases that contain occurrence data for many species as well as environmental data over long time periods have facilitated studies of species distributions on a broad spatial scale (Sexton et al., 2009). However, these datasets are still in short supply hindering further investigation into these factors, which, when understood, can serve to test and forecast ecological responses to large-scale, anthropogenic habitat changes (Sexton et al., 2009).

Many previous distributional studies have focused on free-living species, but there is a need to consider parasitic species ranges because disease-causing agents, including parasites and pathogens, have been recognized as a major component of communities and underlying ecological mechanisms (Szuroczki and Richardson, 2009; Poulin, 1999; Hudson et al., 2006). For instance, Lafferty et al. (2006) found that parasites play a major role in increasing the connectivity and nestedness of food webs, possibly leading to more stable ecosystems when parasites are diverse and abundant. The loss of parasite diversity and abundance in freshwater, estuarine, and marine ecosystems has been used as a bioindicator of anthropogenic pollution (Thielen et al., 2004; Huspeni et al., 2005; Malek et al., 2007), and other anthropogenic stresses such as eutrophication, pesticides, land-use change, and global climate change (Johnson et al., 2007; King et al., 2007; McKenzie and Townsend, 2007; Rohr et al., 2008a; Rohr et al., 2008b; Marcogliese, 2008).
However, understanding parasite species’ distributions is challenged by the fact that parasites depend not just on the physical environment, but also on their host environment. Parasites exist only in areas that have the suitable combination of abiotic and biotic conditions for their hosts (fundamental niche) in which the parasite can maintain its populations without immigration (Holt and Gomulkiewicz, 1996). Therefore, parasites that reside within a host should match their host ranges, although this may not be true for parasites that require multiple hosts where the parasite’s range would be the overlap of all their host ranges. Not all parasites utilize one host in their life cycle, which adds in complexity of multiple overlapping host distributions as well as adding the intricacies of the parasite’s transmission dynamics. Moreover, if a parasite has a free-living stage then the parasite is exposed to the environment, which increases the direct influence of the environment on a parasite’s survival and limits the realized niche of a parasite or where a parasite does exist. Understanding what environmental conditions affect parasite species’ distributions will help assess whether host dynamics are important for their parasites’ distributions and will allow for better assessments of disease risk in humans and wildlife.

Modeling the spatial distributions of parasite species of wildlife can potentially illuminate the environmental determinants of infection. Species Distribution Models (SDMs) use machine-learning algorithms to discover complex non-linear relationships between infection presence data and environmental predictor variables overlaid on a Geographic Information System (GIS) platform. The outcome is a prediction of the spatial distribution of a species, but there are a few, notable potential weaknesses of SDMs. First, SDMs use presence-only data, which does not enable analysis of bias or estimation of prevalence (Phillips et al. 2009). But, absence records can potentially be misleading because of detection biases (Elith and Leathwick, 2009). Also,
species are affected by small-scale habitat variables, such as topography, that are likely to be overlooked in large grid cell (1 km²) data used by SDMs (Williams et al., 2003). Although, there is an ongoing discussion about the relative influence of small or large scale environmental variables on species distributions (Case et al., 2005); large-scale environmental variables may be enough to define species ranges. Finally, abiotic variables used in SDMs can be associated with various ecological dynamics, indicating SDMs show correlations, not causations, between species occurrences and environmental variables (Sinclair et al., 2010). Still, SDMs offer initial corollary investigation into the environmental determinants of species ranges and can predict areas for further study.

Many studies have applied SDMs to investigate the spatial distribution of free-living species, but SDMs have not been used to study parasites as often. SDMs can become a useful tool to understand the spatial distributions of parasites when not every aspect of the parasite’s history and patterns is known (Peterson et al., 2002; Peterson et al., 2004). SDMs have been used to describe spatial patterns of a range of pathogenic human infections including Malaria, monkeypox, Lyme disease, Japanese Encephalitis Virus, and Leshmaniasis (Moffett et al., 2007; Levine et al., 2007; Mak et al., 2010; González et al., 2010; Miller et al., 2012), but few have used SDMs to study a multi-host parasite of wildlife and humans.

Populations of flatworm parasites within the class Trematoda (“trematodes”) and their amphibian hosts are a system well suited to study the species ranges of parasites. Parasites in the genus Alaria are found in freshwater amphibian hosts (Panel 1; Johnson et al., 1999). Similar to other flatworm parasites, Alaria has a complex life cycle that involves three to four distinct hosts at discrete stages (Panel 2; Olsen, 1974). Because of this complex lifecycle and its low pathology in amphibian hosts, Koprivnikar et al. (2012) proposed Alaria can be an effective bioindicator of
host diversity when present, but this has not been investigated on a large scale. Additionally, humans can become infected by ingesting uncooked meat that contains larval worms (Beaver et al., 1977) or by handling infected meat and eyes (McDonald et al., 1994). One fatal case was caused by pulmonary hemorrhage (Freeman et al., 1976). Understanding the distribution of Alaria will be useful for conservation biology and human health. And while the spatial distribution of the amphibian fungus pathogen Batrachochytrium dendrobatidis has been studied using SDMs (Kaiser and Pollinger, 2012), no amphibian macroparasite, such as Alaria, has been similarly investigated.

Questions, Hypotheses, Predictions

Using occurrences of Alaria, environmental variables, and a species distribution model, we asked the question: What are the relative roles of abiotic variables (climate, geology, land cover) for Alaria presence across its range?

At present, the ecology of Alaria is poorly understood. Although many ecological variables can affect the complex life cycle of Alaria (Panel 2), there is only one previous study on the ecology of Alaria. This study reported that the presence of definitive hosts at sites with the necessary snail host is influenced by the amount of forest in the surrounding habitat (Koprivnikar et al., 2006). Forested habitat, determined from remote sensing databases, near a site provides a suitable habitat and promotes the dispersal of infected definitive hosts, which are more commonly found in forested habitats away from human development (Kuehl and Clark, 2002). Nevertheless, this claim has not been widely studied and Alaria uses a large diversity of definitive hosts that may be affected by forested habitat in varying degrees. In the present study, we are going beyond a local-level ecological study by utilizing a large Alaria occurrence
database, which allows us to use SDMs to study the environmental determinants of a countrywide distribution of *Alaria*.

In exploring patterns of trematode distribution generally, there are at least two other environmental variables that likely affect parasite species distributions. Climatic variables including temperature and precipitation can be predictors of amphibian, snail, and definitive hosts occurrence or abundances (Johnson *et al.*, 2007; Holt *et al.*, 2009; Kaiser and Pollinger, 2012). Additionally, pH may be another environmental variable that influences trematodes. The overall relationship between pH and trematode infection is not well understood (Koprivnikar *et al.*, 2010). However, one study found that lower pH has a slight detrimental effect on cercariae survival (Shostak, 1993), meaning there are less viable cercariae in an acidic pond to infect amphibians. Buller (unpublished) found higher pH (i.e. more alkaline ponds) negatively correlated with average *Alaria* abundance suggesting cercaria are sensitive to pH outside of neutral. Pond habitat pH can be remotely sensed using geology (i.e. rock type) as a proxy variable because certain rock types can leach into water bodies and affect water pH (Newton *et al.*, 1987).

Here, we sought to compare the relative importance of abiotic variables including climate, geology, and land cover. We hypothesized that land cover will be a major contributing factor to the species distribution of *Alaria*, but that it would not completely explain all of the variation in *Alaria* occurrence; climate and geology will have a role because they have been shown to affect parasites similar to *Alaria*. 
METHODS

Wetland surveys

A total of 624 wetlands across the contiguous United States were surveyed between 1997 and 2011 for trematode infections in amphibians. This included a wide range of habitat types (from small ponds to large lakes) and a diversity of amphibian host species. Many of these collections involved a long-term collaboration with the US Fish and Wildlife Service, in which agency personnel sampled amphibians from National Wildlife Refuges across the country to quantify the occurrence of amphibian malformations. Additional collections stemmed from research projects and opportunistic sampling events associated with the Johnson Lab at the University of Colorado.

Laboratory procedures

Amphibian specimens were euthanized in lab using a buffered solution of MS-222 anesthetic (IACUC Protocol #1002.02). Parasites were removed from amphibian tissues and identified to genus using “Keys to the Trematoda” (Gibson et al., 2002). Parasite infection was quantified for *Alaria* following standard procedures. These procedures included inspection of outer and inner skin surface and all internal organs of amphibians for parasite infection. *Alaria* mesocercariae were found alive and were commonly located under the skin, in the muscles, and within the body cavity of amphibian hosts. Both mobility behavior and the distinct forked digestive tract distinguished *Alaria* mesocercaria from other macroparasites inside necropsied specimens. Identification was confirmed by fixing *Alaria* with distilled water onto a glass slide and examination using a light-compound microscope. Data were added to a multiyear (>ten years) database called the Amphibian Parasite Observatory (www.amphibianparasites.org). We
were unable to identify *Alaria* to species given lack of morphological traits in the larval mesocercariae stage. These traits are described in the adult stage found in definitive hosts. Although we quantified infection abundance (number of parasites per individual host), here we only focus on occurrence because species distribution models use the latitude and longitude positions of species occurrences to predict species distributions.

**Distribution modeling**

We used Maxent, a maximum entropy-based species distribution prediction-modeling tool, to develop distribution models for *Alaria* and to assess the environmental variables that contribute to the best-fitting model. A bias-controlled dataset (described below) was used with Maxent 3.3.3k (Phillips et al., 2006), because Maxent has been found to perform best among many different modeling methods (Elith et al., 2006; Ortega-Huerta and Peterson, 2008). Maxent is a machine-learning program that uses occurrence only points to create the least biased spatial distribution of a species by estimating the probability distribution for a species occurrence based on environmental conditions (Phillips et al., 2006). We considered 21 remotely sensed climatic, geological, and land cover type data variables as potential predictors of *Alaria* habitat distribution (Table 1). Nineteen bioclimatic variables were obtained from WorldClim dataset (Hijmans et al., 2005; www.worldclim.or/bioclim). Geology data were obtained from the US Department of Agriculture (datagateway.nrcs.usda.gov) and land cover data from the Multi-Resolution Land Characteristic Consortium National Land Cover Database (www.mrlc.gov). There was no *a priori* selection of climate variables used in the models because the species distribution of *Alaria* or any North American amphibian trematode has not been studied. Geology (i.e. rock type) was used as a proxy for water body pH. The geological and land cover
data layers were resampled from 1.25km$^2$ and 0.03km$^2$, respectively, to match the 1 km$^2$ climate data layers resolution using ArcGIS 10.1 to give all the layers the same grid size.

**Selection of occurrence points**

*Alaria* occurrence points were extracted from the Amphibian Parasite Observatory. Any site that had *Alaria* at least once in the database was considered a positive site. Thus *Alaria* positive sites with repeat sample or any site with at least one amphibian sampled were included. For the 624 sites sampled across the United States, only 90 sites were positive for *Alaria* (Fig. 1). Because there was an apparent sampling bias for *Alaria* in the San Francisco Bay area, we controlled for the sampling bias by equalizing sampling density across the extent of the contiguous United States. We calculated the average nearest neighbor distance between all positive sites (85 km) using ArcMap 10.1 (Environmental Systems Research Institute, 2011), and randomly selected the maximum number of sites that were at least 85 km apart. Distribution models were generated using this subset of *Alaria* occurrence sites, which consisted of 36 positive sites (Fig. 2). This new sample size is well above the minimum number of sites required (n=25) to run an effective Maxent model (Hernandez *et al.*, 2006), and reduced model bias because points were relatively evenly distributed across space.

**Maxent settings**

We ran all Maxent models with 5,000 iterations per replicate (Young *et al.*, 2011) and only used Linear, Quadratic, and Hinge features in the model (Phillips and Dudik, 2008). We tested model fit using default settings, but adjusted replication type and the number of replications to improve model fit. Therefore, we carried out a model comparison of a climate
model with all 19 BIOCLIM variables with *Alaria* presence sites for all three replication run types at n=5, 10, 15, 20, 25 and 30 replications. The three types of replication were cross-validation, repeated subsampling, and bootstrapping, which are explained in Table 2.

**Model selection**

To evaluate model fit, we randomly partitioned occurrence points into training (75% of points) and testing (25% of points) datasets (Phillips and Dudik, 2008). The Area Under the Curve (AUC) of the Receiver Operating Curve (ROC) was used to evaluate model fit. AUC, which ranges from 0 to 1, is a composite value of model fit, and is a good measure of model quality (Elith, 2006). We compared model fit for a suite of models representing our different predications for environmental factors constraining *Alaria* distribution (below). To select the most informative Maxent model for *Alaria*, we carried out a stepwise model selection (Panel 3) among models that included combinations of three types of environmental variables (climate, geology, and land cover). We created a model with all 19 climate variables as a baseline and then pruned the resulting model by removing the variables with low explanatory value (that explained the bottom 10% of the model). We added geology and land cover variables to the slimmed climate model and compared AUC to see how geology and land cover improved the climate model. Finally, the best model with a combination of climate, geology, and land cover was pruned, removing the variables that explained the bottom 10% of the model. This final model was compared against its predecessor with AUC to see how simplifying the model improved the model fit.
Assessing the most informative model

The final *Alaria* potential habitat map was generated based on the best-fitting Maxent model. The map was qualitatively assessed highlighting potential new locations to sample *Alaria*. We used ‘lowest presence threshold’ (LPT, equal to the lowest probability at the species presence locations) to assess if the best-fitting model significantly fit *Alaria* occurrence data (Pearson *et al.*, 2007). The relative contribution of variables within the most informative model was chosen based on their permutation importance (Graham *et al.*, 2011).

RESULTS

In a survey of 624 sites in the contiguous United States, *Alaria* was found in 90 sites, mostly in the western United States, Ohio River Valley, and Midwestern United States (Fig. 1), but also found in the eastern United States to a lesser extent. *Alaria* was not detected in central and southern United States. Also, 15 different amphibian host species were infected with *Alaria* (Table 3). The types of amphibians ranged from toads to frogs and a newt. To the best of our knowledge, this is the first *Alaria* host record for eight amphibian species (Table 3).

From among 26 models considered, we identified the best-fitting *Alaria* model (Mean Test AUC: 0.829 ± 0.070 SD; average of 10 ensemble models) as one comprised of nine variables including climate, geology, and land cover. The components and summary of each model are presented in Table 4. Geology (i.e. rock type) was the most important variable (Fig. 3) explaining nearly 25% of the model variance followed by precipitation in the wettest month (15%, “BIO 13”) and mean temperature of the driest quarter (14%, “BIO 9”). Land cover, however, was not a substantial contributor to *Alaria* occurrence (7%). Out of the 174 different geological rock types included in the geology suite of variables, 19 were predictive of *Alaria* occurrence (Table 5); of these predictive rock types almost 60% are associated with neutral water...
body pH (Fig. 4). As precipitation in the wettest month (BIO 13) reached a certain threshold, the probability of *Alaria* occurrence increased above 50% (Fig. 5) and as the mean temperature of the driest quarter (BIO 9) increased the probability of *Alaria* occurrence also increased (Fig. 6).

Based on the Omission/Commission Curve (Fig. 7), there is still some spatial autocorrelation inherent in *Alaria* occurrence across the United States because the omission line does not match the null 1:1 unbiased line. However, the model significantly predicts over 82% (p=0.02) of *Alaria* presences in the United States at the lowest presence threshold (Fig. 8). The predicted *Alaria* occurrence areas (Fig. 9) included the western and Midwestern United States and the Ohio River Valley. The best-fitting model predicted little to no *Alaria* occurrence in the Rocky Mountains and southern United States. Potential areas where *Alaria* may occur but were not detected in our survey include central Arizona, southern California, Idaho, Arkansas, and Wisconsin.

Before assessing the best-fitting model, a parameterization of Maxent was carried out, testing for the best replication type and replication number for later model selection of *Alaria* distribution. Bootstrapping was the replication run type chosen (Table 2) for the *Alaria* distribution model selection based on its higher grand average test AUC score than the other two types (Table 6). The number of replications was chosen at 20 for the *Alaria* model selection because it had the lowest standard deviation between bootstrapping replication runs (Table 7).

During the model selection (Panel 3), two models (Full Climate and Full Global) were “slimmed” of variables that explained the lower 10% of the full models. The removed variables were no longer used in further model selection. Only 12 of the 19 BIOCLIM variables accounted for 90% of the Full Climate model (0.831 ± 0.051 SD) resulting in the Slim Climate Model (0.824 ± 0.087 SD). Including both geology and land cover into the Slim Climate Model reduced...
model performance (AUC 0.815 ± 0.068 SD), but after five climate variables were removed from this global model, the resulting Slim Global model was the best performing Maxent model for *Alaria* presence in the contiguous United States (Table 4).

**DISCUSSION**

Using a 14-year, nationwide dataset of 624 ponds and a machine-learning algorithm with GIS-derived data, we provide the first example of a species distribution model for a human and wildlife trematode parasite (*Alaria*). Our results showed that *Alaria* has a distinct US distribution with high probabilities of occurrence in the western and Midwestern United States. Using model selection, the best combination of remotely sensed climatic, geological, and land cover data for predicting *Alaria* distribution included climatic, geological, and land cover variables. However, in contrast to our hypothesis, land cover was not the most important explanatory variable.

Our results were robust based on two metrics of model fit. First, the high average test AUC score (0.829) is well above the null expectation (AUC: 0.50), which means our best-fitting model predicts almost 83% of *Alaria* occurrence variation. Second, a significant p-value (p < 0.02) at the lowest presence threshold (LPT) shows our model is has a good fit to the test dataset. Maxent randomly selects 25% (n=9) of the known *Alaria* occurrence points (Fig. 2), creates a new “test dataset,” and then evaluates the best-fitting model that was created using the other 75% known *Alaria* occurrence points (called the “training dataset”). The LPT is a conservative metric for model fitness that assesses whether or not the best-fitting model significantly predicts the lowest probability of *Alaria* at its “test” locations. Our best-fitting model significantly predicts (p < 0.02) the “test dataset.”
The role of geology (i.e. rock type) to the distribution of *Alaria* may manifest through various pathways at multiple stages of its life cycle. Geology directly influences water chemistry and nutrient availability in soil, which then impacts the biotic composition of aquatic communities and land cover types. Certain rock types (bedrock) can leach into surface water and affect water chemistry, including nutrients and pH (Holloway *et al.*, 1998; Newton *et al.*, 1987). Differences in nutrient variability can be affected by the extent of a particular rock type, where large basins of constant rock type can have relatively constant water chemistry and mountain streams that drain different rock types from one basin to another can have highly variable water chemistry (Meybeck, 2005).

These water chemistry variations, or lack thereof, can affect the freshwater community at the ponds where *Alaria* spends its larval stages inside an amphibian and snail, including its free-living stage. Amphibians are highly sensitive to non-neutral pH levels, especially embryos (Dale *et al.*, 1986). Nutrient availability influences the macroinvertebrate (susceptible snail host) community through food availability and quality, macroinvertebrate survival, and stream temperature (Huryn *et al.*, 1995; Leland and Porter, 2000; Willoughby and Mappin, 1988). Water pH is known to be a limiting factor of freshwater gastropods and *Helisoma trivolvis*, a snail that serves as the primary intermediate of *Alaria*, experiences mortality and reduced growth at low pH (Hunter, 1989; Watson and Ormerod, 2004). Moreover, parasites outside their hosts can resist broad deviations in pH (Ford *et al.*, 1998; Nollen *et al.*, 1979; Sawabe and Makiya, 1995), but this trait may be species-specific (Pietrock and Marcogliese, 2003). For instance, trematode richness in a parasite community including *Alaria* was significantly negatively affected by pH (Richgels *et al.*, 2013). Almost 60% of predictive rock types for the best-fitting *Alaria*
distribution model are associated with neutral water body pH; therefore, pH may influence where susceptible intermediate host populations can be present and be infected by *Alaria* larvae.

However, the direct relationship between rock type and *Alaria* may not only be through water chemistry, because geology can also influence the immediate soil, which regulates the plant community composition. Geology provides the parent material for the soil that offers certain environmental conditions (nutrients, pH, and radiation), which then dictates what types of plants and microorganisms can thrive (Kruckeberg, 2004). At the pond-level, rock type can affect the plant community through effects on the medium and mineral ability (Kruckeberg, 2004). A vegetated pond may offer protection for tadpoles from cercaria, thus reducing trematode infection. At the landscape level, climate is the primary determinant of soil characteristics (Cain, 1944), but climate and soil are products of geology because geology is highly heterogeneous in terms of topology (elevation) and rock type across the landscape (Kruckeberg, 2004; Verheye, 2009). There is evolutionary pressure for plants to develop a tolerance to unique soils and climates in order to colonize new niches and avoid competition (Gankin and Major, 1964). The community composition of plants in some areas and not others determines the type of land cover in an area and also influences the available types of land use, which could limit the availability and distribution of hosts at nearby ponds.

The identified importance of the climatic variables could also be linked to factors associated with host availability and abundance. Climatic variables are often sufficient enough to predict the distribution of terrestrial, free-living species (Araújo and Guisan, 2006) by defining the suitable habitat a species can occupy based on their physiological tolerances and resource limiting requirements. Therefore, the two major climate variables in the best-fitting *Alaria* distribution model, precipitation of the wettest month (BIO 13) and mean temperature of the
driest quarter (BIO 9), may define the likely habitat where the hosts for *Alaria* overlap or occupy so long as a site is present in a suitable rock type. Taking the two climactic variables together, they appear to affect the hydroperiod or the permanence of a body of water in a yearly cycle and the size of a body of water. If a pond in this study system dries up during the summer, the density of *Alaria* cercariae in the water increases, amplifying amphibian exposure. Size of a water body can influence many aspects of ponds including host richness and water chemistry (Zelmer and Campbell, 2011). As habitat area increases there is more habitat available for hosts (Halpern et al., 2005) and larger, less fragmented habitats have been found to reduce the chance a host species becomes locally extinct at a site when it becomes infected with a pathogen (McCallum and Dobson, 2002). This allows more infected individual hosts to persist at a site, which would increase the likelihood of *Alaria* being detected when surveyed.

Although Koprivnikar et al. (2006) found that local *Alaria* occurrence was influenced by the amount of forest in the surrounding habitat, we did not find a similar pattern. Discrepancy could be a matter of study design. The present study is nationwide while Koprivnikar et al. (2006) was a local, correlative study in Manitoba, which hypothesized that forested regions were attractive for definitive hosts, thus increasing the chance of their presence at a pond and depositing *Alaria* eggs in their feces. *Alaria* is a genus of 8 species that uses a wide range of over 30 species of definitive hosts in the United States (Buller, unpublished review) ranging from highly dispersive coyotes to territorial badgers. Each definitive host has a different preference to environmental conditions and a particular pond may be more attractive for some but not all of the definitive hosts for *Alaria*. Therefore, land cover type may matter for *Alaria* occurrence in north Midwestern United States because of the definitive hosts that reside there. Assessing *Alaria*
distribution regionally across the United States may reveal different important environmental variables than countrywide or locally because of this host effect.

Recent studies have used SDMs to describe the distribution of a pathogen, usually infecting humans, and how disease risk will change over time with climate change. Many studies of vector-borne diseases use SDMs to predict the distribution of their vector hosts as a metric for disease risk (Mak et al., 2010; Miller et al., 2012; Moffett et al., 2007). Daszak et al. (2013) used Maxent to predict future potential distribution of the bat reservoir host of Nipah Virus, an emerging infectious disease of humans, in order to inform efforts for disease control. The study found the predicted change in climatic variables would lead to an expansion of the geographic range of Nipah Virus into new habitats as it tracks its climatic niche (Daszak et al., 2013). Levine et al. (2007) used case reports of monkeypox (Orthopoxviridae spp.) and found that high mean annual precipitation and low elevations were correlated with monkeypox occurrence using another SDM method called GARP (Genetic Algorithm for Rule-set Production). Our study included other environmental variables in addition to climate, and examined a parasite with a free-living stage at its individual occurrence points. However, our study neither predicted future distributions for Alaria based on changing variables nor used topography as a possible predictor of Alaria distribution. Although, in our study we used a superior SDM method (Maxent) based on a model comparison by Elith et al. (2006). To the best of our knowledge, the present study is novel because it uses a SDM to study the distribution of a multi-host parasite of wildlife and humans.

There were a few limitations to the present study that should be addressed. The environmental variables in the best-fitting model affect various aspects of the Alaria lifecycle and are statistically correlative in their association with the distribution of Alaria. SDMs are not
mechanistic models and only reveal statistical relationships between species occurrences and environmental variables. Investigators must be prudent with their use of statistical models to predict outcomes outside of their initial study area or with environmental changes, such as climate change. For example, Ruiz-Moreno et al. (2012) recently predicted the United States might see outbreaks of Chikungunya virus as early as 2014 based on a statistical modeling approach using climate-driven mosquito population data derived from other areas of the world. But, Gubler et al. (2001) explained many vector-borne diseases are not present in the United States because the built environment and human behavior play a large role in disease mitigation even though the United States has an ideal climate for many vector-borne diseases. Future mechanistic studies should be carried out to tease apart the causative link between the environmental variables and *Alaria* occurrence.

Additionally, SDMs are only as reliable as the algorithms and data used to train them. Maxent trains models based upon the realized niche or the actual habitat of a species instead of the fundamental niche; therefore, the *Alaria* distribution may be overpredicted in some areas (Pearson, 2007). Maxent provides a mean *Alaria* distribution in order to address this bias, but the provided *Alaria* map should be considered a possible, not a definitive, distribution. A limitation of assessing different types of environmental variables is the availability of similar, high-resolution data. When resampling land cover and geology to match the climate data resolution, some specificity of the variables was lost. This issue will change with the increasing availability of high-resolution remote-sensing databases. Finally, in addition to abiotic gradients, a species’ range is governed by a myriad of biotic interactions involving competitors, predators and pathogens (Brown *et al.*, 1996). Species distribution models rely heavily on abiotic variables and integrating biotic interactions into species distribution framework is often neglected (Araújo and
Future studies could examine how host richness and abundances improve species distribution models to assess the contribution of biotic versus abiotic variables influence species distributions.

From an applied perspective, the potential species distribution map for *Alaria* can help in discovering new populations, identifying top-priority survey sites, or setting priorities to restore the natural habitat of *Alaria* for more effective conservation of its hosts. Koprivnikar *et al.* (2012) proposed *Alaria* could be a bioindicator of host diversity, so this *Alaria* distribution map may be an indication of areas with high host diversity. The methodology presented here could be used for Maxent parameterization in future studies that require controlling for sampling density, and can choose which replication type and replication number to use in the program. Species distribution models can be used to quantify habitat distribution patterns for other parasitic diseases and may aid in field surveys and allocation of disease control or conservation programs.
BIBLIOGRAPHY


APPENDIX

**Table 1**: Selected environmental variables in Maxent model for *Alaria* in the United States

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIO 1</td>
<td>Annual Mean Temperature</td>
<td></td>
</tr>
<tr>
<td>BIO 2</td>
<td>Mean Diurnal Range</td>
<td></td>
</tr>
<tr>
<td>BIO 3</td>
<td>Isothermality (Temperature evenness)</td>
<td></td>
</tr>
<tr>
<td>BIO 4</td>
<td>Temperature Seasonality</td>
<td></td>
</tr>
<tr>
<td>BIO 5</td>
<td>Max Temperature of Warmest Month</td>
<td></td>
</tr>
<tr>
<td>BIO 6</td>
<td>Min Temperature of Coldest Month</td>
<td></td>
</tr>
<tr>
<td>BIO 7</td>
<td>Temperature Annual Range</td>
<td></td>
</tr>
<tr>
<td>BIO 8</td>
<td>Mean Temperature of Wettest Quarter</td>
<td></td>
</tr>
<tr>
<td>BIO 9</td>
<td>Mean Temperature of Driest Quarter</td>
<td></td>
</tr>
<tr>
<td>BIO 10</td>
<td>Mean Temperature of Warmest Quarter</td>
<td></td>
</tr>
<tr>
<td>BIO 11</td>
<td>Mean Temperature of Coldest Quarter</td>
<td></td>
</tr>
<tr>
<td>BIO 12</td>
<td>Annual Precipitation</td>
<td></td>
</tr>
<tr>
<td>BIO 13</td>
<td>Precipitation of Wettest Month</td>
<td></td>
</tr>
<tr>
<td>BIO 14</td>
<td>Precipitation of Driest Month</td>
<td></td>
</tr>
<tr>
<td>BIO 15</td>
<td>Precipitation Seasonality</td>
<td></td>
</tr>
<tr>
<td>BIO 16</td>
<td>Precipitation of Wettest Quarter</td>
<td></td>
</tr>
<tr>
<td>BIO 17</td>
<td>Precipitation of Driest Quarter</td>
<td></td>
</tr>
<tr>
<td>BIO 18</td>
<td>Precipitation of Warmest Quarter</td>
<td></td>
</tr>
<tr>
<td>BIO 19</td>
<td>Precipitation of Coldest Quarter</td>
<td></td>
</tr>
<tr>
<td>Geology</td>
<td>174 categories types of rock types, proxy for water body pH</td>
<td></td>
</tr>
<tr>
<td>Land Cover</td>
<td>13 categories of land cover types</td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Explanation of the three types of replication in Maxent. These definitions come from Phillips, 2006. The replication type that was chosen for the present study was bootstrapping because it had a higher Area Under the Curve (AUC) score than the other two types when predicting the distribution of *Alaria* using climate data.

<table>
<thead>
<tr>
<th>Replication Type</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-Validation</td>
<td>Occurrence data is randomly split into a number of equal-size groups called “folds”, and models are created leaving out each fold in turn. The left-out folds are then used for evaluation of the better model.</td>
</tr>
<tr>
<td>Repeated Subsampling</td>
<td>Occurrence points are repeatedly split into random training and testing subsets. The training dataset is selected by sampling <em>without</em> replacement from the occurrence points, with the number of samples equaling the total number of presence points. The best model created with the training dataset is tested on how well it predicts the testing dataset.</td>
</tr>
<tr>
<td>Bootstrapping</td>
<td>Occurrence points are split into random training and testing subsets. The training data is selected by sampling <em>with</em> replacement from the occurrence points, with the number of samples equaling the total number of occurrence points. The number of occurrence points in each set equals the total number of occurrence points, so the training datasets will have duplicate records. The best model created with the training dataset is tested on how well it predicts the testing dataset.</td>
</tr>
</tbody>
</table>
Table 3: List of surveyed amphibian hosts with *Alaria*.

<table>
<thead>
<tr>
<th>Name</th>
<th>Common Name</th>
<th>New Host Record?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaxyrus boreas</td>
<td>Western Toad</td>
<td>Y</td>
</tr>
<tr>
<td>Bufo americanus</td>
<td>American Toad</td>
<td>N</td>
</tr>
<tr>
<td>Hyla versicolor</td>
<td>Grey Tree Frog</td>
<td>Y</td>
</tr>
<tr>
<td>Pseudacris maculata</td>
<td>Boreal Chorus Frog</td>
<td>Y</td>
</tr>
<tr>
<td>Pseudacris regilla</td>
<td>Pacific Tree Frog</td>
<td>N</td>
</tr>
<tr>
<td>Pseudacris triseriata</td>
<td>Western Chorus Frog</td>
<td>Y</td>
</tr>
<tr>
<td>Rana aurora</td>
<td>Northern Red-legged Frog</td>
<td>Y</td>
</tr>
<tr>
<td>Rana blairi</td>
<td>Plains Leopard Frog</td>
<td>Y</td>
</tr>
<tr>
<td>Rana catesbeiana</td>
<td>American Bullfrog</td>
<td>N</td>
</tr>
<tr>
<td>Rana clamitans</td>
<td>Green Frog</td>
<td>N</td>
</tr>
<tr>
<td>Rana palustris</td>
<td>Pickerel Frog</td>
<td>Y</td>
</tr>
<tr>
<td>Rana pipiens</td>
<td>Northern Leopard Frog</td>
<td>N</td>
</tr>
<tr>
<td>Rana sphenocephala</td>
<td>Southern Leopard Frog</td>
<td>N</td>
</tr>
<tr>
<td>Rana sylvatica</td>
<td>Wood Frog</td>
<td>N</td>
</tr>
<tr>
<td>Taricha torosa</td>
<td>California Newt</td>
<td>Y</td>
</tr>
</tbody>
</table>
Table 4: The components of each model. A "1" in a column signifies the variable included in a model. The best model has a high average test AUC score and a low standard deviation (SD). Under "Replication Type" the three types were cross validation (CV), subsampling (SS), and bootstrapping (BS). Slim models were created with the variables explaining 90% of the full model.
Table 5: The predictive rock types of the best-fitting *Alaria* distribution model and their effect on water pH for each rock type.

<table>
<thead>
<tr>
<th>Rock Type</th>
<th>Effect on pH</th>
<th>Percent of Observed Alaria Occurrence Points (n=36)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alluvium</td>
<td>Neutral</td>
<td>14</td>
</tr>
<tr>
<td>Clay or mud</td>
<td>Alkaline</td>
<td>11</td>
</tr>
<tr>
<td>Limestone</td>
<td>Alkaline</td>
<td>9</td>
</tr>
<tr>
<td>Conglomerate</td>
<td>Acidic</td>
<td>6</td>
</tr>
<tr>
<td>Basalt</td>
<td>Alkaline</td>
<td>6</td>
</tr>
<tr>
<td>Granite</td>
<td>Alkaline</td>
<td>6</td>
</tr>
<tr>
<td>Dolomite</td>
<td>Neutral</td>
<td>6</td>
</tr>
<tr>
<td>Sand</td>
<td>Neutral</td>
<td>6</td>
</tr>
<tr>
<td>Sandstone</td>
<td>Neutral</td>
<td>6</td>
</tr>
<tr>
<td>Shale</td>
<td>Neutral</td>
<td>6</td>
</tr>
<tr>
<td>Black shale</td>
<td>Acidic</td>
<td>3</td>
</tr>
<tr>
<td>Andesite</td>
<td>Alkaline</td>
<td>3</td>
</tr>
<tr>
<td>Tholeiite</td>
<td>Alkaline</td>
<td>3</td>
</tr>
<tr>
<td>Dacite</td>
<td>Neutral</td>
<td>3</td>
</tr>
<tr>
<td>Gravel</td>
<td>Neutral</td>
<td>3</td>
</tr>
<tr>
<td>Melange</td>
<td>Neutral</td>
<td>3</td>
</tr>
<tr>
<td>Meta-argillite</td>
<td>Neutral</td>
<td>3</td>
</tr>
<tr>
<td>Phyllite</td>
<td>Neutral</td>
<td>3</td>
</tr>
<tr>
<td>Troctolite</td>
<td>Neutral</td>
<td>3</td>
</tr>
</tbody>
</table>
Table 6: Comparing Resampling Types by Mean Test AUC Score. Bootstrapping has substantially higher Mean Test AUC scores than the other two resampling types.

<table>
<thead>
<tr>
<th>Number of Replications</th>
<th>Resampling Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cross Validation</td>
</tr>
<tr>
<td>5</td>
<td>0.781</td>
</tr>
<tr>
<td>10</td>
<td>0.757</td>
</tr>
<tr>
<td>15</td>
<td>0.742</td>
</tr>
<tr>
<td>20</td>
<td>0.770</td>
</tr>
<tr>
<td>25</td>
<td>0.721</td>
</tr>
<tr>
<td>30</td>
<td>0.716</td>
</tr>
</tbody>
</table>
Table 7: Comparing Number of Bootstrap Replications. The smallest standard deviation occurred when Maxent was replicated 20 times.

<table>
<thead>
<tr>
<th>Number of Replications</th>
<th>Mean Test AUC Score</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.854</td>
<td>0.065</td>
</tr>
<tr>
<td>10</td>
<td>0.821</td>
<td>0.071</td>
</tr>
<tr>
<td>15</td>
<td>0.806</td>
<td>0.073</td>
</tr>
<tr>
<td>20</td>
<td>0.831</td>
<td>0.051</td>
</tr>
<tr>
<td>25</td>
<td>0.812</td>
<td>0.072</td>
</tr>
<tr>
<td>30</td>
<td>0.784</td>
<td>0.098</td>
</tr>
</tbody>
</table>
Figure 1: All survey sites and which sites were positive for *Alaria*. Absences are included to justify studying the entire extent of the contiguous United States in Maxent and show there were areas we sampled where we did not find *Alaria*. 
Figure 2: The *Alaria* occurrence sites used in the Maxent models which were controlled for sampling bias by equalizing sampling density using the average nearest neighbor distance between all positive sites (85 km)
**Figure 3**: The Importance of Environmental Variables in the Slim Global Model. The definition of the BIOCLIM variables can be found in Table 1.
**Figure 4**: The Effect on Water pH relative to Predictive Rock Types. Majority (58%) of the 19 predictive rock types had a neutral effect on water pH.
Figure 5: The Response Curve of BIO 13 (Precipitation of the Wettest Month). This shows how BIO 13 affects the Maxent prediction for Alaria occurrence distribution, when keeping all other 8 variables in the model at their average sample value. The red curve is the mean response of the 20 replicate Maxent runs. The blue area is the mean ± 1 standard deviation. There are no units for BIO 13, only the overall trend can be assessed.
Figure 6: The Response Curve of BIO 9 (Mean Temperature of the Driest Month). This shows how BIO 9 affects the Maxent prediction for *Alaria* occurrence distribution, when keeping all other 8 variables in the model at their average sample value. The red curve is the mean response of the 20 replicate Maxent runs. The blue area is the mean ± 1 standard deviation. There are no units for BIO 9, only the overall trend can be assessed.
**Figure 7:** Omission/Commission Curve for the Slim Global Model. This curve tests for biases in the occurrence data. There may be some spatial autocorrelation in the *Alaria* presence points (n=36) even when accounting for sampling bias because the mean omission curve is not matching up with the predicted 1:1 omission line. Sampling effort was higher in North and Western US than East and Southern US, the spatial autocorrelation may be a relic of the study design. Omission (false negatives) is underprediction of presence, while commission (false positives) is overprediction of presence. The slim global model is the best-fitting model that was created with the variables that explained 90% of the full global model.
**Figure 8**: Receiver Operating Characteristic (ROC) curve for the best-fitting model. The curve is close to a sensitivity of 1, which signifies the high model fitness to *Alaria* presences. The black line is the null hypothesis that a model predicts an *Alaria* presence about 50% of the time. Having an area under the ROC closer to one creates the red curve seen above, which indicates the best-fitting model has good fit with the predicted distribution of *Alaria*. The blue area is the average standard deviation of the 20 replication runs.
Figure 9: Map of the average likelihood of Alaria presence in the contiguous United States. This is based on the best-fitting model with climate, geological, and land cover variables. Warmer colors signify a higher likelihood of Alaria presence.
Panel 1: Photo of *Alaria* mesocercaria. Under a light-compound microscope (x100)
Panel 2: The general life cycle of the genus *Alaria*. Steps 1 through 5 signify a complete three-stage life cycle involving only three hosts, which is commonly found in other aquatic trematodes and can be true for *Alaria*. In step (1), the parasite is sexually mature in the definitive host digestion tract and produces eggs that are deposited through feces into water. In step (2), the eggs hatch into miracidia, a free-swimming worm stage that seeks out and infects a snail host. In step (3), a miracidium migrates to the reproductive glands of a snail, castrates the host, and asexually produces sporocysts, a stage which asexually reproduces cercariae. In step (4), cercariae, another free-swimming worm stage, are released into the water and infect an amphibian host. In step (5), the cercariae have matured into mesocercariae, a non-reproductively mature larval worm, and wait for the amphibian host to be eaten by a definitive host, thus completing the life cycle. Steps A through C involve an additional path to the life cycle of *Alaria*. In step (A), a paratenic host consumes an amphibian host and collects a high density of mesocercariae. In step (B), different paratenic hosts can consume each other. In step (C), the paratenic host is either eaten by a definitive host thus completing the cycle or the mesocercariae can be passed from a pregnant mother to offspring through blood or breast milk via vertical transmission.
**Panel 3:** Flowchart of the model selection approach. The boxes signify the types of models that were run with their component environmental variables. Dashed lines signify a model was “slimmed” or re-ran with the variables that explained the top 90% of the model. A dot signifies a un-chosen model. The better model was selected based on a higher Area Under the Curve (AUC) score unless slimmed models were within an AUC 0.010 range because simpler models are more powerful than complex models. In the first stage, model “C” had the higher AUC score (0.831). In the second stage, model “C” was slimmed to create model “S” with an AUC score (0.824) within the acceptable range. In the third stage, land cover and geology were added to model “S” with the better model being a global model including all three variables (AUC 0.815). In the final stage, this global model was slimmed to create the best fitting model “B” (AUC 0.829).

Legend:
B = Slim Global Model
C = Climate
G = Geology
L = Land cover
S = Slim Climate Model